The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2083–2100
https://doi.org/10.1214/21-AOAS1554

SEMIPARAMETRIC BAYESIAN FORECASTING OF SPATIOTEMPORAL EARTHQUAKE OCCURRENCES

BY GORDON J. ROSS[a] AND ALEKSANDAR A. KOLEV[b]

School of Mathematics, University of Edinburgh, [a]*gordon.ross@ed.ac.uk,* [b]*aleksandar.kolev@ed.ac.uk*

The Epidemic Type Aftershock Sequence (ETAS) model is a self-exciting point process which is used to model and forecast the occurrence of earthquakes in a geographical region. The ETAS model assumes that the occurrence of mainshock earthquakes follows an inhomogeneous spatial point process, with their aftershock earthquakes modelled via a separate triggering kernel. Most previous studies of the ETAS model have relied on point estimates of the model parameters, due to the complexity of the likelihood function and the difficulty in estimating an appropriate spatial mainshock distribution. In order to take estimation uncertainty into account, we instead propose a fully Bayesian formulation of the ETAS model, which uses a nonparametric Dirichlet process mixture prior to capture the spatial mainshock process, and show how efficient parameter inference can be carried out using auxiliary latent variables. We demonstrate how our model can be used for medium-term earthquake forecasts in a number of geographical regions.

1. Introduction. Self-exciting point process models, such as the Hawkes process (Hawkes (1971)), are widely used to describe events which are clustered in time and/or space. Applications of the Hawkes process span a diverse range of fields, such as criminology (Mohler et al. (2011)) and neuroscience (Chornoboy, Schramm and Karr (1988)). A particular class of self-exciting processes, known as the Epidemic-Type Aftershock Sequence (ETAS) model, is often used in seismology to forecast earthquake occurrences and quantify seismic risk (Fox, Schoenberg and Gordon (2016), Ogata (1988)).

The standard ETAS model has two components: an inhomogeneous spatial background process, which describes the long-term rate of mainshock activity in a geographical area, and a triggering component which describes the rate and location of aftershock events. The self-exciting nature of the model allows it to capture the empirical fact that earthquakes tend to be clustered in time and space. This is due to the underlying physical mechanisms which produce earthquakes (Ogata (1988)), since large magnitude earthquakes tend to causally trigger a substantial number of near-by aftershocks.

Parameter estimation in the ETAS model is known to be difficult for two reasons (Veen and Schoenberg (2008), Zhuang, Ogata and Vere-Jones (2002)): first, the likelihood function is complex and exhibits multimodality and extended flat regions. Second, the model requires a specification of the inhomogeneous background process which can be challenging when only a limited number of mainshock events are available. Due to these complications, the majority of the ETAS literature focuses only on point estimates of the model parameters and ignores any estimation uncertainty, although some recent exceptions are Fox, Schoenberg and Gordon (2016) and Wang, Schoenberg and Jackson (2010).

While the majority of previous work on the ETAS model is carried out in a frequentist framework, the Bayesian paradigm provides a natural alternative for incorporating parameter uncertainty. However, the application of Bayesian methods to the ETAS model have been limited by the complexity of the likelihood function, and many of the approaches, which purport

Received August 2021.
Key words and phrases. Dirichlet process, ETAS, Hawkes process, KDE, spatial analysis, Italy, Bayesian analysis, seismology, aftershocks.

to be Bayesian, still rely on point estimates of model parameters (Ebrahimian et al. (2014)) with the prior only used for regularisation purposes. Recently, Markov chain Monte Carlo (MCMC) has been applied to perform Bayesian inference for the temporal ETAS model (without a spatial component) (Ebrahimian and Jalayer (2017), Omi et al. (2016), Vargas (2012), Kumazawa and Ogata (2014)), although it was pointed out by Ross (2021) that such direct MCMC schemes can suffer from serious autocorrelation problems due to the high correlation of parameters in the likelihood function which results in a low effective sample size. An alternative approach to inference, using latent variables, was introduced by (Ross (2021)) who also applied it to the spatial version of the ETAS model with a constant background rate; however, a constant background rate is extremely unrealistic and will not lead to accurate spatial forecasts on real-world earthquake catalogs. Bayesian estimation of an inhomogeneous background rate has not, to our knowledge, been previously considered.

In this paper we present the first fully Bayesian treatment of the spatial ETAS model. Our main innovation is the use of a nonparametric Dirichlet process mixture prior to estimate the inhomogeneous spatial background rate. Due to the complexity of the ETAS likelihood function, direct estimation of the resulting nonparametric model is not feasible. As such, we develop an estimation procedure using auxiliary latent variables which is an extension of recent work by Markwick and Ross (2021) originally proposed for temporal Hawkes processes.

We begin in Section 2 with a brief introduction to the spatial ETAS model and its use in modelling clustered event sequences. Then, in Section 3 we introduce the Dirichlet process prior that will be used for nonparametric spatial estimation. We next discuss posterior inference strategies using auxiliary latent variables in Section 4. Issues relating to prior choice and forecasting are discussed in Sections 5 and 6. Finally, we study the performance of the resulting spatial ETAS model on real earthquake catalogs in Sections 7 and 8.

2. Spatiotemporal ETAS model. In the ETAS formulation (Ogata (1998)), earthquakes are modelled as events (t_i, m_i, x_i, y_i) from a marked point process on the temporal interval $[0, T]$ and spatial region Σ, where t_i denotes the time at which the ith earthquake occurred, with $m_i \in \mathbb{R}^+$, and $(x_i, y_i) \in \Sigma$ denoting its magnitude and spatial location. It is well known that earthquakes cluster together in both space and time, since large earthquakes tend to trigger further earthquakes nearby, known as aftershocks. To capture this behaviour, the ETAS point process uses the following conditional intensity function:

$$\lambda(t, m, x, y \mid \mathscr{H}_t) = \mu\phi(x, y) + \sum_{t_i < t} k(m_i) h(t - t_i) s(x - x_i, y - y_i), \tag{1}$$

where $\mathscr{H}_t = \{(t_i, m_i, x_i, y_i); t_i < t\}$ denotes the history of the process before time t. The magnitudes m_i are assumed to be independent and identically distributed according to the usual Gutenberg–Richter (G–R) law: $m_i - M_0 \sim \text{Exponential}(\beta)$. As a point of notation, we will suppress all explicit dependence on $\mathscr{H}_t$ and simply write $\lambda(\cdot)$ for the conditional intensity function.

The function $\phi(\cdot) > 0$ is chosen to be a probability density that integrates to 1 over the spatial region Σ. Along with the scaling constant μ, it is known as the background process and specifies the baseline rate at which earthquakes occur, while the functions $k(\cdot)$, $h(\cdot)$, $s(\cdot)$ determine the contribution of each previous earthquake at $t_i < t$ to the intensity at time t. The temporal function $h(\cdot)$ is typically chosen to be monotonically decreasing, which implies that each earthquake causes the process intensity to temporarily increase for a period of time, producing local clusters of events.

The ETAS model can, equivalently, be viewed as a branching process, as was first noted by (Hawkes and Oakes (1974)). At each time t, suppose that n_t events occurred prior to t. From equation (1) the conditional intensity at time t is a linear superposition of $n_t + 1$ independent inhomogeneous Poisson processes, where the first is the background process which contributes intensity $\mu\phi(x, y)$ and the remainder are indexed by each of the n_t previous events with each contributing intensity $k(m_i) \times h(t - t_i) \times s(x - x_i, y - y_i)$, respectively. Since these processes are independent, each event time t_i can be assumed to have been generated either by the background process with intensity $\mu\phi(x, y)$ or by one of the processes triggered by one of the previous events. As a point of convention, we will refer to the events from the background process as immigrant events or mainshocks, with the triggered events being aftershocks or offspring.

2.1. *Specific functional form.* To complete the specification of the ETAS model, we must specify the functions $\phi(\cdot)$, $k(\cdot)$, $h(\cdot)$ and $s(\cdot)$. It is common to take $h(\cdot)$ to be the modified Omori law, which has been empirically shown to capture the temporal decay of earthquake productivity (Utsu and Ogata (1995)), and to take $k(m)$ to be exponential (Ogata (1988), Fox, Schoenberg and Gordon (2016)),

$$h(z) = (p-1)c^{p-1}\frac{1}{(z+c)^p}, \qquad k(m) = Ke^{\alpha(m_i - M_0)},$$

where (c, p, K, α) are model parameters and M_0 is the magnitude of completeness of the catalog which is determined empirically and corresponds to the minimum magnitude above which all earthquakes are successfully detected (Wiemer and Wyss (2000)). Several forms have been proposed for the spatial kernel $s(\cdot)$ (Ogata (1998), Fox, Schoenberg and Gordon (2016)), such as power law decay $s(x - x_i, y - y_i) = \{(x - x_i)^2 + (y - y_i)^2 + d\}^{-q}$ or a spherical bivariate Gaussian,

$$s(x - x_i, y - y_i) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\left[\frac{(x - x_i)^2}{\sigma_x^2} + \frac{(y - y_i)^2}{\sigma_y^2}\right]\right) \tag{2}$$

with parameters σ_x and σ_y. Without loss of generality, we will use this Gaussian kernel throughout. Finally, the background intensity $\phi(x, y)$ defines the long-term distribution of seismicity over the spatial region being considered. This is the most difficult function to choose, since it will depend on the particular spatial region being studied. Many studies assume that seismicity is constant over space so that $\phi(x, y) \propto 1$; however, this is highly unrealistic since it is known that earthquakes tend to occur along geological fault lines. As such, several methods have been proposed to model and estimate the spatial dependence of $\phi(x, y)$, such as Chiodi and Adelfio (2017), Fox, Schoenberg and Gordon (2016), Gordon (2017), Helmstetter (2006), Marsan and Lengline (2008), Zhuang, Ogata and Vere-Jones (2002), which typically uses some kind of kernel smoothing. However the limitation of this approach is that it is difficult to capture the inherent uncertainty in the estimate which can be especially problematic in regions where only a small number of historical earthquakes are mainshocks from $\mu\phi(x, y)$ rather than triggered events.

The primary goal of this paper is to develop a novel nonparametric Bayesian method for the efficient estimation of $\phi(x, y)$ which captures all underlying uncertainty and allows for a fully probabilistic estimate of the level of background seismicity in a geographical region as well as probabilistic forecasting of future earthquakes. We will take this up in Section 3 after first discussing the likelihood function for the ETAS model.

2.2. *(Log-)likelihood.* The log-likelihood function for a general space-time point process on $[0, T] \times \Sigma$ with intensity $\lambda(\cdot)$ and n observed earthquakes $Y = \{(t_i, m_i, x_i, y_i)\}$ and parameter vector $\Theta = (\mu, K, \alpha, c, p, \sigma_x, \sigma_y, \phi)$ is given by the following expression (Daley and Vere-Jones (2003)):

$$l(Y|\Theta) = \sum_{i=1}^{n}\log\big(\lambda(t_i, m_i, x_i, y_i|\Theta)\big) - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^T \lambda(t, x, y|\Theta)\,dt\,dx\,dy. \tag{3}$$

The evaluation of this triple integral for the ETAS model is computationally challenging and can be numerically unstable; thus, several approximations are provided in the literature (Ogata (1998), Schoenberg (2013), Lippiello et al. (2014)). Noting that the functions $\phi(\cdot)$, $h(\cdot)$ and $s(\cdot)$ are probability densities that integrate to 1, the triple integral in equation (3) can be approximated as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty}\lambda(t,m,x,y|\Theta)\,dt\,dx\,dy \approx \mu T + \bar{K}\sum_{i=1}^{n} e^{\alpha(m_i-M_0)}.$$

This approximation assumes that the temporal and spatial domains are infinite, when, in fact, the point process is defined on the finite region $[0,T]\times\Sigma$. In Schoenberg (2013) it was shown that the assumption of an infinite spatial domain has negligible effect on the likelihood, and so we will use this approximation to reduce the computational cost of evaluating the triple integral. However, in Kolev and Ross (2019) it was shown that the infinite time assumption can be inaccurate; thus, we prefer to avoid it. As such, we instead work on the finite temporal region $[0,T]$. Using straightforward integration techniques, it can be to shown that the log-likelihood of the ETAS model, based on the finite time and infinite space assumptions, can be approximated by the following expression:

$$\begin{aligned} l(Y|\Theta) = {} & \sum_{i=1}^{n}\log\bigl(\lambda(t_i,m_i,x_i,y_i|\Theta)\bigr) - \mu T \\ & - K\sum_{i=1}^{n} e^{\alpha(m_i-M_0)}\biggl(1-\frac{c^{p-1}}{(T-t_i+c)^{p-1}}\biggr). \end{aligned} \tag{4}$$

3. Nonparametric estimation of background intensity. We propose estimating $\phi(x,y)$ using a Dirichlet process mixture prior, since this allows the shape of $\phi(x,y)$ to be learned in a nonparametric data driven manner without the need to specify functional forms in advance. Unlike other methods, such as the kernel density estimation of Zhuang, Ogata and Vere-Jones (2002) and Marsan and Lengline (2008), it also allows for fully probabilistic inference for the ETAS model parameters which then can be propagated forward to make probabilistic forecasts about future earthquake occurrences in a manner which incorporates all parameter uncertainty.

The Dirichlet process (DP) (Ferguson (1973), Antoniak (1974)) is a probability distribution over probability distributions and is commonly used as a prior in Bayesian nonparametric modelling. If a probability distribution G has a DP prior, then we write

$$G \sim \mathrm{DP}(\chi, G_0),$$

where G_0 is the base distribution, which defines the expected value of the DP, and $\chi > 0$ is a measure of the variance. The DP is a conjugate prior in the following sense: suppose that $\theta_1,\dots,\theta_n \sim G$ where $G \sim \mathrm{DP}(\chi, G_0)$. Then, the posterior distribution of G is

$$G|\theta_1,\dots,\theta_n \sim \mathrm{DP}\biggl(\chi+n, \frac{\chi G_0 + \sum_{i=1}^{n}\delta_{\theta_i}}{\chi+n}\biggr), \tag{5}$$

where δ_{θ_i} denotes a point mass located at θ_i.

A constructive definition of the DP was given by Sethuraman (1994), who showed that samples from a DP can be written in stick breaking form,

$$G = \sum_{i=1}^{\infty}\pi_i\delta_{\psi_i}, \qquad \psi_k \sim G_0,$$

where $\{\beta_i\}_{i=1}^{\infty} \sim \text{Beta}(1, \chi)$, $\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i)$. This provides a practical method for drawing a sample from a DP, by approximating the stick breaking as a finite (truncated) sum: Combining this with the conjugacy result above, we can sample G from its posterior distribution, given some observed data $\theta_1, \dots, \theta_n$, as

$$G|\theta_1, \dots, \theta_n = \sum_{i=1}^{N} \pi_i \delta_{\psi_i}, \qquad \psi_k \sim \frac{\chi G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\chi + n}, \tag{6}$$

where $\{\beta_i\}_{i=1}^{N} \sim \text{Beta}(1, \chi + n)$ and $\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i)$.

An alternative representation of the DP is based on the Chinese restaurant process (Neal (2000)) which shows that the marginal prior distribution of the samples $\theta_1, \dots, \theta_n$ (with G integrated out) can be written as

$$\theta_i|\theta_1, \dots, \theta_{i-1} \sim \frac{1}{i-1+\chi} \sum_{j=1}^{i-1} \delta_{\theta_i} + \frac{\chi}{i-1+\chi} G_0, \quad \text{where } \theta_i \sim G_0.$$

From the above results we can see that samples from a DP follow a discrete distribution, so it cannot be directly used as a prior for the continuous background ETAS density $\phi(x, y)$. In order to adapt the DP to continuous data, it is common to instead use the DP as a prior distribution for a continuous mixture model (Lo (1984)). Since $\phi(x, y)$ is a two-dimensional spatial distribution, we will model it as a mixture of bivariate Gaussian distributions $N(x, y|\theta)$, where $\theta = (\boldsymbol{\mu}, S)$ with $\boldsymbol{\mu}$ denoting the mean vector and S denotes denoting the precision matrix. For conjugacy we choose G_0 to be the Normal Wishart distribution. This leads to the following model, which we will refer to as DP ETAS:

$$\begin{aligned} (x_i, y_i) &\overset{\text{iid}}{\sim} \phi(x, y), \\ \phi(x, y) &= \int N(x, y|\theta) G(\theta)\, d\theta, \\ G &\sim \text{DP}\big(\chi, \text{NIW}(\boldsymbol{\mu}_0, \kappa_0, v_0, T_0)\big), \end{aligned}$$

where $\boldsymbol{\mu}_0, \kappa_0, v_0, T_0$ are the parameters of the Normal Wishart distribution,

$$G_0(\boldsymbol{\mu}, S) = N\big(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa_0 S)^{-1}\big) Wi(S|v_0, T_0).$$

4. Posterior simulation. Since the posterior distribution of the DP ETAS model parameters is intractable, we will use simulation methods to approximate this posterior instead. It may initially seem feasible to use direct Metropolis–Hastings MCMC for this purpose. However, this is potentially problematic. It is known from Schoenberg (2013) that, even when the background rate $\phi(\cdot)$ is constant, the ETAS likelihood function is multimodal and the components of the parameter vector are highly correlated. Since MCMC methods can also suffer from serious convergence issues when the parameters are correlated, they can inherit the same difficulties, This was confirmed by (Ross (2021)) who showed that a direct approach based on MCMC tended to suffer from low effective sample sizes due to severe correlation in the posterior samples. Since this problem is already present in the simple parametric case with constant $\phi(\cdot)$, it will be even worse in the more nonparametric setting which is substantially more complex

As such, we instead use a reparametrisation of the model, based on the latent variable formulation of Rasmussen (2011), for temporal Hawkes processes that aims to break the parameter correlation in the likelihood function and leads to an efficient Metropolis–Hastings algorithm for posterior sampling. This is an extension of the method proposed by Rasmussen (2013) for the parametric temporal Hawkes process. We now describe this sampling scheme, followed by a discussion of the choice of prior distribution in Section 5.

4.1. *Latent variable formulation.* As discussed in Section 2, the ETAS model can be reinterpreted as a branching process. Suppose that at time t there have been n_t previous events. Then, equation (1) can be interpreted as showing that the local behavior of the process at time t is equivalent to the superposition of $n_t + 1$ different Poisson processes. The first is a time-homogenous Poisson process with intensity $\mu\phi(x, y)$, while each event at time t_j triggers an inhomogeneous Poisson process with intensity

$$\lambda_j(t, x, y) = Ke^{\alpha(m_j - M_0)} \frac{(p-1)c^{p-1}}{(t - t_j + c)^p} \frac{1}{2\pi\sigma_x\sigma_y} \times \exp\left(\frac{-1}{2}\left[\frac{(x - x_j)^2}{\sigma_x^2} + \frac{(y - y_j)^2}{\sigma_y^2}\right]\right). \tag{7}$$

Now, consider an event that occurs at time t_i, when there have been n_{t_i} previous events. Based on standard results about the superposition of Poisson processes (Daley and Vere-Jones (2003)), we can interpret event t_i as having been generated by a single one of these $n_{t_i} + 1$ processes. For an earthquake catalog that contains n events in total, we introduce the latent branching variables $B = \{B_1, \dots, B_n\}$ where $B_i \in \{0, 1, \dots, n_{t_i}\}$ indexes the process which generated t_i,

$$B_i \sim \begin{cases} 0 & \text{if } t_i \text{ was produced by the background process,} \\ j & \text{if } t_i \text{ was triggered by the previous earthquake at time } t_j. \end{cases}$$

Conditional on knowing B, we can partition the earthquakes into $n + 1$ sets $S_0, \dots, S_n$, where

$$S_j = \{t_i; B_i = j\}, \quad 0 \le j < n,$$

so that S_0 is the set of immigrant events, which were not triggered by previous earthquakes, and S_j is the set of direct aftershocks triggered by the earthquake at time t_j. It is clear that these sets are mutually exclusive and that their union contains all the earthquakes in the catalog. Additionally, we can see that the earthquakes in set S_0 are generated by an inhomogeneous Poisson process with intensity $\mu(x, y)$, while the events in each set S_j for $j > 0$ are generated by a single inhomogeneous Poisson process with intensity given by equation (7). The ETAS likelihood function from equation (4) can hence be factorised (conditional on the latent branching variables) into a product of the following four blocks:

$$\begin{aligned} p(Y|\Theta, B) = {} & \left(e^{-\mu T} \prod_{t_i \in S_0} \phi(x_i, y_i)\right) \\ & \times \left(\prod_{j=1}^{n} e^{-(Ke^{\alpha(m_j - M_0)}[1 - \frac{c^{p-1}}{(T - t_i + c)^{p-1}}])} \{Ke^{\alpha(m_j - M_0)}\}^{|S_j|}\right) \\ & \times \left(\prod_{j=1}^{n} \prod_{t_i \in S_j} \frac{(p-1)c^{p-1}}{(t_i - t_j + c)^p}\right)\left(\prod_{j=1}^{n} \prod_{t_i \in S_j} \frac{1}{2\pi\sigma_x\sigma_y} e^{\frac{-1}{2}[\frac{(x_i - x_j)^2}{\sigma_x^2} + \frac{(y_i - y_j)^2}{\sigma_y^2}]}\right). \end{aligned} \tag{8}$$

From this factorisation we can see that μ and $\phi(x, y)$ are independent of the other model parameters in the likelihood and so will be independent in the posterior assuming prior independence. This latent variable formulation thus breaks the dependency, which would have made the DP part of the model difficult to estimate, and further weakens the dependence between $\{c, p\}$, $\{K, \alpha\}$ and $\{\sigma_x, \sigma_y\}$ which will allow for more efficient MCMC sampling.

Based on this idea, we propose a Metropolis-within-Gibbs sampler which samples the parameters in the following condtionally independent blocks: $\{B\}$, $\{\phi(\cdot)\}$, $\{\mu\}$, $\{K, \alpha\}$, $\{c, p\}$, $\{\sigma_x, \sigma_y\}$. We now discuss each of these steps in turn. The choice of prior distributions $\pi(\cdot)$ will be discussed in Section 5.

4.1.1. *Updating B.* Note that each B_i can take values only in the discrete set $\{0, 1, \dots, i-1\}$, that is, each earthquake can be triggered only by either a previous earthquake or the background process. Assuming a discrete uniform prior on each B_i, the probability of an event at t_i being caused by any of these i processes is simply the proportion of the overall intensity that can be attributed to that process, that is,

$$p\big(B_i^{(k+1)} = j|Y, \Theta^{(k)}\big) = \begin{cases} \dfrac{\mu\phi(x_i, y_i)}{\lambda(t_i, m_i, x_i, y_i|\Theta^{(k)})} & \text{for } j = 0, \\ \dfrac{k(m_i)h(t_i - t_j)s(x_i - x_j, y_i - y_j)}{\lambda(t_i, m_i, x_i, y_i|\Theta^{(k)})} & \text{for } j \neq 0. \end{cases} \tag{9}$$

Each B_i can thus be drawn independently from the discrete uniform distribution on $\{0, \dots, i-1\}$ with weights given by the above (Kolev and Ross (2019)). Note that, unlike the MCMC sampler used in Rasmussen (2013) for Hawkes processes, this samples $B^{(k+1)}$ exactly from its conditional posterior, which drastically improves computational efficiency.

4.1.2. *Updating $\phi(x, y)$.* Recall from Section 2.2 that

$$\phi(x, y) = \int N(x, y|\theta)\, d\theta, \quad \theta \sim G, G \sim \mathrm{DP}(\chi, G_0),$$

where $\phi(x, y)$ is the spatial density followed by the $|S_0|$ immigrant earthquakes that are assigned to the background process based on the current branching structure. In order to simulate a value of $\phi(\cdot)$ from its conditional posterior, we first simulate values of $\theta_i, i \in \{1, 2, \dots, |S_0|\}$ from their posterior distributions, given the earthquakes which are assigned to the background process, using the usual Chinese restaurant process sampler. Given these values, the conditional posterior for G is (from equation (5))

$$G|\theta_1, \dots, \theta_{|S_0|} \sim \mathrm{DP}\bigg(\chi + n, \frac{\chi G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\chi + n}\bigg).$$

We can thus sample a value of G from its posterior, using truncated stick breaking as in equation (6), that is,

$$G = \sum_{i=1}^N \pi_i \delta_{\psi_i}$$

which deterministically gives a realisation of $\phi(\cdot)$. Note that the reason why we need to sample a realisation of $\phi(x, y)$ (and hence G) rather than working only with the θ_i samples is that we need to have a realisation of $\phi(x, y)$ to evaluate the branching posterior in equation (9).

4.1.3. *Updating μ.* From equation (8), μ depends only on the number of events in the background process S_0, and so

$$p(\mu|Y, \Theta, B) \propto \pi(\mu)e^{-\mu T}\mu^{|S_0|}.$$

This is equivalent to estimating the intensity μ of a homogeneous Poisson process on $[0, T]$ with event times S_0. In this case the Gamma distribution is the conjugate prior: $\pi(\mu) = \mathrm{Ga}(\alpha_\mu, \beta_\mu)$. With this prior the posterior distribution is then $p(\mu|Y, \Theta, B) = \mathrm{Ga}(\alpha_\mu + |S_0|, \beta_\mu + T)$ which can be sampled from directly.

4.1.4. *Updating K and α.* Similar to the update for μ, we can sample new values of K and α from $p(K, \alpha | Y, \Theta, B)$. Based on equation (8), we have that

$$p(\alpha, K | Y, \Theta, B) \propto \pi(K, \alpha) \prod_{j=1}^{n} \left(e^{-K e^{\alpha(m_j - M_0)}(1 - \frac{c^{p-1}}{(t_n - t_i + c)^{p-1}})} \{K e^{\alpha(m_j - M_0)}\}^{|S_j|}\right).$$

Although there is no conjugate prior in this case, it is straightforward to use, for example, Metropolis–Hastings with random walk proposals to draw a sample from this distribution. Alternatively, a more sophisticated scheme such as Hamiltonian Monte Carlo could also be used, although we found that Metropolis–Hastings was sufficient.

4.1.5. *Updating c and p.* Again, based on equation (8), the conditional posterior distribution of σ_x and σ_y is

$$p(c, p | Y, \Theta, B) \propto \pi(c, p) \prod_{j=1}^{n} \left(e^{-K e^{\alpha(m_j - M_0)}(1 - \frac{c^{p-1}}{(t_n - t_i + c)^{p-1}})} \prod_{t_i \in S_j} \frac{(p-1)c^{p-1}}{(t_i - t_j + c)^p}\right).$$

This can again be sampled from using Metropolis–Hastings with random walk proposals.

4.1.6. *Updating σ_x and σ_y.* Based on the infinite space and finite time approximation to the likelihood, which was discussed in Section 2.2, the conditional distribution for σ_x and σ_y is

$$p(\sigma_x, \sigma_y | \mathscr{H}_t, \Theta, B) \propto \pi(\sigma_x, \sigma_y) \prod_{j=1}^{n} \prod_{t_i \in S_j} \frac{1}{2\pi \sigma_x \sigma_y} e^{\frac{-1}{2}[\frac{(x_i - x_j)^2}{\sigma_x^2} + \frac{(y_i - y_j)^2}{\sigma_y^2}]}.$$

This posterior is identical to that of estimating the diagonal elements of a spherical bivariate Gaussian covariance matrix. As such, if $\pi(\sigma_x, \sigma_y) = \pi(\sigma_x)\pi(\sigma_y)$ are chosen to be independent conjugate Inverse-Gamma priors, then this distribution can be sampled from exactly.

5. Prior choice and implementation details. Completing the Bayesian model requires specification of a prior distribution for all model parameters. In cases where there is strong existing beliefs about particular model parameters, this can be directly incorporated into the prior distribution. However, often, this will not be the case. Indeed, the specification of prior distributions can sometimes be difficult, particularly for practitioners without a background in Bayesian statistics. As such, we give the following recommendations for default prior distributions, which we found to be appropriate to most catalogs we studied. These priors are all noninformative or weakly informative and hence do not encode existing information about the model parameters. Inference under these will hence be robust to small changes in prior specification.

We first recommend rescaling the spatial coordinates (x, y) of the earthquakes in the catalog to have mean 0 and standard deviation 1. This allows a default prior to be specified on this range without needing to worry about the precise spatial scaling of the catalog. Then, as mentioned in the above discussion, it is possible to choose priors for μ, σ_x and σ_y that lead to conditionally conjugate inference that allows exact sampling from their conditional posteriors. We recommend the following noninformative choices:

$$\pi(\mu) = \text{Gamma}(0.01, 0.01), \qquad \pi(\sigma_x) = \pi(\sigma_y) = \text{Inverse-Gamma}(0.01, 0.01).$$

For the β parameter of the Gutenberg–Richter law governing the magnitude distribution, we use a noninformative conjugate Gamma(0.01, 0.01) prior. There is, however, no conjugate prior for the parameters (c, p) of the temporal triggering kernel. In this case we recommend Uniform priors

$$\pi(c) = \text{Uniform}(0, 10), \qquad \pi(p) = \text{Uniform}(1, 10)$$

with the prior on p chosen to restrict it to be greater than 1, since smaller values are not physically meaningful. We note that these priors are weakly informative in the sense that they rule out very large values of c. This is required since Holschneider et al. (2012) showed in a Bayesian study of the Omori law that c and p are only weakly identifiable, in the sense that there are many different combinations (c, p) which give essentially the same value of the likelihood.

For K, α we choose a Uniform prior on a parameter range that avoids supercriticality. The phenomena of super-criticality was first noted by Helmstetter and Sornette (2002) who pointed out that particular choices of k and α result in catalogs that have an infinite number of expected earthquakes. In order for the catalog to be finite, the average number of direct aftershocks produced by a "typical" earthquake must be less than 1. By integrating over the Gutenberg–Richter magnitude distribution, this holds if

$$\beta < \alpha \quad \text{and} \quad \frac{\bar{K}\beta}{\beta - \alpha} < 1.$$

The prior we use is Uniform on the simplex of values for which this condition holds. The final choice is the specification of the DP(α, G_0) prior for $\phi()$. For the base measure G_0, we use a conjugate Normal Wishart prior $G_0(\boldsymbol{\mu}, S) = N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa_0\Lambda)^{-1}))Wi(S|\upsilon_0, T_0)$ where, as discussed in Section 5, $\boldsymbol{\mu}$ and $\mathbf{S}$ are the cluster means and preceision matrices. We make the following choice for the prior parameters: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\kappa_0 = 2$, $\upsilon_0 = 2$ and $T_0 = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Note that these are fairly noninformative choices for the spatial data that has been rescaled to have mean 0 and standard deviation 1. They are the recommended default values used in the *Dirichlet process* R package (Ross and Markwick (2018)), which we used to perform the DP posterior sampling.

Finally, in the above discussion of the MCMC sampler we stated that random walk Metropolis–Hastings was used to update some blocks of parameters. For these we used a Normal proposal distribution with standard deviation of 0.1.

In Section 7 we study four different real earthquake catalogs taken from different geographical regions. For each of these catalogs, we use the above noninformative prior specifications and do not alter the prior for each particular catalog. Since this prior choice is shown to work well for several different catalog, we are comfortable recommending it as a default choice.

6. Forecasting. An important application of earthquake models is forecasting the occurrence of future earthquakes based on the previous seismicity. A key advantage of our fully Bayesian model is that it allows the easy construction of a forecast distribution which takes into account all uncertainty in the ETAS parameter estimation. Suppose that we have observed n earthquakes $Y = \{(t_i, m_i, x_i, y_i)\}$ on the time interval $[0, T]$ and we wish to assess how well the estimated ETAS model has performed when forecasting on a future interval $[T, \tilde{T}]$. Suppose that $\tilde{n}$ earthquakes $\tilde{Y} = \{(\tilde{t}_i, \tilde{m}_i, \tilde{x}_i, \tilde{y}_i)\}$ are observed on $[T, \tilde{T}]$. The (out-of-sample) predictive likelihood is

$$p(\tilde{Y}|Y) = \int p(\tilde{Y}|\Theta)p(\Theta|Y)\,d\Theta \approx \frac{1}{M}\sum_{r=1}^{M} p(\tilde{Y}|\Theta^{(r)}),$$

where $\Theta^{(r)}$ denotes the parameter values sampled from $p(\Theta|Y)$ using the above MCMC prodcedure. Using standard point process theory, it can be shown that

$$
\begin{aligned}
p(\tilde{Y}|\Theta) = {} & \sum_{i=1}^{\tilde{n}} \log \lambda(\tilde{t}_i, \tilde{x}_i, \tilde{y}_i|\Theta) - \mu(\tilde{T} - T) \\
& - \sum_{i=1}^{\tilde{n}} K e^{\alpha(M_0 - \tilde{m}_i)} \bigl[1 - c^{p-1}(\tilde{T} - \tilde{t}_i + c)^{1-p}\bigr] \\
& - \sum_{i=1}^{n} K e^{\alpha(M_0 - m_i)} c^{p-1} \bigl[(T - t_i + c)^{1-p} - (\tilde{T} - t_i + c)^{1-p}\bigr].
\end{aligned}
\tag{10}
$$

Next, suppose that rather than assessing how well the model performs at predicting the earthquake $\tilde{Y}$, we instead wish to produce a forecast for seismic behaviour on $[T, \tilde{T}]$. Typically, this will involve forecasting some particular quantity, such as the total number of earthquakes which will occur, or the number that will occur within some spatial subregion or the probability of there being a large earthquake with magnitude greater than some threshold $\tilde{M}$. Let $\tilde{N}$ denote the particular quantity which is being forecast. Its posterior predictive distribution is

$$
p(\tilde{N}|Y) = \int p(\tilde{N}|\Theta) p(\Theta|Y) \, d\Theta \approx \frac{1}{M} \sum_{r=1}^{M} p(\tilde{N}|\Theta^{(r)}).
\tag{11}
$$

Unlike the predictive likelihood in equation (10), the distribution $p(\tilde{N}|\Theta^{(r)})$ will generally not be easy to compute for most forecasted quantities. However, it is straightforward to approximate it via simulation. For each $\Theta^{(r)}$, we simulate a set of events on the interval $[T, \tilde{T}]$ using $\Theta^{(r)}$ as the ETAS parameter. The value of the statistic $\tilde{N}^{(r)}$ is then computed using these simulated events, and the set of these values gives an approximation to the forecast distribution $p(\tilde{N}|Y)$. To give a concrete example, suppose we wish to determine the probability that an earthquake with magnitude greater than $\tilde{M}$ will occur in $[T, \tilde{T}]$. For each set of simulated events r with parameter $\Theta^{(r)}$, let $E_r = 1$, if these events contains an earthquake with magnitude greater than $\tilde{M}$, and $E_r = 0$ otherwise. A point estimate of the probability of obtaining an earthquake greater than magnitude M is then given by $1/M \sum_{r=1}^{M} E_r$, and credible intervals can easily be constructed based on the simulated E_r values.

Implementing this procedure requires an efficient way to simulate events from an ETAS model on $[T, \tilde{T}]$, given the earthquakes that have already occurred on $[0, T]$. It is common in the ETAS literature to use the methodology of thinning for this purpose (Kumazawa and Ogata (2014), Lombardi (2015), Lombardi (2017)) which is essentially a type of rejection sampling. However, rejection sampling is in this context is inefficient and unnecessary. Instead, we suggest an algorithm, based on exploiting the cluster process representation for Hawkes processes (Hawkes and Oakes (1974)), which simulates events directly without any need for rejections. This method is extremely fast, and we found it typically take less than a second to simulate several thousand events:

1. Sample I mainshocks from a Poisson$(\mu(\tilde{T} - T))$ distribution. For each mainshock, sample its temporal location independently from a Uniform$(T, \tilde{T})$ distribution and its magnitude from the Gutenberg–Richter law with parameter β.
2. For each earthquake (t_i, m_i) in $[0, T]$, sample its direct offspring on $[T, \tilde{T}]$ by first simulating its total number O_i of direct offspring on $[t_i, \infty)$ from a Poisson$(K e^{\alpha(m_i - M_0)})$ distribution. For each such offspring $j \in (1, 2, \dots, O_i)$, let z_j be an independent sample from the modified Omori distribution $h(z) = (p-1)c^{p-1}\frac{1}{(z+c)^p}$ on $[0, \infty]$. The offspring

location is then given by $t_i + z_j$. If this falls outside the range $[T, \tilde{T}]$, then it is discarded. As such, the total number of offspring produced by t_i on $[T, \tilde{T}]$ will be equal to or less than O_i. Note that this procedure works correctly since the Poisson nature of the point process means that the direct offspring that (t_i, m_i) will produce on $[T, \tilde{T}]$ is independent of the direct offspring that it has already produced on $[0, T]$. Again, the magnitude of each offspring can be simulated independently from the Gutenberg–Richter law.

3. Let E be the combined set of mainshock and offspring on $[T, \tilde{T}]$ that were generated in the above two steps. For each earthquake $(t_i, m_i) \in E$, simulate its number of direct offspring O_i from a Poisson($Ke^{\alpha(m_i - M_0)}$) distribution. The locations of these offspring are then generated in the same manner as in the previous step. For each offspring $j \in (1, 2, \dots, O_i)$, simulate its offset z_j from the modified Omori distribution on $[0, \infty]$. Its location is then given by $t_i + z_j$, and it is discarded if it falls outside $[T, \tilde{T}]$.
4. Repeat the previous step to find the offspring of the generated offspring, and repeat until no further offspring are generated.

The above procedure repeatedly requires efficient simulation of values from the modified Omori law $h(z) = (p-1)c^{p-1}\frac{1}{(z+c)^p}$. This can be obtained by inverse-transform sampling. It is straightforward to show that if u is a sample from a Uniform(0, 1) distribution and $z = c(1-u)^{1/(1-p)} - c$, then z is a sample from the modified Omori law as required.

The above procedure produces earthquakes from the temporal ETAS model. The next step is to simulate their spatial locations. For the mainshocks produced in Step 1 above, their location can be simulated directly from $\phi(x, y)$ since this is a bivariate Gaussian mixture model. For each triggered earthquake at time t_j, let (t_i, x_i, y_i) be the earthquake which produced it as an offspring. Its spatial offset (x_j^*, y_j^*) is sampled independently from the spatial kernel $s(x, y)$, and its location is then $x_j = x_i + x_j^*$, $y_j = y_i + y_j^*$.

7. Application to earthquake catalogs. We now assess the performance of the DP ETAS model when forecasting future seismicity using real earthquake catalogs taken from four different geographical regions. The DP ETAS model is implemented in the *bayesianETAS* R package available at https://cran.r-project.org/web/packages/bayesianETAS/index.html. To show the importance of an inhomogeneous background rate, we compare performance to a version of the ETAS model where the background rate $\phi(x, y) \propto 1$ is constant. We will also compare performance to a frequentist version of the space-time ETAS model which uses kernel density estimation (KDE) to estimate the background rate. For this purpose we follow the approach of Zhuang, Ogata and Vere-Jones (2002) who perform simultaneous estimation of $\phi(x, y)$ and the ETAS parameters using stochastic declustering (SD). This approach has been widely used in the literature with other examples being Zhuang and Mateu (2019) and Reinhart and Greenhouse (2018). It is essentially an EM algorithm which alternates between estimating the ETAS parameters, estimating the probabilities p_i of each observation being a mainshock and computing a weighted kernel estimate of $\phi(x, y)$ from these using a bivariate Gaussian kernel with diagonal covariance matrix.

A potential drawback of kernel density estimation is that using fixed bandwidths (i.e., diagonal covariance matrix elements) can result in over/under-smoothing in some spatial regions, so we follow Mohler et al. (2011) and Zhuang, Ogata and Vere-Jones (2002) and use adaptive bandwidths based on k-nearest neighbours. Specifically, for each observation (x_i, y_i) we take the bandwidth of $K(x - x_i, y - y_i)$ to be equal to the distance to the 100th nearest neighbour of (x_i, y_i). As well as this approach using stochastic declustering, we also compare to an alternative frequentist approach which simply estimates $\phi(x, y)$ using the same kernel and the entire earthquake catalog without declustering.

We now describe the four earthquake catalogs, which we will study. For each one we will fit the models on an initial training set representing the historical seismicity and then

TABLE 1
Out-of-sample forecast log-likelihoods for the DP ETAS, KDE, stochastic declustering and constant background models. Large (less negative) values of the forecast log-likelihood indicate superior performance, and the best performing model for each catalog is shown in bold. *Note that the high values of the likelihood for the Central Italy catalog are due to the very large number of events compared to the other catalogs*

	DP	KDE	SD	Constant
Vroacea	**−317**	−336	−327	−334
Zakynthos	−98	**−96**	**−96**	−98
Friuli	**−207**	−209	−210	−211
Central Italy	**5591**	5560	5578	4787

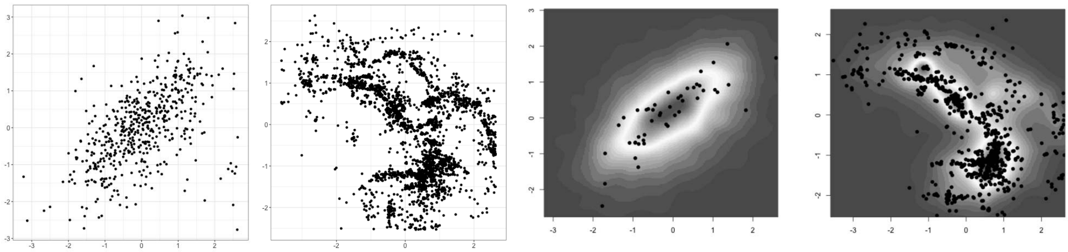
assess performance by out-of-sample forecasting on a test set consisting of the most recent earthquakes.

7.1. *Vrancea*, *Romania*. Vrancea is an area in Romania that has a strong seismic influence on Southeastern Europe. On 4/3/1977 a large magnitude 7.2 earthquake occurred which caused substantial destruction and human loss in both Bulgaria and Romania. We analysed the earthquake catalog from 01/01/1974 to 01/01/2019 that covers the spatial region 46°, 45°18′N and 27°, 26°E with a magnitude of completeness of $M_0 = 2.5$. The data was split into a training set over the period 01/01/1975–01/01/2014 with 529 events and a test set over the period 01/01/2014–01/01/2019 with comprises 46 events. The data was obtained from the United States Geological Survey (USGS) catalog (http://earthquake.usgs.gov/).

7.2. *Zakynthos and Kefalonia*, *Greece*. Zakynthos and Kefalonia are subject to prolonged seismic activity. The area of interest spans 38°33.54′, 47°14.34′N and 21°36.96′, 19°39.96′E. The most important event in the region is the 6.8 M Ionian earthquake that occurred on 12/08/1953. Including data from this period is very challenging due to the high magnitude of catalog completeness caused by poor quality seismic detection equipment. For this reason we focused our study on a more recent time period from (1/1/2069 to 1/1/2019) and chose a magnitude of completeness of $M_0 = 4.5$ to ensure consistency throughout the catalog. The data was split into a training set 01/01/1969–01/01/2018 with 343 events and a test set 01/01/2018–01/01/2019 that comprises 109 events. The data was obtained from the United States Geological Survey (USGS) catalog (http://earthquake.usgs.gov/).

7.3. *Friuli*, *Italy*. Friuli is an area in Italy that is primarily known for the 6.5 M earthquake that occurred on 06/05/1976, followed by multiple aftershocks with considerably large magnitudes. We based our study on the earthquake occurrence from 01/01/1975 to 01/01/2019 that covers the area within 46°36′, 46°N and 12°18′, 13°30′E with minimum magnitude of $M_0 = 3$. For inferential purposes we split the data into a train set 01/01/1975–01/01/2004 with 310 events and a test set 01/01/2004–01/01/2019 that comprises 20 events. The data was obtained from the United States Geological Survey (USGS) catalog (http://earthquake.usgs.gov/).

7.4. *Central Italy*. In 2006, Central Italy suffered a particularly damaging magnitude 6.2 earthquake that caused the death of nearly 300 people. Although the type of point process models discussed in this paper are not appropriate for predicting the occurrence of individual earthquakes, they are particularly useful for forecasting patterns in aftershock sequences. To this end, we obtained earthquake data from the Italian National Institute of Geophysics and Volcanology (Instituto Nazionale Di Geofisica e Vulcanalogia http://www.ingv.it) from



(a) Vrancea (b) Central (c) Vrancea (d) Central

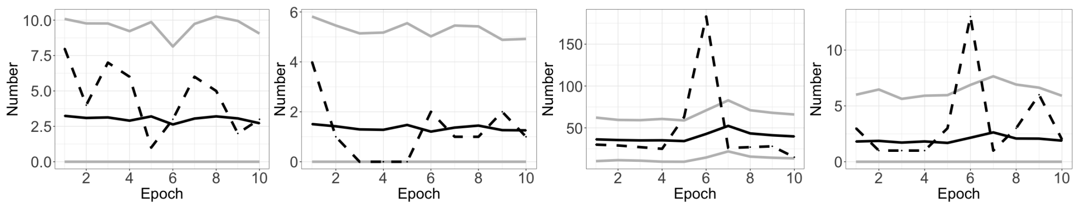
FIG. 1. *The first two plots shows the locations of the historical* (*i.e., training set*) *earthquakes in the Vrancea and Central Italy catalogs. The second two show the DP ETAS forecast for the most probable locations of magnitude* 3.5 *or greater earthquakes during the five-year forecast period. The true locations of the magnitude* > 3.5 *earthquakes during this period are shown as black dots.*

01/04/1999 to 01/04/2019 that spans Italy within 35°, 49°N and 5°, 20°E with magnitude of completeness taken to be $M_0 = 3$. Then, we split the data into a training set from 01/04/1999 to 01/04/2014 (4669 events), which was used to estimate the model parameters, followed by a test set from 01/04/2014 to 01/04/2019 (2171 events) used to measure performance.

7.5. *Results*. For each of the four catalogs, the parameters of the ETAS models were estimated using only the earthquakes in the training set. For this purpose, 20,000 samples were drawn from the DP ETAS posterior, and the the predictive log-likelihood in equation (10) was then computed on the test sets. For the frequentist approaches the predictive log-likelihoods were computed using the point estimates of the model parameters. Table 1 shows the resulting forecast performance for the DP, SD, KDE and constant background models. As expected, the constant background model generally gives a poor fit which demonstrates the importance of modelling spatial inhomogeneity in $\phi(x, y)$. Unlike previous Bayesian approaches to the ETAS model, our method is able to capture this inhomogeneity and hence provides more accurate earthquake forecasts. The comparison to KDE and SD is less drastic, with both methods giving very similar performance on the Zakythos and Friuli catalogs, with the Bayesian DP method being superior on Vrancea and Central Italy. Of course, we do not suggest that Bayesian methods will always outperform frequentist methods on every earthquake catalog, and in some cases the choice between the two will depend on the philosophical preferences of the modeller.

8. Detailed study. One of the primary uses of ETAS-type models is to produce forecasts and hazard maps which can quantify the degree of seismic risk over a given time period in a geographical region. In this section we will study two particular regions and show how various types of forecasts and hazard maps can be generated from the DP ETAS model. Using the time periods and spatial boundaries described in the previous section, the two regions we study are Vrancea and Central Italy. The first two plots of Figure 1 shows the spatial locations of all the earthquakes that occurred during the training set period in both regions. It can be seen that the regions are very different, with Vrancea having a fairly simple spatial distribution, while Central Italy is substantially more complex.

The Bayesian DP ETAS model was fit to the training set catalogs taken from both regions, with 20,000 parameters drawn from the posterior distribution using MCMC. A spatial forecast density can then be produced by simulating from the fitted model, as discussed in Section 6. For each region the test set consists of the most recent five years of data and the second two plots of Figure 1 shows the spatial forecast distribution for the most likely locations of magnitude ≥ 3.5 earthquakes during this period. The true locations of the magnitude



(a) Vranccea, $M > 3.5$ (b) Vranccea, $M > 4.5$ (c) Central, $M > 3.5$ (d) Central, $M > 4.5$

FIG. 2. *Forecast distribution for the number of earthquakes above a magnitude threshold of* 3.5 *and* 4.5 *in each six-month Epoch. The forecast is the mean of the posterior forecast distribution* (*solid black line*) *plus/minus three standard deviations* (*grey lines*). *The true number of earthquakes in each Epoch is shown in as a dotted line.*

≥ 3.5 earthquakes in the test set period are shown as black dots. It can be seen that the DP ETAS model has successfully learned the spatial inhomogeneity in $\phi(x, y)$.

This forecast for the five-year test set period was produced immediately at the end of the training set period. However, in practice, it is desirable to update forecasts in response to recent earthquake activity. This is because the occurrence of large earthquakes will change the future probabilities of further aftershocks occurring nearby. To study this, we break the five-year test set down into 10 different epochs, each consisting of *six* months (so that Epoch 1 is the first six months of the test set period, Epoch 2 is the next six months, and so on). For each epoch, we produce a forecast, which uses the ETAS parameters estimated on the training set but which takes into account triggering from both the training set earthquakes and from the earthquakes that occurred during previous epochs. Figure 2 shows the model forecasts for the number of earthquakes with magnitude greater than 3.5 and 4.5 that were predicted to occur during each six-month epoch. It can be seen that the forecast distribution is very close to the true number of earthquakes that occurred in almost all epochs. The exception is Epoch 6 of the Central Italy catalog, where the model drastically underestimates the number of earthquakes that occurred.

A closer study of the Central Italy catalog shows that a number of large magnitude earthquakes occurred during the test set period which is responsible for the large number of aftershocks in Epoch 6. First, a magnitude 6.0 earthquake occurred during Epoch 5. This led to a substantial number of aftershocks which also triggered nearby large earthquakes with magnitudes 6.5, 5.9 and 5.5 during Epoch 6. The probability of getting four large earthquakes so close together is extremely small. Indeed, during the 15-year training set period, there were only five earthquakes with magnitude ≥ 5.5 in the entire Central Italy region. Since this earthquake pattern was so unlikely, it is assigned a low probability by the ETAS model which explains why the number of earthquakes in Epoch 6 falls far outside the credible interval. Nonetheless, it can be seen that the forecasts made by the DP ETAS model do respond to the events in Epoch 6, and the number of earthquakes forecast for Epoch 7 is substantially higher as a result.

To investigate this further, we also tried forecasting the future seismicity month-by-month rather than in six-month blocks. The model forecasts for each of the 60 months in the test set are shown in Figure 3. It can clearly be seen that the occurrence of the successive large magnitude earthquakes causes the forecasted number to sharply increase in the following month. Although the true seismicity in the following month drops substantially, it is still higher than average but falls within the prediction interval due to the higher number of earthquakes that are expected under the model.

As well as forecasting the number of earthquakes in each epoch, the DP ETAS model also produces a forecast for their spatial locations. The top row of Figure 4 shows the true spatial location of the earthquakes in the (six-month) test set Epochs 5, 6 and 7 which correspond

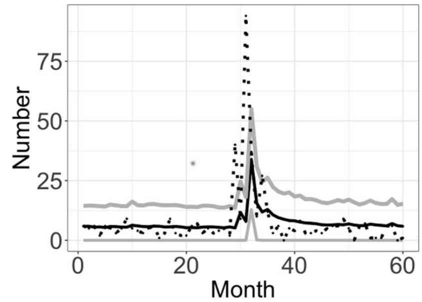
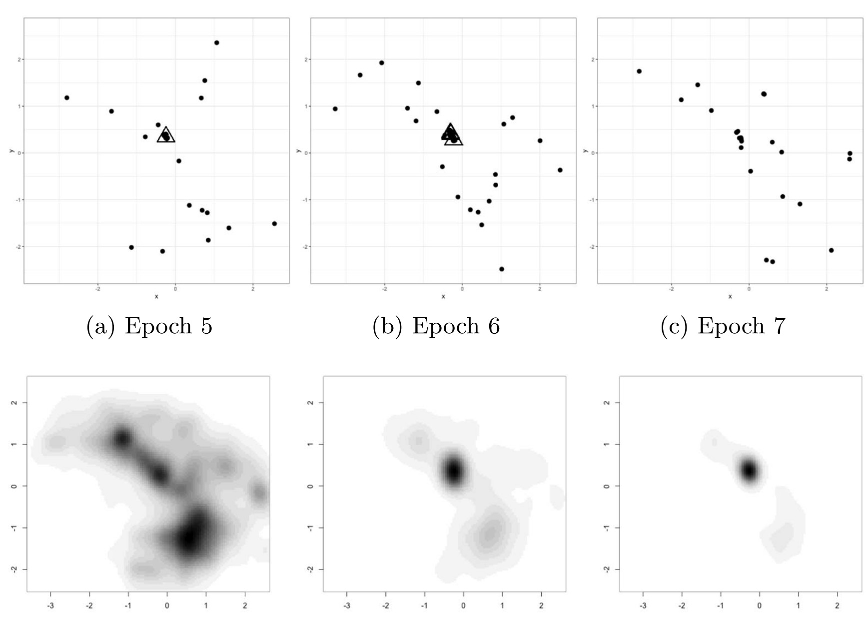


FIG. 3. *Forecast distribution for the number of earthquakes above magnitude* 3.5 *during each test set month in Central Italy. The forecast is the mean of the posterior forecast distribution* (*solid black line*) *plus/minus three standard deviations* (*grey lines*). *The true number of earthquakes in each month is shown as a dotted line.*

to the time periods surrounding the occurrence of the large magnitude events. The bottom row shows the forecast distribution of the locations of earthquakes with magnitude > 3.5 where the forecast for each epoch again takes into account the earthquakes that occurred in all previous epochs. It can be seen that the forecast distribution for Epoch 5 roughly follows the general spatial distribution of earthquakes that can be observed in Figure 1 and is largely determined by the background intensity $\phi(x, y)$. However, due to the large earthquake that occurred during Epoch 5, the forecast distribution for Epoch 6 becomes concentrated around this event, due to the expectation of local aftershock clustering, which does indeed occur. Similarly, the forecast distribution for Epoch 7 is even more concentrated around this spatial region due to the multiple additional large earthquakes that occurred nearby during Epoch 6. This shows that the model is successfully adapting to local seismic activity, as new data becomes available, and is able to adapt and update the forecast distribution in response.



(a) Epoch 5 (b) Epoch 6 (c) Epoch 7

(d) Epoch 5 (e) Epoch 6 (f) Epoch 7

FIG. 4. *Top row*: *True location of the earthquakes with magnitude* > 3.5 *in each epoch for the Central catalog. Large magnitude* > 5.5 *earthquakes are shown as triangles. Bottom row*: *DP ETAS forecast distribution for the location of magnitude* > 3.5 *earthquakes based on all previous seismicity.*

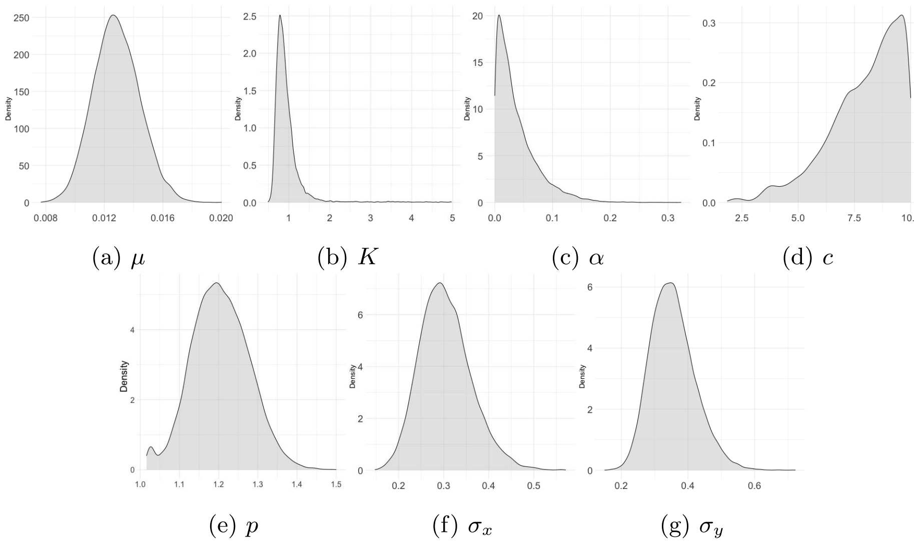


FIG. 5. *Posterior density of the DP ETAS model parameters on the Vrancea earthquake catalog.*

In general, comparing Figures 1 and 4 shows that the background rate $\phi(x, y)$ makes a large contribution to the forecast density during normal circumstances, since this controls the average long-run rate of seismicity in each area. However, there is substantial modulation of this intensity based on aftershock triggering from the more recent earthquakes in the catalog, particularly when large magnitude events occur. This suggests that the short-term prediction is neither solely driven by the Dirichlet process nor by recent seismic activity but is, instead, a combination of both.

Finally, Figure 5 shows the posterior distribution of the ETAS model parameters on the Vrancea catalog, with the Central Italy posteriors being similar. It can be seen that the posterior distributions for μ, K and α are relatively narrow, indicating that there is little uncertainty about these parameters. The posterior for c and p is substantially wider since these parameters are only weakly identifiable, as has previously been pointed out by Holschneider et al. (2012) and Ross (2021). Finally, Figure 6 shows the MCMC trace-plots for each of the model parameters, demonstrating that the Markov chain is mixing well.

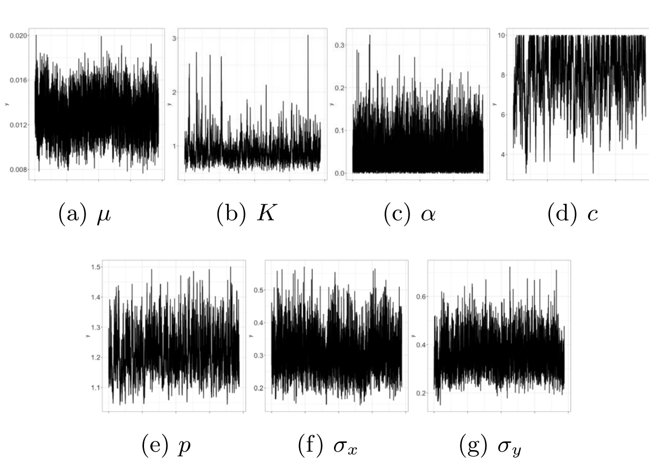


FIG. 6. *MCMC trace-plots of the* 20,000 *posterior samples simulated from the DP ETAS posterior on the Vrancea earthquake catalog.*

9. Conclusions. Traditional frequentist methods commonly used in seismic forecasting typically assume that model parameters are known exactly which can result in forecasts which are overly confident. Further, they typically use heuristic kernel estimates of the background spatial intensity rather than being based on full probability models. Alternative approaches to earthquake modelling that use Bayesian method are becoming more common in seismology. Despite being one of the most popular forecasting models, the ETAS framework has not yet received a fully Bayesian treatment and particularly scant attention has been paid to the estimation of the inhomogeneous background rate. As our experiments have shown, having an accurate estimate of this rate is crucial to providing spatially sensitive earthquake forecasts. In this work we have introduced a new nonparametric model for this background rate which gives the first fully Bayesian ETAS-based forecasting scheme. We have demonstrated an efficient routine for sampling from the posterior of the resulting model and showed how this model can be used to produce informative hazard maps and forecast distributions for geographical regions of interest based on the Bayesian forecasting distribution.

Acknowledgements. We thank the anonymous reviewers who made several helpful comments which improved the manuscript. We also thank Miguel de Carvalho and Finn Lindgren, who gave us very useful feedback on an early manuscript draft.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric Bayesian forecasting of SpatioTemporal earthquake occurrences" (DOI: 10.1214/21-AOAS1554SUPP; .zip). Full code implementing the methods in this paper is available in Ross and Kolev (2022).

REFERENCES

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

CHIODI, M. and ADELFIO, G. (2017). Mixed non-parametric and parametric estimation techniques in R package etasFLP for earthquakes' description. *J. Stat. Softw.* **76** 1–29.

CHORNOBOY, E. S., SCHRAMM, L. P. and KARR, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biol. Cybernet.* **59** 265–275. MR0961117 https://doi.org/10.1007/BF00332915

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

EBRAHIMIAN, H. and JALAYER, F. (2017). Robust seismicity forecasting based on Bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models. *Sci. Rep.* **7** 1–15.

EBRAHIMIAN, H., JALAYER, F., ASPRONE, D., LOMBARDI, A. M., MARZOCCHI, W., PROTA, A. and MANFREDI, G. (2014). Adaptive daily forecasting of seismic aftershock hazard. *Bull. Seismol. Soc. Amer.* **104** 145–161.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FOX, E. W., SCHOENBERG, F. P. and GORDON, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.* **10** 1725–1756. MR3553242 https://doi.org/10.1214/16-AOAS957

GORDON, J. S. (2017). *Nonparametric Estimation, Forecasts, and Model Evaluation of Spatial Temporal Point Process Models for California Seismicity*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of California, Los Angeles. MR3697731

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HAWKES, A. G. and OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. MR0378093 https://doi.org/10.2307/3212693

HELMSTETTER, A. (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *Bull. Seismol. Soc. Amer.* **96** 90–106.

HELMSTETTER, A. and SORNETTE, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J. Geophys. Res., Solid Earth* **107**.

HOLSCHNEIDER, M., NARTEAU, C., SHEBALIN, P., PENG, Z. and SCHORLEMMER, D. (2012). Bayesian analysis of the modified Omori law. *J. Geophys. Res., Solid Earth* **117**.

KOLEV, A. A. and ROSS, G. J. (2019). Inference for ETAS models with non-Poissonian mainshock arrival times. *Stat. Comput.* **29** 915–931. MR3994609 https://doi.org/10.1007/s11222-018-9845-z

KUMAZAWA, T. and OGATA, Y. (2014). Nonstationary ETAS models for nonstandard earthquakes. *Ann. Appl. Stat.* **8** 1825–1852. MR3271355 https://doi.org/10.1214/14-AOAS759

LIPPIELLO, E., GIACCO, F., DE ARCANGELIS, L., MARZOCCHI, W. and GODANO, C. (2014). Parameter estimation in the ETAS model: Approximations and novel methods. *Bull. Seismol. Soc. Amer.* **104** 985–994.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519 https://doi.org/10.1214/aos/1176346412

LOMBARDI, A. M. (2015). Estimation of the parameters of ETAS models by simulated annealing. *Sci. Rep.* **5** 8417.

LOMBARDI, A. M. (2017). SEDA: A software package for the statistical earthquake data analysis. *Sci. Rep.* **7** 44171. https://doi.org/10.1038/srep44171

MARKWICK, D. and ROSS, G. (2021). The Hierarchical Nonparametric Hawkes Process. Preprint.

MARSAN, D. and LENGLINE, O. (2008). Extending earthquakes' reach through cascading. *Science* **22** 1076–1079.

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

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

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

OMI, T., OGATA, Y., HIRATA, Y. and AIHARA, K. (2016). Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches. *J. Geophys. Res.* **120** 2561–2578.

RASMUSSEN, J. G. (2011). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642.

RASMUSSEN, J. G. (2013). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642. MR3085883 https://doi.org/10.1007/s11009-011-9272-5

REINHART, A. and GREENHOUSE, J. (2018). Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1305–1329. MR3873709 https://doi.org/10.1111/rssc.12277

ROSS, G. J. (2021). Bayesian estimation of the ETAS model for earthquake occurrences. *Bull. Seismol. Soc. Amer.* **111** 1473–1480.

ROSS, G. J. and KOLEV, A. A. (2022). Supplement to "Semiparametric Bayesian Forecasting of SpatioTemporal Earthquake Occurrences." Available at https://cran.r-project.org/web/packages/bayesianETAS/index.html. https://doi.org/10.1214/21-AOAS1554SUPP

ROSS, G. J. and MARKWICK, D. (2018). Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models Technical report.

SCHOENBERG, F. P. (2013). Facilitated estimation of ETASShort note. *Bull. Seismol. Soc. Amer.* **103** 601–605.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

UTSU, T. and OGATA, Y. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *J. Phys. Earth* **43** 1–33.

VARGAS, N. (2012). Bayesian Point Process Modelling of Earthquake Occurrences Technical report, Ruprecht-Karls Univ. Heidelberg.

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998 https://doi.org/10.1198/016214508000000148

WANG, Q., SCHOENBERG, F. and JACKSON, D. (2010). Standard errors of parameter estimates in the ETAS model. *Bull. Seismol. Soc. Amer.* **106** 1989–2001.

WIEMER, S. and WYSS, M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Amer.* **90** 859–869.

ZHUANG, J. and MATEU, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *J. Roy. Statist. Soc. Ser. A* **182** 919–942. MR3955503

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459 https://doi.org/10.1198/016214502760046925