Statistics and Computing (2022) 32:29
https://doi.org/10.1007/s11222-022-10085-3

GP-ETAS: semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model

Christian Molkenthin[1] · Christian Donner[2] · Sebastian Reich[1] · Gert Zöller[1] · Sebastian Hainzl[3] · Matthias Holschneider[1] · Manfred Opper[4]

Received: 27 August 2020 / Accepted: 29 January 2022 / Published online: 8 March 2022

Abstract
The spatio-temporal epidemic type aftershock sequence (ETAS) model is widely used to describe the self-exciting nature of earthquake occurrences. While traditional inference methods provide only point estimates of the model parameters, we aim at a fully Bayesian treatment of model inference, allowing naturally to incorporate prior knowledge and uncertainty quantification of the resulting estimates. Therefore, we introduce a highly flexible, non-parametric representation for the spatially varying ETAS background intensity through a Gaussian process (GP) prior. Combined with classical triggering functions this results in a new model formulation, namely the GP-ETAS model. We enable tractable and efficient Gibbs sampling by deriving an augmented form of the GP-ETAS inference problem. This novel sampling approach allows us to assess the posterior model variables conditioned on observed earthquake catalogues, i.e., the spatial background intensity and the parameters of the triggering function. Empirical results on two synthetic data sets indicate that GP-ETAS outperforms standard models and thus demonstrate the predictive power for observed earthquake catalogues including uncertainty quantification for the estimated parameters. Finally, a case study for the l'Aquila region, Italy, with the devastating event on 6 April 2009, is presented.

Keywords Self-exciting point process · Hawkes process · Spatio-temporal ETAS model · Bayesian inference · Sampling · Earthquake modeling · Gaussian process · Data augmentation

1 Introduction

Point process models are often used in statistical seismology for describing the occurrence of earthquakes (point data) in a spatio-temporal setting. The most widely used one is the epidemic type aftershock sequence (ETAS) model, first introduced as a temporal point process model (Ogata 1988), and later enhanced to the currently predominantly employed spatio-temporal version (Ogata 1998). Main applications are seismic forecasting or the characterisation of earthquake clustering in a particular geographical region and topics alike (e.g., Jordan et al. 2011). The ETAS model is an example of a self-exciting, spatio-temporal, marked point process, which is a particular Hawkes process model, extending the temporal Hawkes process (Hawkes 1971). Self-excitation means

Christian Molkenthin and Christian Donner: Equal contribution.

✉ Christian Molkenthin
molkenth@uni-potsdam.de

✉ Christian Donner
christian.research@mailbox.org

Sebastian Reich
sebastian.reich@uni-potsdam.de

Gert Zöller
zoeller@uni-potsdam.de

Sebastian Hainzl
sebastian.hainzl@gfz-potsdam.de

Matthias Holschneider
hols@uni-potsdam.de

Manfred Opper
manfred.opper@tu-berlin.de

1 Institute of Mathematics, University of Potsdam, 14476 Potsdam, Germany

2 Swiss Data Science Center, ETH Zürich, 8006 Zurich, Switzerland

3 GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany

4 Artificial Intelligence Group, TU Berlin, 10587 Berlin, Germany

that one event can trigger a series of subsequent follow-up events (offspring), as in the case of earthquakes, main shocks and aftershocks. The ETAS model as a variant of a Hawkes process model assigns the earthquake magnitude as a mark to each event, and it usually employs a specific mark depending excitation kernel (Ogata 1988, 1998). Besides its primary application in seismology, the Hawkes process is utilised in several other domains, e.g. finance (Bacry et al. 2015; Filimonov and Sornette 2015), crime (Porter and White 2010; Mohler et al. 2011), neuronal activities (Gerhard et al. 2017), social networks (Zhou et al. 2013; Zhao et al. 2015), genomes (Reynaud-Bouret and Schbath 2010), transportation (Hu and Jin 2017).

The ETAS model is characterized by its conditional intensity function, that is, the rate of arriving events conditioned on the history of previous events. This time-dependent conditional intensity function itself consists of two parts, (i) a background intensity μ of a Poisson process, which models the arrival of spontaneous (exogenous) events, and (ii) a time-dependent triggering function φ which encodes the form of self-excitation by adding a positive impulse response for each event, that is, an instantaneous jump which decays gradually at time progresses. An alternative approach interprets the stationary Hawkes process (e.g. the ETAS model) as a Poisson cluster process or branching process (Hawkes and Oakes 1974), which leads to the concept of a non-observable, underlying branching structure (a latent variable). One assumes that each event is either (i) a spontaneous, exogenous background event, or (ii) an offspring event, i.e., it was triggered by an existing event. Background events are generated independently by a Poisson process with rate μ and form cluster centers. Each event may produce offspring, i.e., may become a parent of triggered events in the future, where each triggered event may produce further offspring. The intensity of these offspring processes is controlled by the triggering function φ. Thus, each event has either a direct parent from which it was generated (offspring events) or is background (exogenous events with no direct parent); this yields an ordered branching structure useful for designing simulation and inference algorithms, e.g. (Zhuang et al. 2002; Veen and Schoenberg 2008).

The fitting of an ETAS model to data entails learning the conditional intensity function. Most currently used ETAS models employ a *parametric* form for the background μ and the triggering function φ. The parameters are then calibrated via maximum-likelihood estimation (MLE), maximising the classical likelihood function for point processes. Unfortunately, MLE has no simple analytical form. Alternatively, different numerical optimisation methods are employed involving, e.g., an Expectation-Maximisation (EM; Dempster et al. 1977) algorithm using the latent branching structure (Ogata 1998; Veen and Schoenberg 2008; Lippiello et al. 2014; Lombardi 2015).

Non-parametric methods have also been suggested previously to fit the conditional intensity function (or parts of it). For example, Zhuang et al. (2002) and Adelfio and Chiodi (2014) fit simultaneously a non-parametric background intensity via kernel density estimation and a classical parametric triggering kernel; Marsan and Lengliné (2008) consider a constant background intensity combined with a non-parametric histogram estimator of the triggering kernel; Mohler et al. (2011) suggest non-parametric kernel density estimators for both the components, background μ and offspring φ; and Fox et al. (2016) propose a non-parametric kernel density estimator for the background and non-parametric histogram estimation for the triggering kernel. Furthermore, Bacry and Muzy (2016) suggest a non-Bayesian, non-parametric way of estimating the triggering function of a Hawkes process based on Wiener Hopf integral equation; Kirchner (2017) presents a non-Bayesian non-parametric estimation procedure for a multivariate Hawkes process based on an integer-valued autoregressive model.

Uncertainty quantification of the ETAS model remains challenging. Most estimation techniques deliver a point estimate for its conditional intensity function and uncertainty quantification is usually achieved by relying on standard errors of estimated ETAS parameters, based on the Hessian (Ogata 1978; Rathbun 1996; Wang et al. 2010). This approach requires that the observational window is long enough (sufficiently large sample size), otherwise it may lead to an underestimation of parameter uncertainties. Moreover, standard errors based on Hessians cannot be obtained in the non-parametric case. Another approach to uncertainty quantification relies on various bootstrap techniques based on many forward simulations, e.g., Fox et al. (2016). Ad hoc variants for quantifying uncertainty have also been devised, e.g., by the solutions of multiple optimisation runs of the MLE, e.g. Lombardi (2015).

None of the aforementioned uncertainty quantification methods are fully satisfactory and we believe that a fully semi-parametric Bayesian framework is worthwhile pursuing, which allows one to incorporate prior knowledge. The posterior distribution effectively encodes the uncertainty of the quantities arising from data and a prior distribution. However, this poses a challenge for a spatio-temporal ETAS model, as there is no known conjugate structure, that is, the posterior can not be obtained in closed-form. One way to deal with this problem is to employ Monte Carlo sampling techniques, e.g. via Markov chain Monte Carlo (MCMC). However implementing MCMC remains challenging for *non- or semi-parametric* conditional intensity functions. Several studies have suggested Bayesian methods for the temporal, or multivariate Hawkes process, either based on parametric forms of the conditional intensity function (Rasmussen 2013; Ross 2018) or for non-parametric versions (Linderman and Adams 2015; Donnet et al. 2018;

Zhang et al. 2019a, b; Zhou et al. 2019). But these studies rarely consider the spatio-temporal ETAS model and only with strong simplifications, e.g., a constant background intensity μ (Rasmussen 2013). Recently, however, Kolev and Ross (2020) considered an inhomogeneous background intensity modelled via a Dirichlet process.

It is desirable to estimate the spatially dependent background μ of an ETAS model fully *non-parametrically* as it is often difficult to specify an appropriate functional form a priori. The background intensity (also called long-term component) is of particular importance for seismic hazard assessment and seismic forecasting. It is often preferred to maintain a specific *parametric* triggering function φ (e.g., modified Omori law Omori 1894; Utsu 1961) as there is a long tradition for interpreting and comparing this particular parametric form in different settings, regions, etc. Thus, one faces two main issues for the development of a suitable Bayesian inference approach:

(i) providing a Bayesian non-parametric way of modelling the background intensity μ, and
(ii) creating a fully Bayesian inference algorithm for the resulting ETAS models including its parametric triggering component φ.

We address these two issues in this paper by first formulating a Bayesian non-parametric approach to the estimation of the background intensity μ via a Gaussian process (GP) prior. Secondly, we propose and implement a computationally tractable approach for the implied Bayesian inference problem by introducing auxiliary variables: a latent branching structure, a latent Poisson process, and latent Pólya–Gamma random variables. More specifically, we suggest to model the background intensity μ *non-parametrically* by sigmoid transformed realisations of a GP prior, i.e., as a Sigmoid-Gauss-Cox-Process (SGCP; Adams et al. 2009), which is a doubly stochastic Poisson process. No specific functional form has to be chosen for the intensity function, and the prior fully specifies the chosen GP. Adams et al. (2009) proposed a Bayesian inference scheme via MCMC for SGCPs. However, the suggested scheme is computationally demanding and convergence is slow. Our paper relies instead on the work of Donner and Opper (2018) who recently enhanced Bayesian inference for SGCPs substantially by data augmentation with Pólya–Gamma random variables (Polson et al. 2013). The triggering function φ is modelled in a classical *parametric* way, which together with the SGCP model for μ leads to a novel semi-parametric ETAS model formulation, which we denote as GP-ETAS. In order to implement such an approach, we need to address a number of computational challenges:

(i) the background intensity μ and the triggering function φ are not directly separable in the likelihood;
(ii) intractable integrals for the posterior computation when μ is modelled as SGCP, and
(iii) handling a non-Gaussian point process likelihood while using a Gaussian process prior.

We show how these challenges can be resolved by data augmentation (introducing auxiliary variables), which strongly simplifies the Bayesian inference problem. It effectively allows us to construct an efficient MCMC sampling scheme for the posterior involving an overall Gibbs sampler (Geman and Geman 1984) consisting of three main steps, each conditioned on the other two:

(a) conditionally sampling the latent branching structure which factorises the likelihood function into background and triggering component;
(b) conditionally sampling the posterior of the background intensity μ from explicit conditional densities easy to sample from; and
(c) conditionally sampling the parameters of the triggering function φ by employing Metropolis–Hastings (MH; Hastings 1970) steps.

The remainder of this paper is structured as follows: First we describe the classical spatio-temporal ETAS model; secondly we introduce our GP-ETAS model including a simulation algorithm; thirdly the Bayesian inference approach is presented; fourthly empirical results based on synthetic and real data illustrate practical aspects of the framework. The paper concludes with a discussion and some final remarks.

2 Background

We start with a review of the classical spatio-temporal ETAS model, which we will use as a benchmark for comparison.

2.1 Classical ETAS model

The ETAS model (Ogata 1998), describes a stochastic process, which generates point pattern over some domain $\mathcal{X} \times \mathcal{T} \times \mathcal{M}$, where $\mathcal{T} \times \mathcal{X}$ is the time-space window and $\mathcal{M}$ the mark space of the process. Realisations of this point process are denoted by $\mathcal{D} = \{(t_i, \boldsymbol{x}_i, m_i)\}_{i=1}^{N_\mathcal{D}}$, which in seismology can be interpreted as an earthquake catalog consisting of $N_\mathcal{D}$ observed events. $\mathcal{D}$ is usually ordered in time (time series), $t_i \in \mathcal{T} \subseteq \mathbb{R}_{>0}$ is the time of the ith event (time of the earthquake), $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^2$ is the corresponding location (longitude and latitude of the epicenter), and $m_i \in \mathcal{M} \subseteq \mathbb{R}$ the corresponding mark (the magnitude of the earthquake).

2.1.1 Interpretations

There are two equivalent interpretations of the ETAS model (Hawkes process). We briefly discuss both.

Conditional intensity function. One way to define the ETAS model is by a conditional intensity function, which models the infinitesimal rate of expected arrivals around $(t, \boldsymbol{x})$ given the history $H_t = \{(t_i, \boldsymbol{x}_i, m_i) : t_i < t\}$ of the process until time t. The marked point process has intensity $\tilde{\lambda}(t, \boldsymbol{x}, m|H_t) = \lambda(t, \boldsymbol{x}|H_t)p_M(m)$, which factorizes under usual assumptions for the ETAS model (e.g., Zhuang et al. 2002; Daley and Vere-Jones 2003) in a ground process $\lambda(t, \boldsymbol{x}|H_t)$ given in (1) and a mark distribution $p_M(m)$ which models earthquake magnitudes $m \geq m_0$ independently following an exponential distribution $p_M(m|\beta) = \beta e^{-\beta(m-m_0)}$, $\beta > 0$. Density $p_M(m)$ corresponds to a Gutenberg-Richter law, and m_0 is the magnitude of completeness, (cut-off magnitude) a threshold above which all events are observed (complete data). . The conditional ETAS intensity function of the ground process can be written as (Ogata 1998)

$$\lambda(t, \boldsymbol{x}|H_t, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi) = \mu(\boldsymbol{x}|\boldsymbol{\theta}_\mu) + \sum_{i:t_i<t} \varphi(t-t_i, \boldsymbol{x}-\boldsymbol{x}_i|m_i, \boldsymbol{\theta}_\varphi), \tag{1}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)$ a set of parameters. Here the background intensity $\mu(\boldsymbol{x}|\boldsymbol{\theta}_\mu) : \mathbb{R}^2 \to [0, \infty)$ defines a non-homogeneous Poisson process in space but stationary in time with $\boldsymbol{\theta}_\mu$ as the required parameters, while $\varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i|m_i, \boldsymbol{\theta}_\varphi) : \mathbb{R}^4 \to [0, \infty)$ is the triggering function, modeling the rate of aftershocks (self-exciting process) following an event at $(t_i, \boldsymbol{x}_i)$ with magnitude m_i, controlled by the parameters $\boldsymbol{\theta}_\varphi$. Specific parametric representations of $\mu(\cdot)$ and $\varphi(\cdot)$ for the ETAS model will be discussed in Sect. 2.1.2.

Latent branching structure. Another interpretation of a Hawkes process (with the ETAS model being a particular example) is as Poisson cluster- or branching process (Hawkes and Oakes 1974), leading to the concept of an underlying branching structure, that is, a non observable latent random variable z_i for each event i. Events are structured in an ensemble of trees, either having a parent, which is one of the previous events or being spontaneous, called background. The latent variable is typically modelled as taking integer values in a discrete set $z_i \in \{0, 1, \ldots, i-1\}$, where

$$z_i = \begin{cases} 0 & \text{event } i \text{ is background} \\ j > 0 & \text{event } i \text{ is direct offspring (aftershock)} \\ & \text{of event } j \text{ at } t_j < t_i \end{cases} \tag{2}$$

Background events defined as $\mathcal{D}_0 = \{(t_i, \boldsymbol{x}_i, m_i, z_i = 0)\}_{i=1}^{N_{\mathcal{D}_0}}$ with $z_i = 0$ occur according to a Poisson process with intensity $\mu(\boldsymbol{x})$ and form cluster centres, i.e., initial points for branching trees. Within each branching tree, an existing event at t_j can produce direct offspring at $t > t_j$ according to an inhomogeneous Poisson process with rate $\lambda_j(t|t_j, \boldsymbol{x}_j, m_j) = \varphi(t - t_j, \boldsymbol{x} - \boldsymbol{x}_j|m_j, \boldsymbol{\theta}_\varphi)$. The overall intensity $\lambda(t, \boldsymbol{x}|H_t)$ is the sum of all the offspring Poisson processes $\sum_j \lambda_j$ with $t_j < t$ and the background Poisson process $\mu(\boldsymbol{x})$ (Poisson superposition), as given in (1). All events, which are not background, are offspring events defined as $\mathcal{D}_\varphi = \{(t_i, \boldsymbol{x}_i, m_i, z_i \neq 0)\}_{i=1}^{N_{\mathcal{D}_\varphi}}$, and $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_\varphi$, where $\mathcal{D}$ are all the observations.

The latent branching structure cannot be observed. However, by its construction (superposition of i Poisson processes at t_i) the probability $p_{i0} = p(z_i = 0)$ (background event) is (see,e.g., Zhuang et al. 2002),

$$p_{i0} = \frac{\mu(\boldsymbol{x}_i|\boldsymbol{\theta}_\mu)}{\lambda(t_i, \boldsymbol{x}_i|H_{t_i}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)}, \tag{3}$$

while the probability $p_{ij} = p(z_i = j)$ (event j triggered event i, $j > 0$) is,

$$p_{ij} = \frac{\varphi(t_i - t_j, \boldsymbol{x}_i - \boldsymbol{x}_j|m_j, \boldsymbol{\theta}_\varphi)}{\lambda(t_i, \boldsymbol{x}_i|H_{t_i}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)}, \tag{4}$$

with $p_{i0} + \sum_j p_{ij} = 1$.

2.1.2 Components of the ETAS model

This section sketches the components (background and triggering function) as given in (1).

Background intensity. The background intensity $\mu(\boldsymbol{x})$ is usually modelled either as piecewise constant function over a rectangular grid (or specific polygones, seismo-tectonic units) with L cells (e.g., in Veen and Schoenberg 2008; Lombardi 2015),

$$\mu(\boldsymbol{x}|\boldsymbol{\theta}_\mu) = \mu_l \tag{5}$$

if $\boldsymbol{x}$ is in grid cell l, $l = 1, \ldots, L$; or via a weighted kernel density estimator with variable bandwidth, as suggested by Zhuang et al. (2002),

$$\mu_{\text{kde}}(\boldsymbol{x}) = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{N_{\mathcal{D}}} p_{i0} k_{d_i}(\boldsymbol{x} - \boldsymbol{x}_i). \tag{6}$$

Here, $|\mathcal{T}|$ is the length of the observational time window, p_{i0} is the probability that event i is background as defined in (3), $d_i = \max\{d_{min}, r_{i,n_p}\}$ is the variable bandwidth determined for event i corresponding to the distance r_{i,n_p} of its number of nearest neighbours n_p, where d_{min} is some minimal bandwidth, and $k_d(\cdot)$ is an isotropic, bivariate Gaussian kernel function. There are different suggestions to select n_p; Zhuang et al. (2002) propose to choose n_p between 10 and 100, and state that estimated parameters only change slightly if n_p is changed in the range of 15–100; Zhuang (2011) suggests based on cross-validation experiments, that an optimal n_p is in the range 3–6 for Japan. The minimal bandwidth is commonly chosen as $d_{min} \in [0.02, 0.05]$ degrees, which is in the range of the localisation error (Zhuang et al. 2002).

Background parameters to be estimated are $\boldsymbol{\theta}_\mu = (\mu_1, \mu_2, \ldots, \mu_L)$ in the first case and the scaled kernel density estimator $\mu_{\text{kde}}(\boldsymbol{x})$ given through estimated background probabilities $\{p_{i0}\}_{i=1}^{N_{\mathcal{D}}}$ in the second case, respectively. For non-parametric models of μ as in (6) we neglect the explicit dependency on $\boldsymbol{\theta}_\mu$ in our notation, but the reader should keep in mind, that in such cases μ depends on a varying (potentially infinite) number of parameters.

Parametric triggering function. The triggering function $\varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_\varphi)$ characterizes the intensity of the offspring processes, thus, the number and distribution of offspring events in space and time. Offspring events (aftershocks) $\mathcal{D}_\varphi = \{(t_i, \boldsymbol{x}_i, m_i, z_i \neq 0)\}_{i=1}^{N_{\mathcal{D}_\varphi}}$ are triggered by previous events and are all events which are not background. The triggering function $\varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_\varphi)$ of the ETAS model is usually a non-negative parametric function, which is separable in space and time, and depends on m_i and $\boldsymbol{\theta}_\varphi$. There are numerous suggested parameterisations. See, for example, Ogata (1998); Zhuang et al. (2002); Console et al. (2003); Ogata and Zhuang (2006). One of the most common parametrisations is provided by

$$\begin{aligned}&\varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_\varphi) \\ &\quad = \kappa(m_i | K_0, \alpha) g(t - t_i | c, p) s(\boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_s).\end{aligned} \tag{7}$$

The first term $\kappa(\cdot)$ is proportional to the aftershock productivity (or Utsu law, Utsu 1970) of event i with m_i,

$$\kappa(m_i | K_0, \alpha) = K_0 e^{\alpha(m_i - m_0)}, \tag{8}$$

and K_0 is called productivity coefficient. Parameters K_0 and α determine the average number of offspring events (aftershocks) of event i per unit time. The second term $g(\cdot)$ describes the temporal distribution of aftershocks (offspring); a power law decay proportional to the modified Omori Utsu law (Omori 1894; Utsu 1961), and $t - t_i > 0$ is the elapsed time since the parent event (main shock), that is,

$$g(t - t_i | c, p) = (t - t_i + c)^{-p}. \tag{9}$$

Finally, the third term $s(\cdot)$ is a probability density function for the spatial distribution of the direct aftershocks (offspring) around the triggering event at $\boldsymbol{x}_i$. Often, one of the following probability density functions are employed. One distinguishes between a short range decay, which uses an isotropic Gaussian distribution with covariance $d_1^2 e^{\alpha(m_i - m_0)} \boldsymbol{I}$ (Ogata 1998; Zhuang et al. 2002); and a long range decay following a Pareto distribution (Kagan 2002; Ogata and Zhuang 2006),

$$\begin{aligned}&s(\boldsymbol{x} - \boldsymbol{x}_i | m_i, d, \gamma, q) \\ &\quad = \frac{q - 1}{\pi \sigma_m(m_i)} \left(1 + \frac{(\boldsymbol{x} - \boldsymbol{x}_i)^\top (\boldsymbol{x} - \boldsymbol{x}_i)}{\sigma_m(m_i)}\right)^{-q},\end{aligned} \tag{10}$$

where $\sigma_m(m_i) = d^2 10^{2\gamma(m_i)}$.

The unknown parameters to be estimated are $\boldsymbol{\theta}_\varphi = (K_0, \alpha, c, p, d_1)$, or $\boldsymbol{\theta}_\varphi = (K_0, \alpha, c, p, d, \gamma, q)$ depending on which version of $s(\cdot)$ is used. Note that $q > 1$ and the rest of the parameters are strictly positive.

2.1.3 Parameter estimation via MLE

The likelihood function observing $\mathcal{D}$ under the spatio-temporal ETAS model is given in (15); it is usually analytically intractable for simple direct optimisation. Numerical optimisation methods (e.g., quasi-Newton methods as in Ogata 1988, 1998, using an EM algorithm Veen and Schoenberg 2008 or simulated annealing Lombardi 2015; Lippiello et al. 2014) are usually employed. Often the integral term related to the triggering function in (15) using (1) is approximated as $\int_{\mathcal{T}_i} \int_{\mathcal{X}} \sum_{i:t_i<t} \varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_\varphi) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \leq \int_{\mathcal{T}_i} \int_{\mathbb{R}^2} \sum_{i:t_i<t} \varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i | m_i, \boldsymbol{\theta}_\varphi) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t$, by integrating over $\mathbb{R}^2$ in space instead of an arbitrary $\mathcal{X}$ (Schoenberg 2013). The introduced bias is small and often negligible (Schoenberg 2013; Lippiello et al. 2014) while the computations are greatly simplified as $\int_{\mathbb{R}^2} s(\boldsymbol{x} - \boldsymbol{x}_i | m_i) \, \mathrm{d}\boldsymbol{x} = 1$. We also use this approximation. Computational and numerical details of MLE using (15) are given in Ogata (1998). Instead of directly maximising (15), one can augment the likelihood function by a the latent branching structure Z and apply an EM algorithm for MLE (Veen and Schoenberg 2008; Mohler et al. 2011),

which is supposed to be advantageous, e.g. regarding stability and convergence (Veen and Schoenberg 2008).

3 Bayesian GP-ETAS model

Our goal is to improve the inference of the spatio-temporal ETAS model in order to allow for comprehensive uncertainty quantification. Despite the availability of powerful MLE based inference methods (see,e.g., Ogata 1998; Veen and Schoenberg 2008; Lippiello et al. 2014; Lombardi 2015), we believe that a Bayesian framework can complement existing methods and will provide a more reliable quantification of uncertainties.

3.1 GP-ETAS model specification

We introduce a novel formulation of the spatio-temporal ETAS model, which models the background rate $\mu(\boldsymbol{x})$ in a Bayesian *non-parametric* way via a GP (Williams and Rasmussen 2006), while the triggering function $\varphi(\cdot)$ assumes still a classical parametric form (modified Omori law (7)). As we will see subsequently, we are able to perform Bayesian inference for this model via Monte Carlo sampling despite its complex form.

While the conditional intensity function of the GP-ETAS model is still given by (1), the background intensity is a priori defined by

$$\mu(\boldsymbol{x}) = \bar{\lambda}\sigma(f(\boldsymbol{x})) = \frac{\bar{\lambda}}{1+e^{-f(\boldsymbol{x})}}, \tag{11}$$

where $\sigma(\cdot)$ is the logistic sigmoid function, $\bar{\lambda}$ a positive scalar, and $f(\boldsymbol{x})$ an arbitrary scalar function mapping $\boldsymbol{x} \in \mathcal{X}$ to the real line $\mathbb{R}$. Since $\sigma : \mathbb{R} \to [0, 1]$ the background intensity of the GP-ETAS model is bounded from above by $\bar{\lambda}$, i.e., $\mu(\boldsymbol{x}) \in [0, \bar{\lambda}]$ for any $\boldsymbol{x} \in \mathcal{X}$.

For the function $f(\boldsymbol{x})$ the GP-ETAS model assumes a Gaussian process prior, which implies that the prior over any discrete set of J function values $\boldsymbol{f} = \{f(\boldsymbol{x}_i)\}_{i=1}^{J}$ at positions $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_J\}$ is a J dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}_f, \boldsymbol{K}_{f,f})$, where $\boldsymbol{\mu}_f$ is the prior mean and $\boldsymbol{K}_{f,f} \in \mathbb{R}^{J\times J}$ is the covariance matrix between function values at positions $\boldsymbol{x}_i$. The matrix $\boldsymbol{K}_{f,f}$ is built from the covariance function (kernel) $k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\nu})$ such that $\boldsymbol{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\nu})$, where $\boldsymbol{\nu}$ are hyperparameters. We set $\boldsymbol{\mu}_f = 0$ and employ a Gaussian covariance function

$$k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\nu}) = \nu_0 \prod_{i=1}^{2} e^{-\frac{(x_i - x_i')^2}{2\nu_i^2}}, \tag{12}$$

where ν_0 is the so called amplitude and (ν_1, ν_2) are the length scales, representing a distance in input space over which the function values become weakly correlated. Note that the parameter $\bar{\lambda}$ and the hyperparameters $\boldsymbol{\nu}$ are also to be inferred from the data. For an in–depth treatment of GPs we refer to Williams and Rasmussen (2006).

The complete specification of the prior model of GP-ETAS including the hyperparameters is now as follows:

$$\boldsymbol{\nu} \sim p_{\boldsymbol{\nu}}, \text{ a prior on } \boldsymbol{\nu} \text{ (exponential distribution)} \tag{13a}$$
$$f \sim \mathcal{GP} \text{ prior with zero mean and a covariance function} \tag{13b}$$
$$\bar{\lambda} \sim p_{\bar{\lambda}}, \text{ a prior on } \bar{\lambda} \text{ (gamma distribution)} \tag{13c}$$
$$\mu|\bar{\lambda}, f, \boldsymbol{\nu} \sim \text{prior model on } \mu \text{ as defined in (11)} \tag{13d}$$
$$\boldsymbol{\theta}_\varphi \sim p_{\boldsymbol{\theta}_\varphi}, \text{ a prior on } \boldsymbol{\theta}_\varphi \text{ of } \varphi(\cdot) \text{ (uniform distribution)} \tag{13e}$$

The corresponding observational model is

$$\mathcal{D}|\mu, \boldsymbol{\theta}_\varphi \sim \text{Hawkes process with GP-ETAS intensity function (1),(11)} \tag{14}$$

where $\mathcal{D}$ is the data. Note that some quantities are independent by construction, e.g., $\boldsymbol{\nu}$ and $\bar{\lambda}$, f and $\bar{\lambda}$.

Without the triggering function in the intensity function (1) the GP-ETAS model would be equivalent to the SGCP model which is used to describe an inhomogeneous Poisson process (Adams et al. 2009) because of its favourable statistical properties (Kirichenko and Van Zanten 2015).

In the following we sketch how to generate data from the GP-ETAS model. A full description of the Bayesian inference problem is provided in Sect. 4.

3.2 Simulating the GP-ETAS model

Data $\mathcal{D} = \{(t_i, \boldsymbol{x}_i, m_i)\}_{i=1}^{N_\mathcal{D}}$ can be easily simulated from the GP-ETAS model using the latent branching structure of the point process. We propose a procedure which consists of two parts:

1. Generate all *background* events $\mathcal{D}_0 = \{(t_i, \boldsymbol{x}_i, m_i, z_i = 0)\}_{i=1}^{N_{\mathcal{D}_0}}$ from a SGCP in Eq. (11) as explained in Adams et al. (2009).
2. Sample all *aftershock* events (offspring) given $\mathcal{D}_0$ in possibly several generations denoted as $\mathcal{D}_\varphi = \{(t_i, \boldsymbol{x}_i, m_i, z_i \neq 0)\}_{i=1}^{N_{\mathcal{D}_\varphi}}$ and add them to obtain $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_\varphi$.

The above procedure can be implemented based on the *thinning* algorithm (Lewis and Shedler 1976); a variant of rejection sampling for point processes.

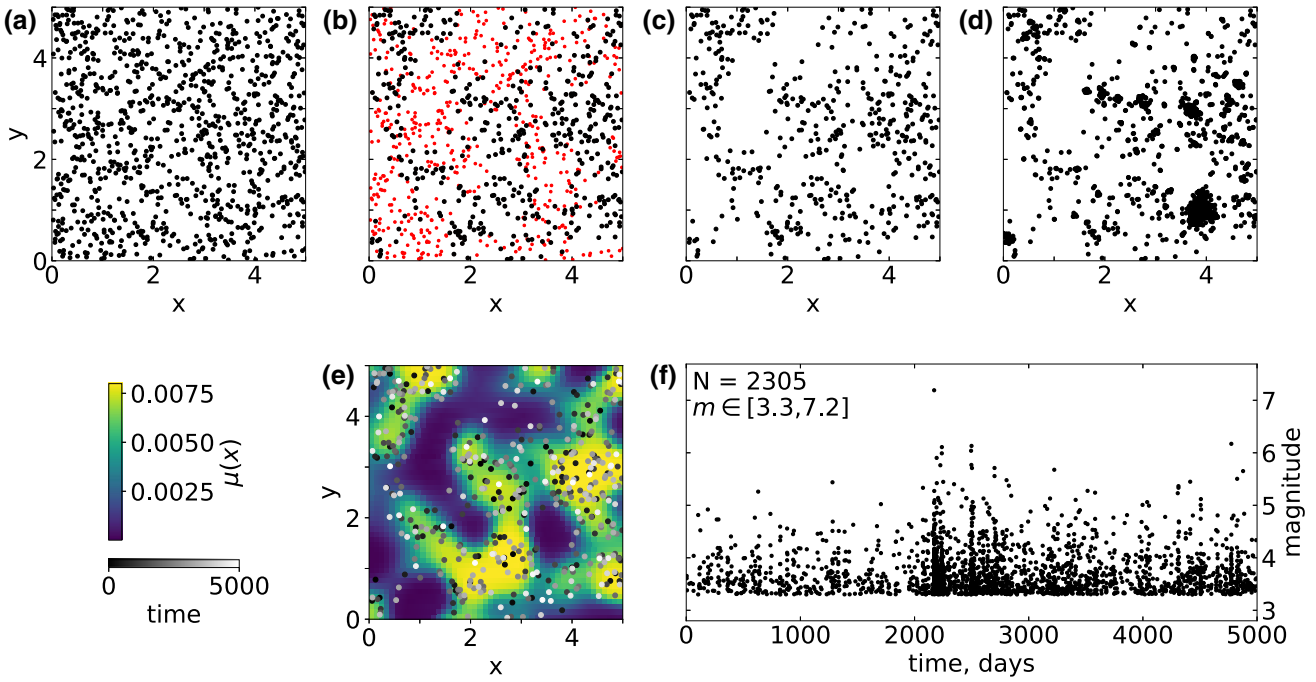

Fig. 1 The Figure depicts the different steps of a forward simulation of the generative GP-ETAS model. **a** Events of a homogeneous Poisson process with intensity $\bar{\lambda}$ are generated ($\bar{\lambda} = 0.008$, $N = 988$). **b** One retains events according to an inhomogeneous Poisson process with the desired intensity $\mu(\boldsymbol{x}) = \bar{\lambda}\sigma(f(\boldsymbol{x}))$ by randomly deleting events (red dots) via *thinning*. **c** The background events (black dots from (**b**)) are denoted by $\mathcal{D}_0$ ($N_{\mathcal{D}_0} = 481$), **d** After adding aftershocks (offspring events) $\mathcal{D}_\varphi$ to $\mathcal{D}_0$ in accordance with the triggering function $\varphi(\cdot)$ one obtains finally the simulated data $\mathcal{D}$ ($N_{\mathcal{D}} = 2305$) of the spatio-temporal GP-ETAS model. **e** Shows the background intensity $\mu(\boldsymbol{x}) = \bar{\lambda}\sigma(f(\boldsymbol{x}))$ together with the generated background events. Gray scaling of the dots refers to the event times. **f** Depicts the simulated data as a synthetic earthquake catalogue in time

After choosing $\bar{\lambda}$, $\boldsymbol{\nu}$, $\boldsymbol{\theta}_\varphi$ and a mark distribution $p(m)$, the simulation procedure of $\mathcal{D} \in \mathcal{X} \times \mathcal{T} \times \mathcal{M}$ can be summarised as follows: *First part:* One uses the upper bound $\bar{\lambda}$ to generate positions $\{\boldsymbol{x}_j\}_{j=1}^J$ of events from a homogeneous Poisson process with mean $|\mathcal{X}||\mathcal{T}|\bar{\lambda}$ which provide candidate background events (Fig. 1a). Subsequently a Gaussian process $\boldsymbol{f}$ is sampled from the prior $\mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{f,f})$ based on $\{\boldsymbol{x}_j\}_{j=1}^J$ using (12). The values $\mu(\boldsymbol{x}_j)$ can be computed using (11). Afterwards events, which do not follow an inhomogeneous Poisson process with intensity $\mu(\boldsymbol{x})$ as given by (11), are randomly deleted via *thinning* (Fig. 1b). The remaining $N_{\mathcal{D}_0}$ events are background events (Fig. 1c). The event times $\{t_i\}_{i=1}^{N_{\mathcal{D}_0}}$ are sampled from a uniform distribution $\mathcal{U}(|\mathcal{T}|)$ and the marks $\{m_i\}_{i=1}^{N_{\mathcal{D}_0}}$ from an exponential distribution, e.g., Gutenberg-Richter relation. Finally one obtains $\mathcal{D}_0$. *Second part:* Given the background events $\mathcal{D}_0$, the aftershock events (offsprings) of all generations are added to $\mathcal{D}_0$ in accordance with the triggering function $\varphi(\cdot)$ and using the mark distribution which yields $\mathcal{D}$ (Fig. 1d).

The overall simulation algorithm is described in detail in the Appendix 1, and is visualised in Fig. 1.

4 Bayesian inference

In this section, we address the Bayesian inference problem of our spatio-temporal GP-ETAS model. The objective is to estimate the joint posterior density $p(\mu, \boldsymbol{\theta}_\varphi|\mathcal{D})$, which encodes the knowledge (including uncertainties) about μ and $\boldsymbol{\theta}_\varphi$ after having seen the data. This is because, the posterior density combines information about μ and $\boldsymbol{\theta}_\varphi$ contained in the data (via the likelihood function) and prior knowledge (information before seeing the data) about μ and $\boldsymbol{\theta}_\varphi$. Here, μ denotes the entire random field of the background intensity as in (13d) and $\boldsymbol{\theta}_\varphi$ are the parameters of the triggering function.

The likelihood of observing a point pattern $\mathcal{D} = \{(t_i, \boldsymbol{x}_i, m_i)\}_{i=1}^{N_\mathcal{D}}$ under the GP-ETAS model (11) is given by the point process likelihood

$$p(\mathcal{D}|\mu, \boldsymbol{\theta}_\varphi) = \prod_{i=1}^{N_\mathcal{D}} \lambda(t_i, \boldsymbol{x}_i|\mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\varphi) \exp\left(-\int_\mathcal{T}\int_\mathcal{X} \lambda(t, \boldsymbol{x}|\mu(\boldsymbol{x}), \boldsymbol{\theta}_\varphi)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t\right), \tag{15}$$

where the intensity $\lambda(\cdot)$ is given by (11), and the dependencies on H_t, H_{t_i} are omitted for notational convenience.

Assuming a joint prior distribution denoted here by $p(\mu, \boldsymbol{\theta}_\varphi)$ for simplicity, the posterior distribution becomes

$$p(\mu, \boldsymbol{\theta}_\varphi|\mathcal{D}) \propto p(\mathcal{D}|\mu, \boldsymbol{\theta}_\varphi)p(\mu, \boldsymbol{\theta}_\varphi). \tag{16}$$

This posterior is intractable in practice and hence standard inference techniques are not directly applicable. More precisely, the following three main challenges arise:

(i) The background intensity μ and triggering function $\varphi(\cdot|\boldsymbol{\theta}_\varphi)$ cannot be treated separately in the likelihood function (15).
(ii) The likelihood (15) includes an intractable integral inside the exponential term due to the GP prior on f in (11), that is the integral of f over $\mathcal{X}$. Furthermore, normalisation of (16) requires an intractable marginalisation over μ and $\boldsymbol{\theta}_\varphi$. Thus, the posterior distribution is *doubly intractable* (Murray et al. 2006).
(iii) We assume a Gaussian process prior for modelling the background rate. However, the point process likelihood (15) is non-Gaussian, which makes the functional form of the posterior nontrivial to treat in practice.

We approach these challenges by data augmentation based on the work of Hawkes and Oakes (1974), Veen and Schoenberg (2008), Adams et al. (2009), Polson et al. (2013), Donner and Opper (2018). We will find that this augmentation simplifies the inference problem substantially. The following three auxiliary random variables are introduced:

(1) A *latent branching structure* Z, as described in Sect. 2.1.1, decouples μ and $\boldsymbol{\theta}_\varphi$ in the likelihood function (e.g., Veen and Schoenberg 2008). (See Sect. 4.1 and Eq. (17) for details.)
(2) A *latent Poisson process* Π enables an unbiased estimation of the integral term in the likelihood function that depends on μ, as the joint distribution of latent and observed data results in a homogeneous Poisson process with constant integral term. (See Sect. 4.2, paragraph Augmentation by a latent Poisson process and Eq. (21) for details.)
(3) We make use of the fact, that the logistic sigmoid function can be written as an infinite scale mixture of Gaussians using latent *Pólya–Gamma random variables* $\omega \sim p_{\mathrm{PG}}(\omega)$ (Polson et al. 2013), defined in Appendix 2. This leads to a likelihood representation, which is conditional conjugate to all the priors including the Gaussian process prior for the background component of the likelihood function (Donner and Opper 2018). (See Sect. 4.2, paragraph Augmentation by Pólya–Gamma random variables and Eqs. (23, 24) for details.)

These three augmentations allow one to implement a Gibbs sampling procedure (Geman and Geman 1984) that produces samples from the posterior distribution in (16). More precisely, random samples are generated in a Gibbs sampler by drawing one variable (or a block of variables) from the conditional posterior given all the other variables. Hence, we need to derive the required conditional posterior distributions as outlined next.

The suggested sampler consists of three modules using the solutions (data augmentations) sketched above: *sampling the latent branching structure*, *inference of the background* μ, and *inference of the triggering* $\boldsymbol{\theta}_\varphi$. Our overall *Gibbs sampling algorithm* of the posterior distribution is summarised in Algorithm 1. After an initial burn-in (a sufficiently long run of the three modules (Sects. 4.1–4.3), the generated samples converge to the desired joint posterior distribution $p(\mu, \boldsymbol{\theta}_\varphi|\mathcal{D})$.

In the following, we discuss some important aspects of the three modules of the Gibbs sampler which the sampler runs repeatedly trough.

4.1 Sampling the latent branching structure

Augmentation by the latent branching structure. We consider an auxiliary variable z_i for each data point i, which represents the latent branching structure as defined in Sect. 2.1.1. Recall that it gives the time index of the parent event. If $z_i = 0$ then the event is a spontaneous background event. Further we define $Z = \{z_i\}_{i=1}^{N_\mathcal{D}}$, which is the overall branching structure of the data $\mathcal{D}$. The likelihood $p(\mathcal{D}, Z|\mu, \boldsymbol{\theta}_\varphi)$ of the augmented model can be written as in 17,

$$p(\mathcal{D}, Z|\mu, \boldsymbol{\theta}_\varphi) = \underbrace{\prod_{i=1}^{N_\mathcal{D}} \mu(\boldsymbol{x}_i)^{\mathbb{I}(z_i=0)} \exp\left(-|\mathcal{T}| \int_\mathcal{X} \mu(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\right)}_{\text{(a)}=p(\mathcal{D}_0|Z,\mu)}$$
$$\times \underbrace{\prod_{i=1}^{N_\mathcal{D}} \prod_{j=1}^{i-1} \varphi_{ij}(\boldsymbol{\theta}_\varphi)^{\mathbb{I}(z_i=j)} \prod_{i=1}^{N_\mathcal{D}} \exp\left(-\int_{\mathcal{T}_i} \int_\mathcal{X} \varphi_i(\boldsymbol{\theta}_\varphi)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t\right)}_{\text{(b)}=\mathrm{p}(\mathcal{D}|\mathrm{Z},\boldsymbol{\theta}_\varphi)} p(Z), \tag{17}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, i.e., $\mathbb{I}(z_i = j)$ takes the value 1 for all $z_i = j$ and 0 otherwise, $\varphi_{ij}(\boldsymbol{\theta}_\varphi) = \varphi(t_i - t_j, \boldsymbol{x}_i - \boldsymbol{x}_j|m_j, \boldsymbol{\theta}_\varphi)$, $\varphi_i(\boldsymbol{\theta}_\varphi) = \varphi(t - t_i, \boldsymbol{x} - \boldsymbol{x}_i|m_i, \boldsymbol{\theta}_\varphi)$, $\mathcal{T}_i = [t_i, |\mathcal{T}|] \subset \mathcal{T}$, and all possible branching structures are equally likely, i.e. $p(Z) = \text{const}$. Furthermore, $\mathcal{D}_0 = \{\boldsymbol{x}_i\}_{i:z_i=0}$ denotes the set of $N_{\mathcal{D}_0}$ background events. Note, that marginalizing over Z in (17) recovers (15), because $\sum_{z_i=0}^{i-1} \mu(\boldsymbol{x}_i)^{\mathbb{I}(z_i=0)} \prod_{j=1}^{i-1} \varphi_{ij}(\boldsymbol{\theta}_\varphi)^{\mathbb{I}(z_i=j)} = \lambda(t_i, \boldsymbol{x}_i|\mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\phi)$. The augmented likelihood factorises into two independent components, (a) a likelihood component for

Algorithm 1 Gibbs Sampler for the posterior distribution of the GP-ETAS model

1: Initialise randomly $\bar{\lambda}^{(0)}$, $\boldsymbol{f}^{(0)}$, $\boldsymbol{\theta}_\varphi^{(0)}$ from the priors
2: **for** $k = 1$ to K **do**
3: **Factorisation of the likelihood (Section 4.1)**
4: $\forall i = 1, ..., N_\mathcal{D}$ Sample latent branching structure $z_i^{(k)}|\mathcal{D}, (\mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)}$ (19)
5: **Inference of the background intensity** $p(\bar{\lambda}, \boldsymbol{f}|\mathcal{D}_0, \boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi, Z)$ **(Section 4.2)**
6: Sample latent Poisson process $\Pi^{(k)}|(\bar{\lambda}, \boldsymbol{f})^{(k-1)}$ (25a)
7: $\forall l = N_\mathcal{D} + 1, ..., N_{\mathcal{D}\cup\Pi}$ Sample Pólya–gamma variables $\omega_l^{(k)}|f_l^{(k-1)}, \Pi^{(k)}$ (25b)
8: $\forall i : z_i = 0$ Sample Pólya–gamma variables $\omega_i^{(k)}|\mathcal{D}, f_i^{(k-1)}, Z^{(k)}$ (25c)
9: Set $\omega_i^{(k)} = 0$ otherwise.
10: Sample upper bound $\bar{\lambda}^{(k)}|(Z, \Pi)^{(k)}$ (25d)
11: Sample Gaussian process $\boldsymbol{f}^{(k)}|\mathcal{D}, (\boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi, Z)^{(k)}$ (25e)
12: Sample hyperparameters $\boldsymbol{v}^{(k)}$ using MH step (26)
13: $\forall i = 1, ..., N_\mathcal{D}$ compute $(\mu(\boldsymbol{x}_i))^{(k)}$ (11)
14: **Inference of the triggering function** $p(\boldsymbol{\theta}_\varphi|\mathcal{D}, Z^{(k)})$ **(Section 4.3)**
15: Sample $\boldsymbol{\theta}_\varphi^{(k)}$ using MH steps (28)
16: **end for**

the background intensity which depends on μ (first two terms on the rhs of (17)) and (b) a likelihood component of the triggering function which depends on $\boldsymbol{\theta}_\varphi$ (last two terms on the rhs of (17)).

From (17) one can derive the conditional distribution of z_i given all the other variables. Note that all z_i's are independent. The conditional distribution is proportional to a categorical distribution,

$$p(z_i|\mathcal{D}, \mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\varphi) \propto [\mu(\boldsymbol{x}_i)]^{\mathbb{I}(z_i=0)} \prod_{j=1}^{i-1} \left[\varphi_{ij}(\boldsymbol{\theta}_\varphi)\right]^{\mathbb{I}(z_i=j)} = \prod_{j=0}^{i-1} p_{ij}^{\mathbb{I}(z_i=j)}, \tag{18}$$

with the probabilities p_{ij} given by (3) and (4) which we collect in a vector $\boldsymbol{p}_i \in \mathbb{R}^i$.

From (18) one can see that the latent branching structure at the kth iteration of the Gibbs sampler is sampled from a categorical distribution, $\forall i = 1, \ldots, N_\mathcal{D}$

$$z_i^{(k)}|\mathcal{D}, (\mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)} \sim \text{Categorical}(\boldsymbol{p}_i). \tag{19}$$

Here $(\mu(\boldsymbol{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)}$ denotes the values of $\mu(\boldsymbol{x}_i)$ and $\boldsymbol{\theta}_\varphi$ from the previous iteration.

4.2 Inference for the background intensity

Given an instance of a branching structure Z, the background intensity in (17) depends on events i for which $z_i = 0$ only. One finds that the resulting term is a Poisson likelihood of the form

$$p(\mathcal{D}_0|f, \bar{\lambda}, Z) = \prod_{i=1:z_i=0}^{N_\mathcal{D}} \bar{\lambda}\sigma(f_i) \exp\left(-|\mathcal{T}| \int_\mathcal{X} \bar{\lambda}\sigma(f(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}\right), \tag{20}$$

where $\mu(\boldsymbol{x})$ has been replaced by (11) and $f_i = f(\boldsymbol{x}_i)$ has been used for notational convenience.

Because of the aforementioned problems in Sect. 4, sampling the conditional posterior $p(f, \bar{\lambda}|\mathcal{D}_0, Z)$ is still non-trivial and requires further augmentations which we describe next.

Augmentation by a latent Poisson process. We can resolve issue (ii) from Sect. 4 by introducing an independent latent Poisson process $\Pi = \{\boldsymbol{x}_l\}_{l=N_\mathcal{D}+1}^{N_{\mathcal{D}\cup\Pi}}$ on the data space with rate $\hat{\lambda}(\boldsymbol{x}) = \bar{\lambda}(1 - \sigma(f(\boldsymbol{x}))) = \bar{\lambda}(\sigma(-f(\boldsymbol{x})))$ using $1 - \sigma(z) = \sigma(-z)$. The points in $\mathcal{D}$, Π form the joint set $\mathcal{D} \cup \Pi$ with cardinality $N_{\mathcal{D}\cup\Pi}$. Note, that the number of elements in Π, i.e. N_Π, is also a random variable. The joint likelihood of $\mathcal{D}_0$ and the new random variable Π is,

$$p(\mathcal{D}_0, \Pi|f, \bar{\lambda}, Z) = \prod_{i=1:z_i=0}^{N_\mathcal{D}} \bar{\lambda}\sigma(f_i) \prod_{l=N_\mathcal{D}+1}^{N_{\mathcal{D}\cup\Pi}} \bar{\lambda}\sigma(-f_l) \exp\left(-|\mathcal{X}||\mathcal{T}|\bar{\lambda}\right), \tag{21}$$

where $f_l = f(\boldsymbol{x}_l)$. Thus, by introducing the latent Poisson process Π, we obtain a likelihood representation of the augmented system, where the former intractable integral inside the exponential term disappears, i.e. reduces to a constant.

We can gain some intuition by reminding ourselves of the aforementioned *thinning* algorithm (Lewis and Shedler 1976) in Sect. 3.2. Considering $\mathcal{D}_0$ as a resulting set of this algorithm, we wish to find the set Π, such that the joint set

$\mathcal{D}_0 \cup \Pi$ is coming from a homogeneous Poisson process with rate $\bar{\lambda}$. Because $\mathcal{D}_0$ is a sample of a Poisson process with rate $\bar{\lambda}\sigma(f)$ and the superposition theorem of Poisson processes (Kingman 1993), one finds that if Π is distributed according to a Poisson process with rate $\bar{\lambda}\sigma(-f)$, the joint set $\mathcal{D}_0 \cup \Pi$ has the rate $\bar{\lambda}\sigma(f) + \bar{\lambda}\sigma(-f) = \bar{\lambda}$. As we will see later, for the augmented model only the cardinality $|\mathcal{D}_0 \cup \Pi|$ will determine the posterior distribution of $\bar{\lambda}$.

Having a closer look at the augmented likelihood (21) and considering only terms depending on the function f, one can find a resemblance with a classical classification problem, namely logistic regression. Having the joint set, $\mathcal{D}_0 \cup \Pi$ the probability of a point belonging to $\mathcal{D}_0$ is $\sigma(f)$ and to Π it is $1 - \sigma(f) = \sigma(-f)$. Since we know, which points belong to which set, the aim is to find the function f, which best classifies/separates these two sets.

While above we provided some intuition, rigorously one can derive the latent Poisson process Π following Donner and Opper (2018). Note that (20) implies

$$\begin{aligned}
&\exp\left(-|\mathcal{T}| \int_{\mathcal{X}} \bar{\lambda}\sigma(f(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\right) \\
&= \exp\left(\int_{\mathcal{T}}\int_{\mathcal{X}} \bar{\lambda}(\sigma(-f(\boldsymbol{x})) - 1)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\right) \\
&= \mathbb{E}_{\bar{\lambda}}\left[\prod_{\boldsymbol{x}_l \in \Pi} \sigma(-f(\boldsymbol{x}_l))\right],
\end{aligned} \tag{22}$$

where the expectation is over random sets Π with respect to a Poisson process measure with rate $\bar{\lambda}$ on the space-time window of the data $\mathcal{T} \times \mathcal{X}$. Here, one uses Campbell's theorem (Kingman 1993). Writing the likelihood parts depending on f and $\bar{\lambda}$ in (17) in terms of the new random variable Π we get (21). Note that marginalisation over the augmented variable Π leads back to the background likelihood in (20) conditioned on the branching structure Z.

Note, that at this stage, with the augmentation in this section, our inference problem became tractable, because the augmented likelihood (21) depends on function f only at a finite set of points. In principle at this stage we could employ acceptance rejection algorithms as in Adams et al. (2009). However, to improve efficiency we will introduce one more variable augmentation in the next paragraph, that will allow rejection-free sampling of f.

Augmentation by Pólya–Gamma random variables. Investigating the augmented likelihood (21) issue (iii) from Sect. 4 is still present, because it is nonconjugate to the GP prior that we assume for function f in the GP-ETAS model. However, we noted before, the relation to a logistic regression problem. Polson et al. (2013) introduced the so-called *Pólya–Gamma random variables*, that allows to efficiently solve the inference problem of logistic GP classifiction (Wenzel et al. 2019). Here, we utilize the same methodology, where make use of the fact that the sigmoid function can be written an infinite scale mixture of Gaussians using latent (Polson et al. 2013), that is,

$$\sigma(z) = \frac{e^{\frac{z}{2}}}{2\cosh(\frac{z}{2})} = \frac{1}{2}e^{\frac{z}{2}} \int_0^\infty e^{-\frac{z^2}{2}\omega} p_{\mathrm{PG}}(\omega|1,0)\,\mathrm{d}\omega, \tag{23}$$

where the new random Pólya–Gamma variable ω is distributed according to the Pólya–Gamma density $p_{\mathrm{PG}}(\omega|1,0)$, see Appendix 2. Inserting the Pólya–Gamma representation of the sigmoid function (23) into (21) yields

$$\begin{aligned}
&p(\mathcal{D}_0, \boldsymbol{\omega}_{\mathcal{D}}, \Pi, \boldsymbol{\omega}_{\Pi}|\boldsymbol{f}, \bar{\lambda}, Z) \\
&= \prod_{i:z_i=0}^{N_{\mathcal{D}}} \frac{\bar{\lambda}}{2} e^{\frac{f_i}{2} - \frac{f_i^2}{2}\omega_i} p_{\mathrm{PG}}(\omega_i|1,0) \prod_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D}\cup\Pi}} \frac{\bar{\lambda}}{2} e^{-\frac{f_l}{2} - \frac{f_l^2}{2}\omega_l} p_{\mathrm{PG}}(\omega_l|1,0) \\
&\quad \times \exp\left(-\bar{\lambda}|\mathcal{X}|T\right),
\end{aligned} \tag{24}$$

where we set the Pólya-Gamma variables of all events $\boldsymbol{\omega}_{\mathcal{D}} = (\omega_1, \ldots, \omega_{N_{\mathcal{D}}})$ to $\omega_i = 0$ if $z_i \neq 0$. For the latent Poisson process the Pólya–Gamma variables are denoted by $\boldsymbol{\omega}_{\Pi} = (\omega_{N_{\mathcal{D}}+1}, \ldots, \omega_{N_{\mathcal{D}\cup\Pi}})$. The likelihood representation of the augmented system (24) has a Gaussian form with respect to $\boldsymbol{f}$ (that is, only linear or quadratic terms of $\boldsymbol{f}$ appear in the exponential function) and is therefore conditionally conjugate to the GP prior denoted by $p(\boldsymbol{f})$. Hence, we can implement an efficient Gibbs sampler for the background intensity function.

Employing a Gaussian process prior over $\boldsymbol{f}$ and a Gamma distributed prior over $\bar{\lambda}$, one gets from (24) the following conditional posteriors for the kth Gibbs iteration:

$$\Pi^{(k)} \mid (\bar{\lambda}, \boldsymbol{f})^{(k-1)} \sim \mathrm{PP}(\bar{\lambda}(\sigma(-f(\boldsymbol{x})))) \tag{25a}$$

$$\forall\, l : N_{\mathcal{D}} + 1, \ldots, N_{\mathcal{D}\cup\Pi}$$

$$\omega_l^{(k)} \mid f_l^{(k-1)}, \Pi^{(k)} \sim p_{\mathrm{PG}}(1, |f_l|) \tag{25b}$$

$$\forall\, i : z_i = 0$$

$$\omega_i^{(k)} \mid f_i^{(k-1)}, \mathcal{D}, Z^{(k)} \sim p_{\mathrm{PG}}(1, |f_i|) \tag{25c}$$

$$\bar{\lambda}^{(k)} \mid Z^{(k)}, \Pi^{(k)} \sim \mathrm{Gamma}\left(N_{\mathcal{D}_0\cup\Pi} + \alpha_0, |\mathcal{X}||\mathcal{T}| + \beta_0\right) \tag{25d}$$

$$\boldsymbol{f}^{(k)} \mid \mathcal{D}, (\boldsymbol{\omega}_{\mathcal{D}}, \Pi, \boldsymbol{\omega}_{\Pi}, Z)^{(k)} \sim \mathcal{N}\Big((\boldsymbol{\Omega} + \boldsymbol{K}^{-1})^{-1}\boldsymbol{u}, (\boldsymbol{\Omega} + \boldsymbol{K}^{-1})^{-1} \Big) \tag{25e}$$

where $\boldsymbol{f} = (\boldsymbol{f}_{\mathcal{D}}, \boldsymbol{f}_{\Pi}) \in \mathbb{R}^{N_{\mathcal{D}\cup\Pi}}$ is the Gaussian process at the data locations $\mathcal{D}$ and Π; and PP$(\cdot)$ denotes an inhomogeneous Poisson process with intensity $\bar{\lambda}(\sigma(-f(\boldsymbol{x})))$; $\boldsymbol{\Omega}$ is a diagonal matrix with $(\boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_{\Pi})$ as diagonal entries. $\boldsymbol{K} \in \mathbb{R}^{N_{\mathcal{D}\cup\Pi}\times N_{\mathcal{D}\cup\Pi}}$ is the covariance matrix of the Gaussian process prior at positions $\mathcal{D}$ and $\Pi^{(k)}$. It can be shown that, the vector $\boldsymbol{u}$ is $1/2$ for all entries in $\mathcal{D}_0$, zero for all entries of the remaining data $\mathcal{D}\backslash\mathcal{D}_0$, and $-1/2$ for the corresponding entries of Π. Gamma$(\cdot)$ is a Gamma distribution, where the Gamma prior has shape and rate parameters α_0, β_0. We used $e^{-\frac{c^2}{2}\omega} p_{\text{PG}}(\omega|1,0) \propto p_{\text{PG}}(\omega|1,c)$ due to the definition of a tilted Pólya–Gamma density (34) as given in (Polson et al. 2013), see Appendix 2. Note that one does not need an explicit form of the Pólya–Gamma density for our inference approach since it is sampling based. In other words, we only need an efficient way to sample from the tilted p_{PG} density (34) which was provided by Windle et al. (2014); Polson et al. (2013). Several p_{PG} samplers are freely available for different computer languages.

In summary, we first introduced a latent Poisson process Π to render the inference problem of f tractable. The additional Pólya–Gamma augmentation allows us to sample f rejection free, given samples of the augmented sets $\Pi, \boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_{\Pi}$. A detailed step-by-step derivation of the conditional distributions is given in the Appendix 3.

Hyperparameters. The Gaussian process covariance kernel given in (12) depends on the hyperparameters $\boldsymbol{\nu}$. Compare Sect. 3. We use exponentially distributed priors on $p(\nu_i) = p_{\nu_i}$, and we sample $\boldsymbol{\nu}$ using a standard MH algorithm as there is no closed form for the conditional posterior available. The only terms where $\boldsymbol{\nu}$ enter are in the Gaussian process prior and hence the relevant terms are

$$\ln p(\boldsymbol{\nu}|\boldsymbol{f}, \mathcal{D}, \Pi, \boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_{\Pi}) = -\frac{1}{2}\boldsymbol{f}^{\top}\boldsymbol{K}_{\boldsymbol{\nu}}^{-1}\boldsymbol{f} - \frac{1}{2}\ln\det\boldsymbol{K}_{\boldsymbol{\nu}} + \ln p(\boldsymbol{\nu}) + \text{const.}, \tag{26}$$

where $\boldsymbol{K}_{\boldsymbol{\nu}}$ is the Gaussian process prior covariance matrix depending on $\boldsymbol{\nu}$ via (12).

4.2.1 Conditional predictive posterior distribution of the background intensity

Given the kth posterior sample $(\bar{\lambda}^{(k)}, \boldsymbol{f}^{(k)}, \boldsymbol{\nu}^{(k)})$, the background intensity $\mu(\boldsymbol{x}^*)^{(k)}$ at any set of positions $\{\boldsymbol{x}_i^*\} \in \mathcal{X}$ (predictive conditional posterior) can be obtained in the following way, see (13d). Conditioned on $\boldsymbol{f}^{(k)}$ and hyperparameters $\boldsymbol{\nu}^{(k)}$ the latent function values $\boldsymbol{f}^*$ can be sampled via the conditional prior $p(\boldsymbol{f}^*|\boldsymbol{f}^{(k)}, \boldsymbol{\nu}^{(k)})$ using (43) with covariance function given in (12) (Williams and Rasmussen 2006). Using (11) one gets $\mu(\boldsymbol{x}^*)^{(k)} = \bar{\lambda}^{(k)}\sigma(\boldsymbol{f}^*)$.

4.3 Inference for the parameters of the triggering function

Given an instance of a branching structure Z, the likelihood function in (17) factorises in terms involving μ and terms involving $\boldsymbol{\theta}_{\varphi}$. The relevant terms related to $\boldsymbol{\theta}_{\varphi}$ are

$$p(\mathcal{D}|Z, \boldsymbol{\theta}_{\varphi}) = \prod_{i=1:z_i\neq 0}^{N_{\mathcal{D}}} \varphi(t_i - t_{z_i}, \boldsymbol{x}_i - \boldsymbol{x}_{z_i}|m_{z_i}, \boldsymbol{\theta}_{\varphi}) \times \prod_{i=1}^{N_{\mathcal{D}}} \exp\Big(-\int_{\mathcal{T}_i}\int_{\mathcal{X}} \varphi(t-t_i, \boldsymbol{x}-\boldsymbol{x}_i|m_i, \boldsymbol{\theta}_{\varphi})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \Big). \tag{27}$$

The conditional posterior $p(\boldsymbol{\theta}_{\varphi}|\mathcal{D}, Z) \propto p(\mathcal{D}|Z, \boldsymbol{\theta}_{\varphi})p(\boldsymbol{\theta}_{\varphi})$ with prior $p(\boldsymbol{\theta}_{\varphi})$ has no closed form. The dimension of $\boldsymbol{\theta}_{\varphi}$ is usually small (≤ 7). We employ MH sampling (Hastings 1970), which can be considered a nested step within the overall Gibbs sampler. We use a random walk MH where proposals are generated by a Gaussian in log space. The acceptance probability of $\boldsymbol{\theta}_{\varphi}^{(k)}$ based on (27) is given by

$$p_{\text{accept}} = \min\left\{1, \frac{p(\mathcal{D}|Z^{(k)}, \boldsymbol{\theta}_{\varphi}^{\text{proposed}})p(\boldsymbol{\theta}_{\varphi}^{\text{proposed}})}{p(\mathcal{D}|Z^{(k)}, \boldsymbol{\theta}_{\varphi}^{(k-1)})p(\boldsymbol{\theta}_{\varphi}^{(k-1)})}\right\}. \tag{28}$$

We take 10 proposals before we return to the overall Gibbs sampler, that is, to step in Sect. 4.1.

5 Experiments and results

We consider two kinds of experiments where we evaluate the performance of our proposed Bayesian approach GP-ETAS (see Sects. 3 and 4). First we look at synthetic data, with known conditional intensity $\lambda(t, \boldsymbol{x})$, i.e. with known background intensity $\mu(\boldsymbol{x})$ and known parameters $\boldsymbol{\theta}_{\varphi}$ of the triggering function. Here, we investigate if GP-ETAS can recover the model underlying the data well. Secondly, we apply our method to observational earthquake data.

Comparison. We compare our approach with the current standard spatio-temporal ETAS model which uses MLE. This classical ETAS model is based on kernel density estimation with variable bandwidths for the background intensity $\mu(\boldsymbol{x})$ as described in Sect. 2.1.2. Two variations are considered: (1) ETAS model with standard choice of the minimal bandwidth (0.05 degrees) and $n_p = 15$ the number of nearest neighbors used for obtaining the individual bandwidths (ETAS–classical; Zhuang et al. 2002), and (2) ETAS model

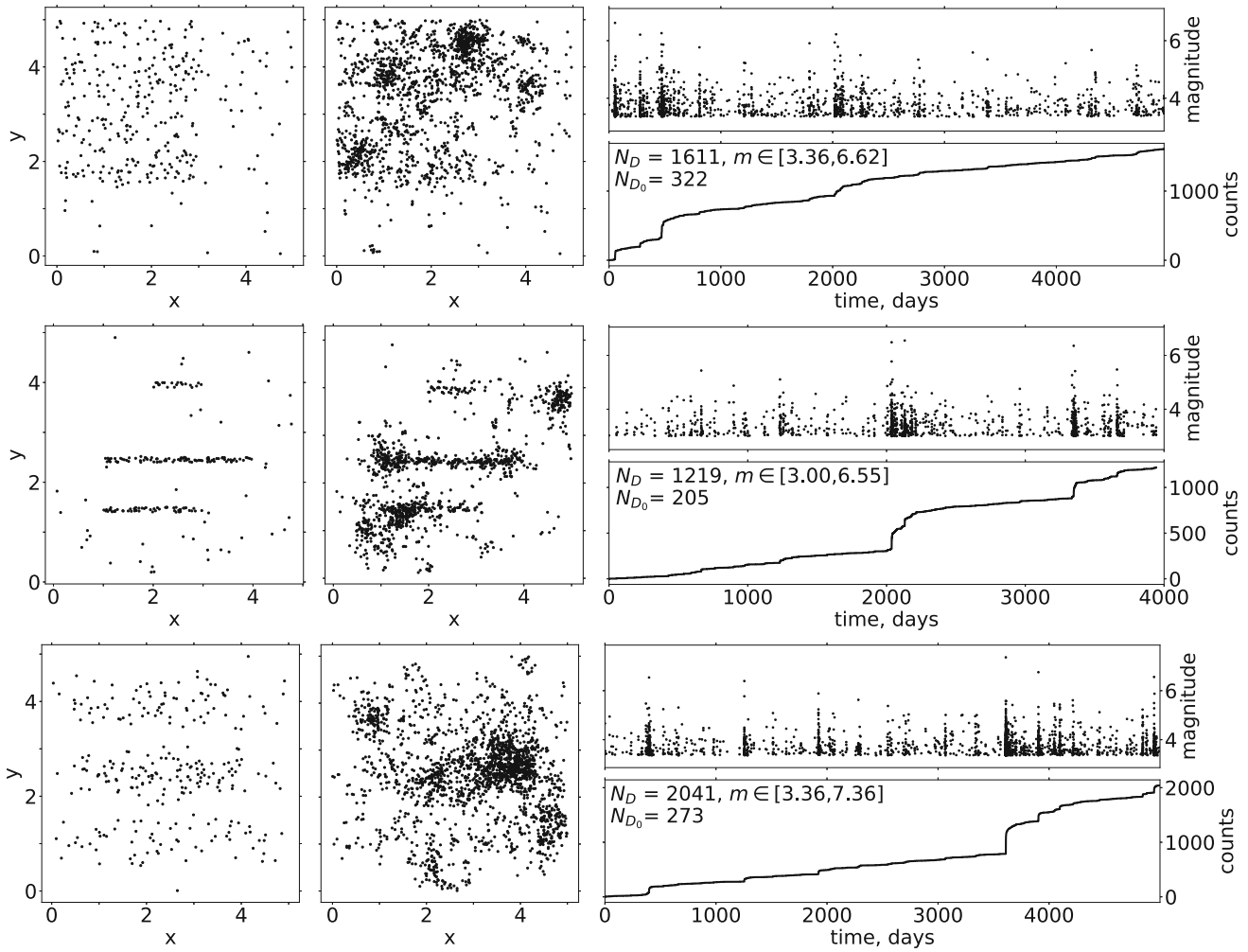

Fig. 2 Setup of synthetic data experiments, Case 1 (first row), Case 2 (second row) and Case 3 (third row): From left to right: background events, data set including background and offspring events, visualisation of the data as earthquake sequence over time

with a minimal bandwidth given by Silverman's rule (Silverman 1986) and $n_p = 15$ (ETAS–Silverman).

Evaluation metrics. Two metrics are used to evaluate the performances. The first metric is the test likelihood, which evaluates the likelihood (15) for a data test set $\mathcal{D}^*$ (unseen data during the inference) given the inferred model on training data $\mathcal{D}$, which is $p(\mathcal{D}^*|\mathcal{D}) = \mathbb{E}_{p(\mu,\boldsymbol{\theta}_\varphi|\mathcal{D})}\left[p(\mathcal{D}^*|\mu,\boldsymbol{\theta}_\varphi)\right]$, where the expectation is over the inferred model posterior. The test likelihood reflects the predictive power of the different modelling approaches. In the case of GP-ETAS we obtain K posterior samples $\{(\mu^k,\boldsymbol{\theta}_\varphi^k)\}_{k=1}^K$ and we evaluate the log expected test likelihood, $\ell_{\text{test}} = \ln p(\mathcal{D}^*|\mathcal{D}) \approx \ln \frac{1}{K}\sum_{k=1}^K p(\mathcal{D}^*|\mu^{(k)},\boldsymbol{\theta}^{(k)})$. In the case of ETAS–classical and ETAS–Silverman we use the MLE point estimate for evaluating ℓ_{test}. The involved spatial integral in (15) is approximated by Riemann sums on a 50×50 point grid. The second metric is the ℓ_2 norm between true background intensity μ and the predicted $\hat{\mu}$, $\ell_2 = \sqrt{\int_{\mathcal{X}}(\mu(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x}))^2\,\mathrm{d}\boldsymbol{x}}$. This is only possible for the experiments with synthetic data.

5.1 Synthetic data

General experimental setups. We simulate synthetic data from three different conditional intensity functions which differ in $\mu(\boldsymbol{x})$. In the first case we consider $\mu_1(\boldsymbol{x})$ to be constant over large spatial regions, e.g. large area sources (*Case 1*, Fig. 2 first row). In a second experiment we consider another particular setting where $\mu_2(\boldsymbol{x})$ is concentrated mainly on small fault-type areas (*Case 2*, Fig. 2 second row). These two settings (area sources and faults) are important, typical limiting cases in analysing seismicity pattern, both used in seismic hazard assessment. In addition, we consider a third Case (*Case 3*, Fig. 2 third row), where the transition between high and low background intensity is smooth

Table 1 GP-ETAS setup: prior choice

Variable, symbol	Prior	Values
Latent function f	Gaussian process prior	Zero mean function; Cov function (12)
Upper bound, $\bar{\lambda}$	Gamma(α_0, β_0)	$\alpha_0 = 1/c_{\bar{\lambda}}^2$, $\beta_0 = 1/c_{\bar{\lambda}}^2/\mu_{\bar{\lambda}}$
		With $c_{\bar{\lambda}} = 1$, $\mu_{\bar{\lambda}} = \frac{\max(n_1,\ldots,n_J)}{\lvert\mathcal{X}\rvert\lvert\mathcal{T}\rvert}$
Hyperparameters of cov function, $\boldsymbol{\nu}$	Exponential distribution	$\beta_{\nu_0} = 1/5$, $\beta_{\nu_1} = c_\nu \mathrm{d}x_1$, $\beta_{\nu_2} = c_\nu \mathrm{d}x_2$
		With $c_\nu = 1/25$, $\mathrm{d}x_i = x_{i,\max} - x_{i,\min}$
Parameters of the triggering function, $\boldsymbol{\theta}_\varphi$	Uniform distribution	$K_0 \in (0, 10)$, $c \in (0, 10)$,
		$p \in (0, 10)$, $\alpha \in (0, 10)$,
		$d \in (0, 10)$, $\gamma \in (0, 10)$,
		$q \in (1, 10)$

Table 2 GP-ETAS setup: control parameters of the Gibbs sampler

Parameter, symbol	Values
Number of posterior samples, K	5000
Burn-in (number of discarded initial iteration), B	2000 or 5000
MH proposal distribution for $\boldsymbol{\nu}$: Gaussian	$\sigma_{p_2} = 0.05$ in log units
MH proposal distribution for $\boldsymbol{\theta}_\varphi$: Gaussian	$\sigma_{p_1} = 0.01$ in log units
Number of MH steps per iteration for $\boldsymbol{\theta}_\varphi$	10

complementary to Case 1 and Case 2. The chosen intensity functions are,

$$\mu_1(\boldsymbol{x}) = \begin{cases} 0.005 & \boldsymbol{x} \in [0,3] \times [1.5,5] \\ 0.001 & \boldsymbol{x} \in [3,5] \times [1.5,5] \\ 0.0005 & \boldsymbol{x} \in [0,5] \times [0,1.5] \end{cases} \tag{29}$$

$$\mu_2(\boldsymbol{x}) = \begin{cases} 0.07035 & \boldsymbol{x} \in [1,3] \times [1.4,1.5] \\ 0.07035 & \boldsymbol{x} \in [1,4] \times [2.4,2.5] \\ 0.03535 & \boldsymbol{x} \in [2,3] \times [3.9,4] \\ 0.00035 & \text{else} \end{cases} \tag{30}$$

$$\mu_3(\boldsymbol{x}) = \frac{1}{100}\left\{ \exp\left(-\frac{(x_1-2.5)^2+(x_2-2.5)^2}{2s^2}\right) \times \cos^2\left(\frac{2\pi}{3}(x_1-2.5)\right) + 0.0001 \right\}, \tag{31}$$

with $s = 1.5$. The triggering function is given in (7–9) with spatial kernel (10) in all cases. The magnitudes are simulated following an exponential distribution $p_M(m_i) = \frac{1}{\beta} e^{-\beta(m_i - m_0)}$ with $\beta = \ln(10)$ which corresponds to a Gutenberg–Richter relation with b-value of 1; $m_0 = 3.36$ in the first and third case, $m_0 = 3$ in the second case. Spatio-temporal domain is $\mathcal{X} \times \mathcal{T} = [0,5] \times [0,5] \times [0,5000]$ for Case 1 and Case 3, and $\mathcal{X} \times \mathcal{T} = [0,5] \times [0,5] \times [0,4000]$ for Case 2. The test likelihood is computed for twelve unseen data sets, simulated from the generative model and averaged. The simulations are done on the same spatial domain $\mathcal{X}_{\text{sim}} = [0,5] \times [0,5]$. The time window is $\mathcal{T}_{\text{sim}} = [0, 1500]$ and ℓ_{test} is evaluated using events with event times $t_i \in [500, 1500]$, all previous events are taken into account in the history H_t.

GP-ETAS setup and model robustness. In GP-ETAS we need to set priors on $\bar{\lambda}$, $\boldsymbol{\nu}$, $\boldsymbol{\theta}_\varphi$. In addition, we have to choose control parameters of the Gibbs sampler: burn-in B; number of posterior samples K; number of MH steps n_{MH} and width of the proposal distribution σ_{p_1} when sampling the offspring parameters $\boldsymbol{\theta}_\varphi$; and the width of the proposal distribution σ_{p_2} of the MCMC sampling of the hyperparameters $\boldsymbol{\nu}$. In the following we give some practical guidance on the model setup and we test the robustness of the model using these specifications. A summary of the setup of GP-ETAS is given in Tables 1 and 2.

Hyperparameters $\boldsymbol{\nu}$ of the covariance function of the GP prior can strongly influence the inference results. We learn $\boldsymbol{\nu}$ from the data via MCMC sampling in GP-ETAS. We suggest exponentially distributed priors $p(\nu_i) = \beta_i e^{-\beta_i \nu_i}$ on $\boldsymbol{\nu} = (\nu_0, \nu_1, \nu_2)$, specified by the prior mean μ_{ν_i}. Such a prior choice is common practice for hyperparameter sampling in GP regression. We set $\mu_{\nu_i} = 1/\beta_i = c_\nu \mathrm{d}x_i$ for the length scales ν_1 and ν_2 , where $0.01 \leq c_\nu \approx 0.1 \leq 1$ is a factor, and $\mathrm{d}x_i = x_{i,\max} - x_{i,\min}$ is the maximum distance in dimension x_i. We choose $\mu_{\nu_0} = 5$ for the amplitude ν_0. Upper bound $\bar{\lambda}$ has a Gamma prior, which we specify via its mean $\mu_{\bar{\lambda}}$ and the uncertainty about this prior mean in terms of the coefficient of variation $c_{\bar{\lambda}}$. As default set up, we choose $c_{\bar{\lambda}} = 1$ and $\mu_{\bar{\lambda}} = \frac{\max(n_1,\ldots,n_J)}{\lvert\mathcal{X}\rvert\lvert\mathcal{T}\rvert}$ where $n_1, \ldots, n_J$ are the number of observations (counts) in spatial bins (2 dimensional histogram count of the positions of the events), and

$N_{\mathcal{D}} = \sum_{j=1}^{J} n_j$. We set $J = 100$ bins to cover the whole spatial domain $\mathcal{X}$ (which corresponds to a 10 by 10 grid). We use uniform priors on the offspring parameters $\boldsymbol{\theta}_\varphi$, given in Table 1. However, expert knowledge of the region of interest can be used to define more appropriate priors on $\boldsymbol{\nu}$, $\bar{\lambda}$ and $\boldsymbol{\theta}_\varphi$. Initial values $(\bar{\lambda}, \boldsymbol{\nu}, \boldsymbol{\theta}_\varphi)^{(0)}$ of the Gibbs sampler can be drawn from the priors. From ML inference of the classical ETAS model it is known that initial values can have an influence on the convergence of the ML optimization and on the obtained results (e.g., Veen and Schoenberg 2008). Unreasonable initial values of $\boldsymbol{\theta}_\varphi$ can prolongate burn-in phase of the GP-ETAS Gibbs sampler. We set following initial values in all our experiments: $\bar{\lambda}^{(0)} = N_{\mathcal{D}}/(2|\mathcal{X}||\mathcal{T}|)$ where $N_{\mathcal{D}}$ is the number of all observed events and $|\mathcal{X}|$, $|\mathcal{T}|$ are the size of the spatial and temporal domain, respectively; $(\nu_1)^{(0)}$ are initialized using Silverman's rule based on all observed events; $\nu_0^{(0)} = 5$; $\boldsymbol{\theta}_\varphi^{(0)} = (K_0, c, p, \alpha_m, d, \gamma, q)^{(0)} = (0.01, 0.01, 1.2, 2.3, \mathrm{d}x_1/100, 0.5, 2.)$ where $\mathrm{d}x_1 = x_{1,\max} - x_{1,\min}$ is the maximum distance in dimension x_1. We choose the control parameters of the Gibbs sampler (see Table 2) based on some pilot simulations. Burn in B is 2000 for Case 1 and Case 2 and $B = 5000$ for Case 3.

In order to test the robustness of our suggested Gaussian process modeling for the background intensity, we perform several synthetic data experiments additionally to those described above. First group of additional experiments: (1) we initialize the Gibbs sampler with a constant $\mu(\boldsymbol{x})^{(0)} = N_{\mathcal{D}}/(2|\mathcal{X}||\mathcal{T}|)$ and a very unlikely $\boldsymbol{\theta}_\varphi^{(0)}$ which leads to a sampling of a completely unrealistic branching structure. Here we investigate two cases: (1a) $\boldsymbol{\theta}_\varphi^{(0,a)}$ is set in such a way that only one observation is allocated to the background and all the other events are offspring (aftershocks); (1b) selecting a $\boldsymbol{\theta}_\varphi^{(0,b)}$ in such a manner that all events are considered to be background. Second group of additional experiments: (2) we scale the spatial domain for Case 1 by a factor of 100 in each dimension, i.e. observations are simulated in a spatial domain $\tilde{\mathcal{X}} = [0, 500] \times [0, 500]$, temporal domain $\mathcal{T}$ is kept as before. The generative model is as in Case 1 with two modifications. Background intensity $\mu_1(\boldsymbol{x})$ is scaled by a factor $\tilde{c} = |\mathcal{X}|/|\tilde{\mathcal{X}}|$ in order to get the same average number of background events on the scaled spatio-temporal window as in the unscaled case, that is, $\tilde{\mu}_1(\boldsymbol{x}) = \tilde{c}\mu_1(\boldsymbol{x})$, where $\tilde{\mu}_1(\boldsymbol{x})$ is the scaled background intensity. In addition, we adjust parameter d of the spatial triggering kernel in (20) in order to obtain a proportional spatial spread of the offspring events, $\tilde{d} = \tilde{c}d$. We employ the same set up of GP-ETAS as in Case 1 (see Tables 1, 2), in particular we use the same settings regarding the hyperparameters as in Case 1. Additionally, we utilize several initialization of the length scales ν_1, ν_2 with values ranging from 5 to 500. In both groups of experiments, i.e. (1) and (2), we test if the generative background model can be retrieved and we evaluate the performance measures ℓ_2 and ℓ_{test}.

Results of these additional experiments related to the robustness of GP-ETAS are shown in Appendix 4. The model of the background intensity underlying the data can be recovered in an acceptable way, in a sensible length of iterations in all experiments. GP-ETAS performs better than standard ETAS models for both metrics ℓ_2 and ℓ_{test}. Beyond the investigated cases there might be scenarios in which the model fails, but under reasonable assumptions (from application point of view) regarding priors, initial values, control parameters of the sampler and based on a sufficient amount of data, our proposed GP-ETAS seems to be robust.

Findings and interpretations. The ground truth and inferred results of $\mu(\boldsymbol{x})$ and $\boldsymbol{\theta}_\varphi$ are given in Figs. 3, 4 and 5 and Tables 3, 4 and 5. Performance metrics: the averaged ℓ_{test} of twelve unseen data sets and the numerically approximated error of the estimated background intensity ℓ_2 are shown in Table 6. Here we describe a few noteworthy aspects. First of all, GP-ETAS recovers well the assumptions both the background intensity $\mu(\boldsymbol{x})$ and the parameters of the triggering function $\boldsymbol{\theta}_\varphi$. GP-ETAS outperforms the standards models for both metrics ℓ_{test} and ℓ_2. The latter fact is of particular importance, as it is common practice to use the declustered background intensity $\mu(\boldsymbol{x})$ for seismic hazard assessment. One may appreciate that ETAS–classical occasionally tends to strongly overshoot the true $\mu(\boldsymbol{x})$ (see Fig. 3); in regions with many aftershocks, e.g. near (2.8,4.5), (1.2,3.8). This effect is less pronounced for ETAS–Silverman, where the minimum bandwidth is broader. In our approach no bandwidth selection has to be made in advance, it is obtained via sampling the hyperparameters. One also observes, that ETAS–classical and ETAS–Silverman suffer more strongly from edge effects than GP-ETAS, which seems to be fairly unaffected.

The parameters of the triggering function $\boldsymbol{\theta}_\varphi$ are roughly correctly identified in all cases and methods. All the values are close to those of the generative model. All the methods overestimate c and K_0 (in Case 2), however the true values are still included in the uncertainty band (credible band) of GP-ETAS. GP-ETAS has the advantage that it provides the whole distribution of the parameters instead of only a point estimate. The median of the obtained upper bound $\bar{\lambda}$ on $\mu(\boldsymbol{x})$ using GP-ETAS overestimates in Case 1 (underestimates in Case 2) the true upper bound, however, it is fairly close to the true value, which is contained in the uncertainty band around $\bar{\lambda}$. The median of $\bar{\lambda}$ is substantially underestimated in Case 3, here it is not contained in the uncertainty band.

Computational costs. GP-ETAS has approximately a complexity of $\mathcal{O}((N_{\mathcal{D}\cup\Pi})^3)$ for the inference of μ. This is due to the matrix inversions involved in Gaussian process modelling. The estimation of $\boldsymbol{\theta}_\varphi$ is less expensive and approx-

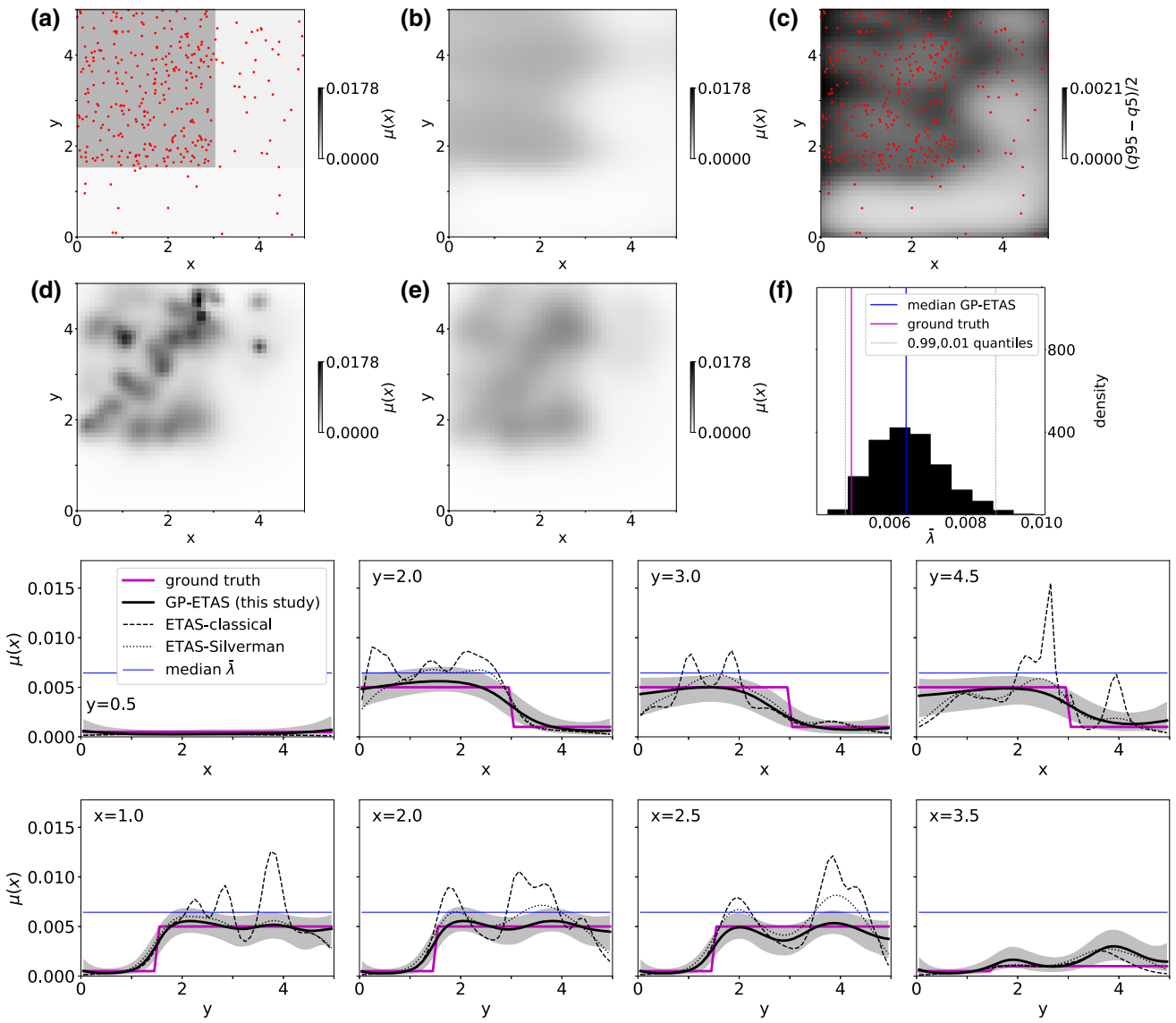

Fig. 3 Experimental results of background intensity $\mu_1(\boldsymbol{x})$ for the synthetic data of Case 1: *First and second row*: **a** generative model, **b** median GP-ETAS, **c** uncertainty GP-ETAS as semi inter quantile 0.05, 0.95 distance **d** ETAS–classical MLE, **e** ETAS–Silverman MLE, **f** normalised histogram of the sampled upper bound $\bar{\lambda}$. Dots are the background events of the realisation. *Third and fourth row*: One dimensional profiles of $\mu_1(\boldsymbol{x})$ (ground truth) and inferred results are shown. The profiles are at $y \in \{0.5, 2, 3, 4.5\}$ and $x \in \{1, 2, 2.5, 3.5\}$

imately of $\mathcal{O}(N_{\mathcal{D}}^2)$, where $N_{\mathcal{D}}$ is the number of data points. In addition, the number of required samples in order to obtain a valuable approximation of the posterior distribution depends on the mixing properties of the Markov chain. From our experience based on the performed experiments one needs $> 10^3$ samples after a burn-in phase of $> 10^3$ iterations. Therefore, our proposed method in its current implementation is computationally expensive but still feasible for small to intermediate data sets with approximately $N_{\mathcal{D}} \lessapprox 10^4$ events; which seems sufficient for many situations where site specific seismic analysis takes place. Run times of GP-ETAS model are given in Appendix 5.

5.2 Case study: L'Aquila, Italy

Now we apply GP-ETAS to real data and compare the performance with the other models ETAS–classical and ETAS–Silverman.

The L'Aquila region in central Italy is seismically active and experiences from time to time severe earthquakes. The most famous example is the $M_w = 6.2$ earthquake on 6 April 2009, which occurred directly below the City of L'Aquila, and caused large damage and more than 300 deaths (Marzocchi et al. 2014). This event was followed by a seismic sequence with a largest earthquake of $M_w = 4.2$, latter

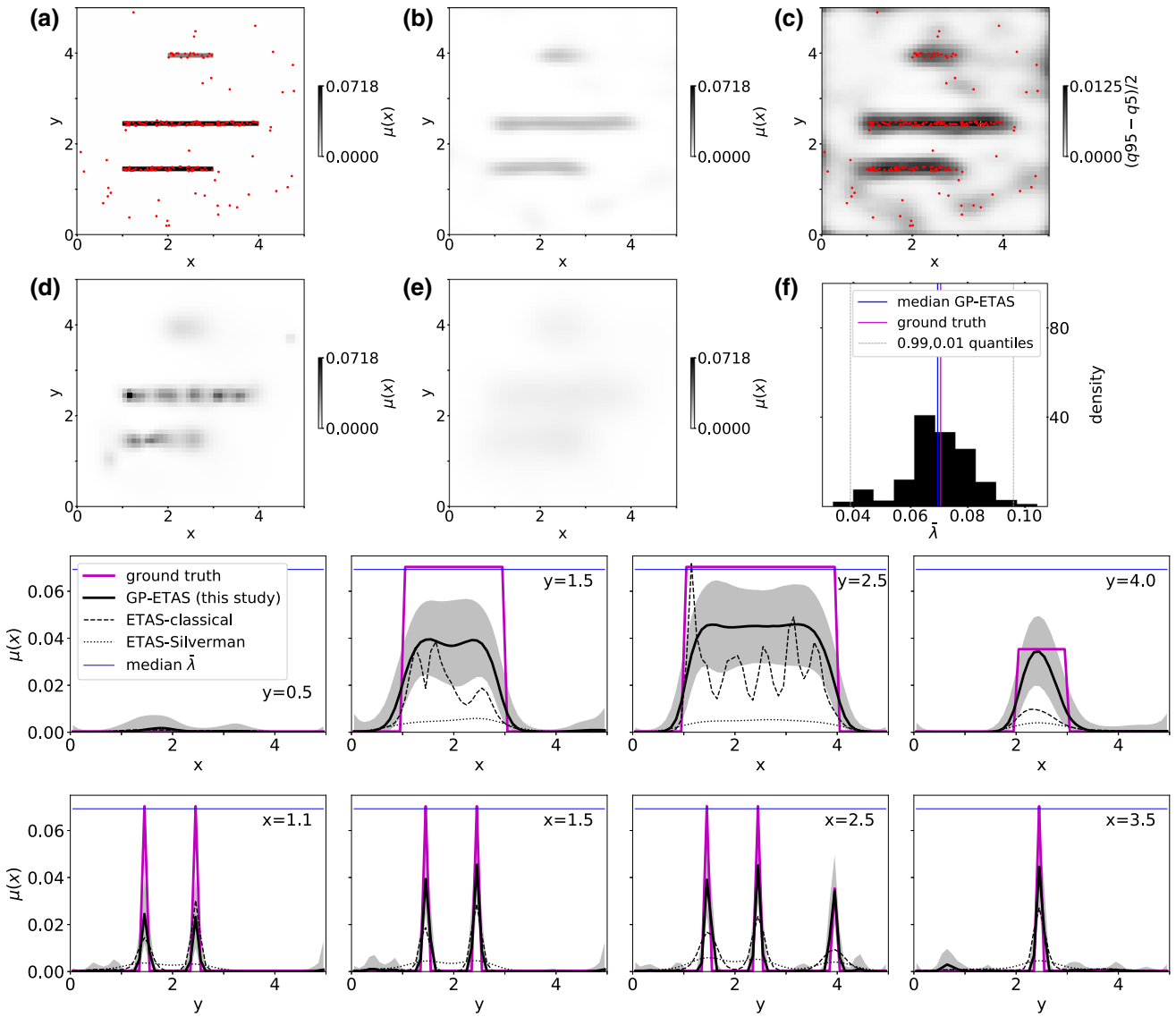

Fig. 4 Same as Fig. 3 but now results of background intensity $\mu_2(\boldsymbol{x})$ for the synthetic data of Case 2. See Fig. 3 for the description of plots and lines. The profiles are at $y \in \{0.5, 1.5, 2.5, 4\}$ and $x \in \{1.1, 1.5, 2.5, 3.5\}$

occurred almost one year later on 30 March 2010 (Marzocchi et al. 2014).

The L'Aquila data set comprises $N = 2189$ events which occurred in a time period from 04/02/2001 to the 28/3/2020, on a spatial domain $\mathcal{X} = [12°E, 15°E] \times [41°N, 44°N]$ with earthquake magnitudes $3.0 \leq m \leq 6.5$. The data was obtained from the website of the National Institute of Geophysics and Vulcanology of Italy (http://terremoti.ingv.it/, Istituto Nazionale della Geofisica e Vulcanologia, INGV). We split the data set into training data, all events with event times $t_i \leq 4000$ days ($N_{\text{training}} = 723$ events, $\mathcal{T}_{\text{training}} = [0, 4000]$ days), and *test data*, all events with $t_i > 4000$ days ($N_{\text{test}} = 1466$, $\mathcal{T}_{\text{test}} = [4000, 6993]$ days), as shown in Fig. 6. The training data is used for the inference and the test data is used to evaluate the performance of the different models.

The inference setup is the same as described for the synthetic data. We simulate 15,000 posterior samples after a burn in of 2000. The priors are—as for the synthetic data—given in Tables 1 and 2 . The inference results for $\mu(\boldsymbol{x})$ and $\boldsymbol{\theta}_{\varphi}$ are shown in Fig. 7 and Table 7. The performance metric ℓ_{test} for different unseen data sets (next 30 days, one year in future, five years in future, and the total test data $\approx$ 8 years) is shown in Table 8. The posterior distribution of the upper bound of the background intensity is shown in Fig. 8.

For the l'Aquila data set GP-ETAS performs slightly better than the other models in terms of ℓ_{test}. Note, that ETAS–classical estimates fairly large values for $\mu(\boldsymbol{x})$ in regions with many aftershocks (Fig. 7). This is similar to Case 1 for

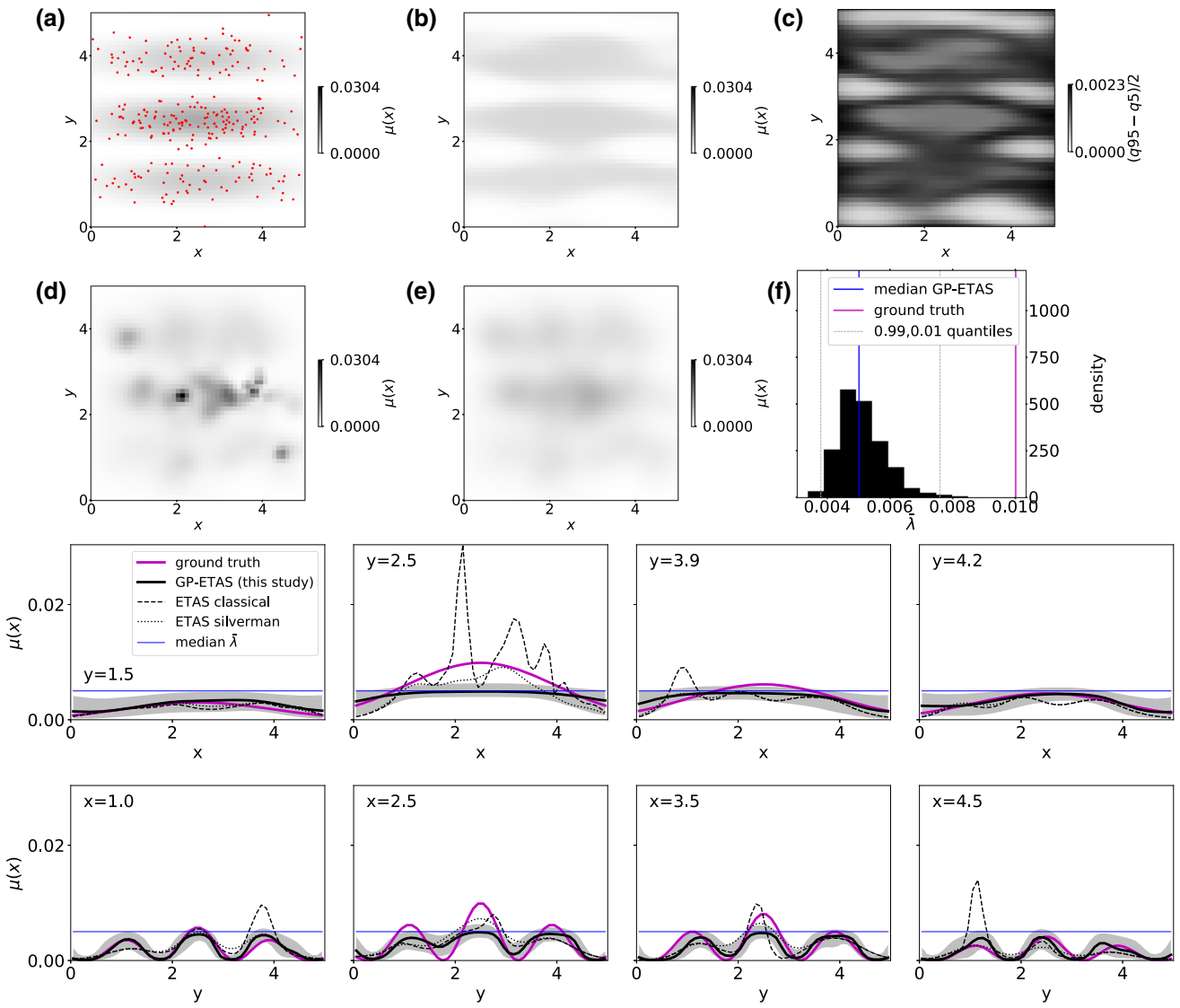

Fig. 5 Same as Fig. 3 but now results of background intensity $\mu_3(\boldsymbol{x})$ for the synthetic data of Case 3. See Fig. 3 for the description of plots and lines. The profiles are at $y \in \{1.5, 2.5, 3.9, 4.2\}$ and $x \in \{1, 2.5, 3.5, 4.5\}$

Table 3 True and inferred parameter values $\boldsymbol{\theta}_\varphi = [K_0, c, p, \alpha, d, \gamma, q]$ of the triggering function of Case 1: gM is generative model (true values of the forward model); E is classical ETAS model and E–S is ETAS Silverman; GP-E is GP-ETAS model (this study) with first line median, second line 0.05 quantile, last line 0.95 quantile of the sampled posterior distribution

Model	K_0	c	p	α	d	γ	q
gM	0.0180	0.0060	1.20	1.690	0.015	0.20	2.00
E	0.0179	0.0075	1.23	1.668	0.018	0.18	2.06
E–S	0.0183	0.0068	1.21	1.668	0.018	0.18	2.07
	0.0184	0.0068	1.21	1.662	0.017	0.19	2.07
GP–E	0.0164	0.0056	1.19	1.595	0.014	0.17	1.93
	0.0203	0.0085	1.24	1.734	0.022	0.21	2.23

Table 4 Same as Table 3 but now for results of Case 2

Model	K_0	c	p	α	d	γ	q
gM	0.0180	0.0060	1.20	1.690	0.015	0.20	2.00
E	0.0192	0.0082	1.21	1.648	0.015	0.20	2.07
E–S	0.0213	0.0058	1.15	1.618	0.015	0.20	2.10
	0.0195	0.0083	1.21	1.645	0.014	0.20	2.00
GP–E	0.0175	0.0063	1.19	1.585	0.011	0.18	1.89
	0.0216	0.0104	1.24	1.699	0.018	0.22	2.16

Table 5 Same as Table 3 but now for results of Case 3

Model	K_0	c	p	α	d	γ	q
gM	0.0180	0.0060	1.20	1.690	0.015	0.20	2.00
E	0.0173	0.0064	1.22	1.693	0.019	0.19	2.25
E–S	0.0176	0.0061	1.21	1.690	0.019	0.19	2.24
	0.0180	0.0062	1.20	1.685	0.020	0.19	2.23
GP–E	0.0165	0.0049	1.18	1.635	0.016	0.18	2.09
	0.0196	0.0074	1.23	1.731	0.022	0.20	2.39

Table 6 Performance comparison between different inference methods: GP-ETAS (GP-E, this study), MLE of the classical ETAS model (E) and ETAS Silverman (E–S). The comparison is done based on two performance measures: averaged test-likelihood ℓ_{test} of twelve unseen data sets (higher = less negative is better) and ℓ_2 norm to the true background intensity (smaller is better)

Experiment	Measure	gM	E	E–S	GP-E
Case 1	ℓ_{test}	−342.1	−352.2	−349.3	−345.1
Case 2	ℓ_{test}	−30.3	−70.6	−93.4	−45.4
Case 3	ℓ_{test}	−277.9	−285.6	−284.4	−280.7
Case 1	ℓ_2		0.0096	0.0060	0.0038
Case 2	ℓ_2		0.0390	0.0482	0.0236
Case 3	ℓ_2		0.0089	0.0068	0.0059

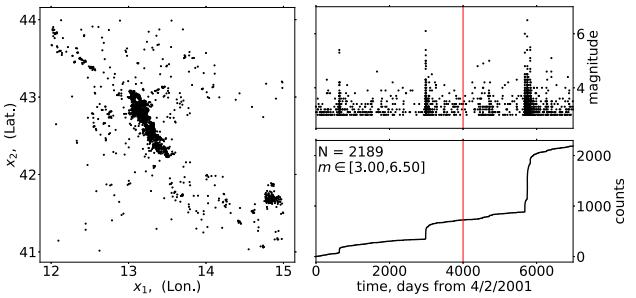

Fig. 6 Earthquake data from central Italy: epicentre plot (left) and visualisation of the data as earthquake sequence over time (right)

the synthetic experiments. Hence, as for the synthetics one may assume that ETAS–classical overshoots in these regions; the posterior of $\bar{\lambda}$ supports this hypothesis, see Fig. 8. The estimated $\boldsymbol{\theta}_\varphi$ are similar for all models. Posterior samples and MLE estimates of d, γ, q are shown in Fig. 9. The spatial kernels of the three models are shown in Fig. 10 for the mean magnitude and a large magnitude.

6 Discussion and conclusions

We have demonstrated that the proposed GP-ETAS model allows for augmentation techniques (Hawkes and Oakes 1974; Veen and Schoenberg 2008; Adams et al. 2009; Donner and Opper 2018) and hence provides the means of

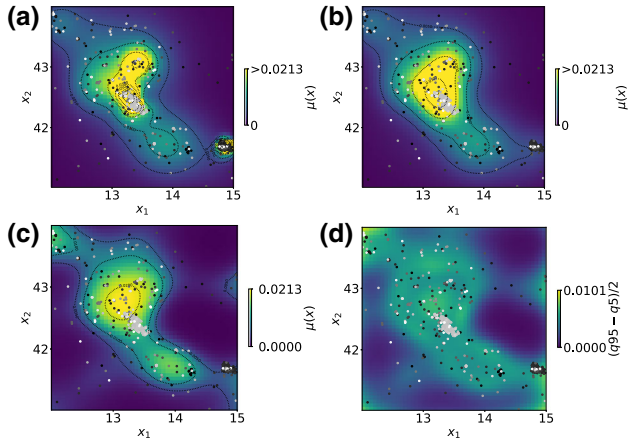

Fig. 7 Results real data, L'Aquila data set: background intensity $\mu(\boldsymbol{x})$ [number of shocks with $m \geq 3$ /day/degree^2] **a** ETAS–classical MLE, **b** ETAS–Silverman MLE, **c** median GP-ETAS, **d** uncertainty GP-ETAS: semi inter quantile 0.05, 0.95 distance, and dots are the events of the training data, where the grey scaling depicts the event times, from black (older events) to white (current events). Note, **a–c** have the same scale

Table 7 L'Aquila data set: inferred parameter values $\boldsymbol{\theta}_\varphi = [K_0, c, p, \alpha, d, \gamma, q]$ of the triggering function; E is classical ETAS model and E–S is ETAS Silverman; GP-E is GP-ETAS model (this study) with first line median, second line 0.05 quantile, last line 0.95 quantile of the sampled posterior distribution

Model	K_0	c	p	α	d	γ	q
E	0.0272	0.0387	1.20	1.813	0.0042	0.20	2.65
E–S	0.0274	0.0305	1.16	1.782	0.0043	0.19	2.50
	0.0269	0.0276	1.16	1.780	0.0044	0.19	2.57
GP-E	0.0224	0.0164	1.11	1.660	0.0029	0.16	2.17
	0.0321	0.0451	1.20	1.887	0.0063	0.23	3.27

Table 8 Test likelihood ℓ_{test} of unseen test data sets

Testing period	N_{test}	ETAS-classical	ETAS-Silverman	GP-ETAS
30 days	2	−13.0	−12.6	−12.3
1 year	18	−78.7	−76.9	−74.7
5 years	1116	5007.7	5013.2	5090.1
Total test period (≈ 8.2 years)	1466	5677.3	5679.5	5769.4

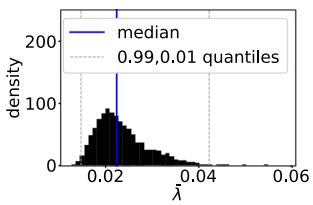

Fig. 8 Normalised histogram of the sampled posterior of the upper bound $\bar{\lambda}$

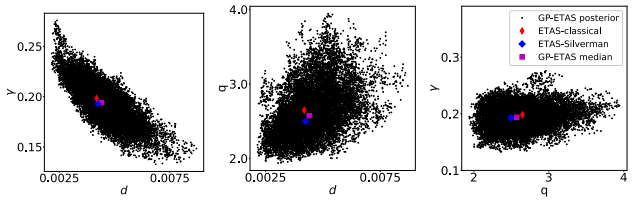

Fig. 9 Scatter plot of the posterior samples of d, γ, q

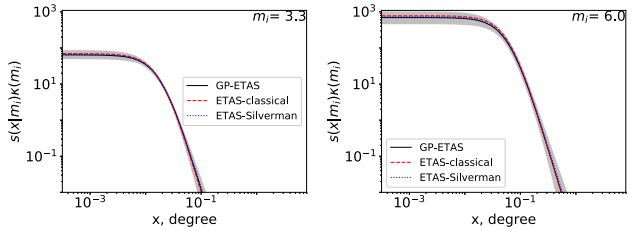

Fig. 10 Estimated spatial kernel (10) multiplied by the productivity (8) for the mean magnitude $m = 3.3$ and for a large $m = 6$ mainshock at position (0,0). Shaded area is 0.05, 0.95 percentile of GP-ETAS

assessing the Bayesian posterior of a semi-parametric spatio-temporal ETAS model. We have shown for three examples that the predictive performance improves over classical methods. In addition, we can quantify parameter uncertainties via their empirical posterior density. The developed framework is flexible and allows for several extensions that deserve consideration in future research, e.g. a time depending on background rate. Another obvious extension of our work would be a Bayesian non-parametric treatment of the triggering function φ. See also (Zhang et al. 2019b).

As mentioned earlier Kolev and Ross (2020) propose an alternative semiparametric framework of the ETAS model, which assumes a Dirichlet process prior over the background intensity. Such a model is more related to the classical KDE approach than the model proposed in our work. It considers, that the background intensity is modelled via density over scaled mixture models. We expect, that inference of the Dirichlet process will be less costly than the for the GP-ETAS model. In contrast, to define the GP-prior seems more intuitive, because one mainly needs to impose some spatial correlation structure, while for the Dirichlet prior we need to have an idea, how many components model well the background intensity function on average. However, which of the two models will be the more suitable one, might depend on the specific data instance.

Future research will deal with a more geology informed choice of the prior GP. Sometimes a given catalog comes with information about, e.g. fault locations, which are not straightforward to incorporate in traditional treatments of the spatio-temporal ETAS model. For the GP-ETAS model, however, incorporation of informative priors is possible within our Bayesian setting. For example, spatial information about fault zones can be incorporated by an adequate choice of the mean of the GP, which was chosen to be 0 throughout this work. While we restricted ourselves to the squared exponential (12) as covariance function for the GP prior of f the framework is not restricted to this either and other covariance function can be used to incorporate prior information, e.g. from the Matérn class, or any other function that ensures that the covariance matrix is positive definite.

Another important issue is the computational effort. Having to sample the GP at all observed events in $\mathcal{D}$ and at positions of the latent Poisson process Π, resulting in a cubic complexity of $\mathcal{O}\left((N_{\mathcal{D}\cup\Pi})^3\right)$, implies an undesirable computational complexity of the current GP-ETAS Gibbs sampler. There are, however, several possibilities to mitigate this complexity via model approximations and/or alteration. For example, one could resort to approximations to the posterior distribution in order to be able to scale the GP-ETAS model to larger catalogues ($N_D \gg 10^3$). While those always come with the sacrifice of asymptotic exactness, some approaches are likely to provide good estimates in the large data regime. One of such approximations is provided by variational inference, which was already proposed for the SGCP by Donner and Opper (2018) utilising sparse GPs (Titsias 2009). This approach makes use of the same model augmentations utilised in this work. The variational posterior of the triggering parameters could be inferred, e.g., via black-box variational inference (Ranganath et al. 2014; Gianniotis et al. 2015). Alternatively, one could restrict the calculations to finding the MAP estimate of the GP-ETAS model. For the background intensity this can be efficiently done by an expectation-maximisation algorithm based on the model augmentations presented here and sparse GPs (Donner and Opper 2018). This can be combined with a Laplace approximation to provide an approximate Gaussian posterior. The limiting factor under such approximations will most likely arise from the branching structure for which the required computations scale like $\mathcal{O}(N_{\mathcal{D}}(N_{\mathcal{D}} - 1)/2)$. Finally, one could also investigate gradient-free affine invariant sampling methods, as proposed by Reich and Weissmann (2019); Garbuno-Inigo et al. (2020, 2019).

We conclude by re-emphasising the importance of semi-parametric Bayesian approaches to spatio-temporal statistical earthquake modelling and the need for developing efficient tools for their computational inference. Within this work, we combined the SGCP model (Adams et al. 2009) for the background intensity with the ETAS model, such that a Gibbs sampling approach could be made tractable and efficient via specific data augmentation. Finally, we have demonstrated the model's applicability to realistic earthquake catalogs.

Acknowledgements This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG, German Science Foundation) - SFB 1294/1 - 318763901.

Appendices

A GP-ETAS generative model

The generative model of GP-ETAS consists of two parts given in Algorithm 2–3, and requires several inputs. Here, we chose a GP with zero mean and covariance function $k(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\nu})$ given in (12) with hyperparameters $\boldsymbol{\nu}$, a triggering function $\varphi(\cdot|\boldsymbol{\theta}_\varphi)$ given in (7–9) with spatial kernel (10) and therefore $\boldsymbol{\theta}_\varphi = (K_0, c, p, \alpha, d, \gamma, q)$, and a mark distribution, that is an exponential distribution $m_i - m_0 \sim \text{Exponential}(\beta)$ (Gutenberg–Richter relation) with parameters β, m_0. The simulation algorithm can be easily adjusted for other choices.

B Definition of the Pólya–Gamma density

Here, we briefly define the Pólya–Gamma density (Polson et al. 2013). First we define the $p_{\text{PG}}(\omega|b, 0)$, which is completely defined through its Laplace transform

$$\int_0^\infty e^{-\omega t} p_{\text{PG}}(\omega|b, 0) d\omega = \cosh^{-b}(\sqrt{t/2}). \tag{32}$$

With this definition it can be shown that

$$\omega \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2}, \tag{33}$$

where $g_k \sim \text{Gamma}(b, 1)$ and the equality is in distribution. With this result one can then define a *tilted* Pólya–Gamma density given by

$$p_{\text{PG}}(\omega|b, c) \propto e^{-\frac{c^2}{2}\omega} p_{\text{PG}}(\omega|b, 0), \tag{34}$$

where $b \in \mathbb{R}^+$ and $c \in \mathbb{R}$, and the normalisation can be straightforwardly obtained with (32). Also for this the tilted density we can derive the Laplace transform

$$\int_0^\infty e^{-\omega t} p_{\text{PG}}(\omega|b, c) d\omega = \frac{\cosh^b(c/2)}{\cosh^b\left(\frac{\sqrt{c^2/2+t}}{2}\right)}. \tag{35}$$

From (35) all moments of the Pólya–Gamma density can be derived analytically and furthermore an acceptance-rejection algorithm with high acceptance rate was derived by Polson et al. (2013).

C Conditional posteriors for the background intensity

Here we derive the conditional posterior distributions given in (25a)–(25e). For each of those our starting point is the augmented likelihood of the background intensity in (24) with the prior of interest. Note, that for the augmented variables $\Pi, \boldsymbol{\omega}_\mathcal{D}, \boldsymbol{\omega}_\Pi$ there are no additional priors, and hence their conditionally distribution will only be determined by (24).

The latent Poisson process Π. To derive the conditional posterior for Π, we consider all terms in (24) that depend on $\Pi = \{x_l\}_{l=1+N_\mathcal{D}}^{N_{\mathcal{D}\cup\Pi^{(k)}}}$ and marginalise over the $\boldsymbol{\omega}_\Pi$. This results in

$$p(\Pi|\bar{\lambda}, \boldsymbol{f}) \propto \prod_{l=N_\mathcal{D}+1}^{N_{\mathcal{D}\cup\Pi}} \bar{\lambda}\sigma(-f_l) \exp\left(-|\mathcal{T}| \int_\mathcal{X} \bar{\lambda}\sigma(-f(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}\right), \tag{36}$$

which we identify as an unnormalised Poisson process density with rate $\bar{\lambda}\sigma(-f_l)$. Again note, that the process is defined over $\mathcal{T} \times \mathcal{X}$. To sample $\Pi^{(k)}$ in the k^{th} iteration, we can utilise the *thinning* procedure (Lewis and Shedler 1976), where we first sample a homogeneous point process with intensity $\bar{\lambda}^{(k-1)}$. At the resulting points we draw the the GP f given the previous sample $\boldsymbol{f}^{(k-1)}$ from the predictive distribution (see below). Then, we keep all events l with probability $\sigma(-f_l)$ which yields $\Pi^{(k)}$ according to (25a).

The Pólya–Gamma random variables $\boldsymbol{\omega}$. Next we sample the Pólya–Gamma random variables at $\mathcal{D}_0$. From (24) we see, that the components in $\boldsymbol{\omega}_{\mathcal{D}_0}$ factorise, meaning that the conditional posteriors are independent. Hence, we get for each $i : z_i = 0$

$$p(\omega_i|\boldsymbol{f}, \mathcal{D}, Z) \propto e^{-\frac{(f_i)^2}{2}\omega_i} p_{\text{PG}}(\omega_i|1, 0) \propto p_{\text{PG}}(\omega_i|1, |f_i|), \tag{37}$$

and hence $\boldsymbol{\omega}_\mathcal{D}^{(k)}$ can be sampled independently from a tilted Pólya–Gamma distribution where $i : z_i = 0$, given branching structure $Z^{(k)}$ and GP $\boldsymbol{f}^{(k-1)}$. For $i : z_i \neq 0$ we set $\omega_i = 0$. The last equivalence follows from a property of an

Algorithm 2 : GP-ETAS generative model Part 1: background events $\mathcal{D}_0$

Input: Spatio-temporal domain $\mathcal{X} \times \mathcal{T}$; GP mean function and covariance function with hyperparameters $\boldsymbol{\nu}$; upper bound $\bar{\lambda}$; parameters of the mark density β, m_0

Output: Background events $\mathcal{D}_0 = \{t_i, x_i, m_i, z_i = 0\}_{i=1}^{N_{\mathcal{D}_0}}$

$N \sim \text{Poisson}(\bar{\lambda}|\mathcal{X}||\mathcal{T}|)$ ▷ Sample number of candidate events from Poisson distribution
$\{\boldsymbol{x}_i\}_{i=1}^{N} \sim \mathcal{U}(\mathcal{X})$ ▷ Distribute candidate events uniformly in $\mathcal{X}$
$\{f(\boldsymbol{x}_i)\}_{i=1}^{N} \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{K}_{\boldsymbol{\nu}})$ ▷ Draw function values from the GP, $\boldsymbol{K}_{\boldsymbol{\nu}} = \{k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\nu})\}_{i,j}^{N}$
$\mathcal{D}_0 \leftarrow \emptyset$ ▷ Initialise the set of background events $\mathcal{D}_0$
$N_{\mathcal{D}_0} \leftarrow 0$ ▷ Initialise number of background events
for $i \leftarrow 1, ..., N$ **do** ▷ *Thinning* procedure
  $r_i \sim \mathcal{U}(0, \bar{\lambda})$ ▷ Draw a uniform random variable on the interval $[0, \bar{\lambda}]$
  if $r_i < \bar{\lambda}\sigma(f(\boldsymbol{x}_i))$ **then** ▷ Acceptance criteria
    $\boldsymbol{x}_i$ is accepted
    $t_i \sim \mathcal{U}(\mathcal{T})$ ▷ Distribute background event uniformly in $\mathcal{T}$
    $m_i - m_0 \sim \text{Exponential}(\beta)$ ▷ Sample a mark
    $z_i \leftarrow 0$ ▷ Assign branching variable z_i, index of parent event
    $\mathcal{D}_0 \leftarrow \mathcal{D}_0 \cup \{t_i, \boldsymbol{x}_i, m_i, z_i\}$ ▷ Add event to accepted background events $\mathcal{D}_0$
    $N_{\mathcal{D}_0} \leftarrow N_{\mathcal{D}_0} + 1$ ▷ Count number of background events
  end if
end for
Sort $\mathcal{D}_0$ by event times t_i
return $\mathcal{D}_0$

Algorithm 3 : GP-ETAS generative model Part 2: offspring events

Input: $\mathcal{D}_0$, $N_{\mathcal{D}_0}$ from Algorithm (2); spatio-temporal domain $\mathcal{X} \times \mathcal{T}$; triggering function $\varphi(\cdot)$ with defining parameters $\boldsymbol{\theta}_{\varphi} = (K_0, c, p, \alpha, d, \gamma, q)$; parameters of the mark distribution β, m_0

Output: Background events $\mathcal{D} = \{t_i, x_i, m_i, z_i\}_{i=1}^{N_{\mathcal{D}}}$

$\mathcal{D} \leftarrow \mathcal{D}_0$ ▷ Initialise the set of simulated events $\mathcal{D}$ with the background events $\mathcal{D}_0$
$N_{\mathcal{D}} \leftarrow N_{S_0}$ ▷ Initialise number of simulated events with the number of background events
$j \leftarrow 1$ ▷ Initialise the index of potential parent event
while $j < N_{\mathcal{D}}$ **do** ▷ Consider all events in $\mathcal{D}$ for producing potentially offspring
  $\{t_j, \boldsymbol{x}_j, m_j, z_j\} \leftarrow \mathcal{D}[j]$ ▷ Obtain entries of the jth simulated event in $\mathcal{D}$
  $\lambda_{\text{max},j} \leftarrow \max(\kappa(m_j)g(t - t_j)) = Ke^{\alpha(m_j - m_0)}c^{-p}$ ▷ Get upper bound $\lambda_{\text{max},j}$ of jth PP
  $|\mathcal{T}_j| \leftarrow |\mathcal{T}| - t_j$ ▷ Compute the size of the time window $|\mathcal{T}_j|$ of direct offspring
  $N_j \leftarrow \text{Poisson}(\lambda_{\text{max},j}|\mathcal{T}_j|)$ ▷ Sample number of candidate offspring events
  if $N_j > 0$ **then** ▷ Check if there are candidate offspring events
    $\{t_i\}_{i=1}^{N_j} \sim \mathcal{U}(0, |\mathcal{T}_j|) + t_j$ ▷ Distribute candidate events uniformly in $[t_j, t_{max}]$
    for $i \leftarrow 1, ..., N_j$ **do** ▷ *Thinning* procedure
      $r_i \sim \mathcal{U}(0, \lambda_{\text{max},j})$ ▷ Draw a uniform random variable on the interval $[0, \lambda_{\text{max},j}]$
      if $r_i < \lambda_{\text{max},j}(t_i - t_j + c)^{-p}$ **then** ▷ Acceptance criteria using (8,9)
        t_i is accepted
        $\boldsymbol{x}_i \sim s(\boldsymbol{x} - \boldsymbol{x}_j|m_j, d, \gamma, q)$ ▷ Sample position of the offspring event (10)
        $m_i - m_0 \sim \text{Exponential}(\beta)$ ▷ Sample marks, see above
        $z_i \leftarrow j$ ▷ Assign branching variable z_i, index of parent event
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{t_i, \boldsymbol{x}_i, m_i, z_i\}$ ▷ Add offspring event to $\mathcal{D}$
        $N_{\mathcal{D}} \leftarrow N_{\mathcal{D}} + 1$ ▷ Count number of simulated events D
      end if
    end for
    $j \leftarrow j + 1$ ▷ Advance to next possible parent event
  end if
end while
Sort $\mathcal{D}$ by event times t_i, while taking care of properly mapping the branching variables
return $\mathcal{D}$

tilted Pólya–Gamma distribution, see (34). In effect, we get (25c).

The conditional posterior $\boldsymbol{\omega}_\Pi$ at the positions of the latent events Π also factorises in all components and hence we get for each $l = N_\mathcal{D} + 1, \ldots, N_{\mathcal{D}\cup\Pi}$

$$p(\omega_l|\boldsymbol{f}, \Pi) \propto e^{-\frac{(f_l)^2}{2}\omega_l} p_{\text{PG}}(\omega_l|1, 0) \propto p_{\text{PG}}(\omega_l|1, |f_l|), \tag{38}$$

and also $\boldsymbol{\omega}_\Pi^{(k)}$ can be sampled independently from a tilted Pólya–Gamma distribution (25b) given the samples of $\Pi^{(k)}$ and $\boldsymbol{f}^{(k-1)}$; we get (25b).

The upper bound on the intensity $\bar{\lambda}$. For $\bar{\lambda}$ we assume a Gamma prior $p(\bar{\lambda}|\alpha_0, \beta_0)$ with shape parameter α_0 and rate parameter β_0. Together with (24) we derive the conditional posterior being,

$$\begin{aligned} p(\bar{\lambda}|\mathcal{D}_0, \Pi, Z) &\propto \bar{\lambda}^{N_{\mathcal{D}_0\cup\Pi}} e^{-\bar{\lambda}|\mathcal{X}||\mathcal{T}|} p(\bar{\lambda}|\alpha_0, \beta_0) \\ &\propto \text{Gamma}(\bar{\lambda}|\alpha_1, \beta_1). \end{aligned} \tag{39}$$

where $\alpha_1 = N_{\mathcal{D}_0\cup\Pi} + \alpha_0$ and $\beta_1 = |\mathcal{T}||\mathcal{X}| + \beta_0$. Hence, given $\Pi^{(k)}$ and $D_0^{(k)}$ we can sample $\bar{\lambda}^{(k)}$, and one gets (25d).

The posterior Gaussian Process f. For the conditional posterior of the Gaussian process f we rewrite (24) with the terms depending on f as follows

$$\begin{aligned} &p(\mathcal{D}_0, \boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi|\boldsymbol{f}, \bar{\lambda}, Z) \\ &\propto \prod_{i:z_i=0}^{N_\mathcal{D}} e^{f_i u_i - \frac{f_i^2}{2}\omega_i} \prod_{i:z_i\neq 0}^{N_\mathcal{D}} e^{f_i u_i - \frac{f_i^2}{2}\omega_i} \prod_{l=N_\mathcal{D}+1}^{N_{\mathcal{D}\cup\Pi}} e^{f_l u_l - \frac{f_l^2}{2}\omega_l} \\ &= e^{-\frac{1}{2}\boldsymbol{f}^\top \boldsymbol{\Omega} \boldsymbol{f} + \boldsymbol{u}^\top \boldsymbol{f}}, \end{aligned} \tag{40}$$

where we define $u_i = \frac{1}{2}$ if $z_i = 0$, $u_i = 0$ if $z_i \neq 0$, and $u_l = -\frac{1}{2}$ for $l = N_\mathcal{D} + 1, \ldots, N_{\mathcal{D}\cup\Pi}$. It follows that $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega}_\mathcal{D}, \boldsymbol{\omega}_\Pi)$. The GP prior $\boldsymbol{f}$ at a finite set $\mathcal{D}, \Pi$ of points is given by

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}) \tag{41}$$

where the entry of row i and column j of $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$ is given by the covariance function (12) $k(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\nu})$. Together with (40) we identify the conditional posterior

$$p(\boldsymbol{f}|\mathcal{D}, \boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi, Z) \propto e^{-\frac{1}{2}\boldsymbol{f}^\top \boldsymbol{\Omega} \boldsymbol{f} + \boldsymbol{u}^\top \boldsymbol{f}} \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}) \tag{42a}$$

$$\propto \mathcal{N}\left(\boldsymbol{f}|\left[\boldsymbol{\Omega} + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}^{-1}\right]^{-1}\boldsymbol{u}, \left[\boldsymbol{\Omega} + \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}^{-1}\right]^{-1}\right), \tag{42b}$$

which defines the conditional posterior at $\boldsymbol{f}$, and it is easy to sample $\boldsymbol{f}^{(k)}$ given instances of $\boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi$ from previous samples. However, the algorithm requires us to sample the posterior also at other points $\boldsymbol{x}^* \notin \mathcal{D} \cup \Pi$, e.g. for sampling the next instance of Π or for visualisation of the background intensity $\mu(\boldsymbol{x})$ on a grid. Here, we denote the GP at all additional points $\boldsymbol{f}^*$. The GP prior defines a predictive distribution (Williams and Rasmussen 2006) of any set $\boldsymbol{f}^*$ given $\boldsymbol{f}$

$$p(\boldsymbol{f}^*|\boldsymbol{f}, \boldsymbol{\nu}) = \mathcal{N}\left(\boldsymbol{f}^*|\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}^{-1}\boldsymbol{f}, \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*} - \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}^{-1}\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}^*}\right), \tag{43}$$

where $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}^*}$ contains the covariances between the points $\boldsymbol{x}^*$ of $\boldsymbol{f}^*$ and $\boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}} = \boldsymbol{K}_{\boldsymbol{f}^*,\boldsymbol{f}}^\top$ between the points $\boldsymbol{x}^*$ of $\boldsymbol{f}^*$ and $\boldsymbol{f}$ at $\boldsymbol{x}$. Note, that the posterior of $\boldsymbol{f}^*$ given $\boldsymbol{f}$ is equal to the conditional prior, since (24) does not depend on $\boldsymbol{f}^*$.

D Additional experiments

Here, we report a summary of the results related to additional synthetic data experiments described in paragraph 'GP-ETAS set up and model robustness' of Sect. 5.1 in the main text.

Results of experiment 1. We investigate two experiments: (1a) an initialization of the Gibbs sampler with almost only offspring events and just one background event; (1b) vice versa an initialization of the Gibbs sampler with almost only background events and no or only a few offspring events. In order to obtain such extreme realizations of the branching structure (1a and 1b), we choose initial offspring parameters $\boldsymbol{\theta}_\varphi^{(0)} = (K_0, c, p, \alpha_m, d, \gamma, q)$ as follows: Experiment (1a) $\boldsymbol{\theta}_\varphi^{(0,\text{a})} = (10., 0.01, 1.2, 10., 0.05, 0.5, 2.)$; Experiment (1b) $\boldsymbol{\theta}_\varphi^{(0,\text{b})} = (1\text{e}-06, 0.01, 1.2, 2.3, 0.05, 0.5, 2.)$; initialization of the background intensity is $\mu(\boldsymbol{x})^{(0)} = N_\mathcal{D}/(2|\mathcal{X}||\mathcal{T}|)$ in both experiments. Underlying generative models are those of Case 1, Case 2 and Case 3.

The Gibbs sampler converges after a sensible number of iterations (< 2000) to the ground truth in terms of the number of events assigned to the background process $\mu(\boldsymbol{x})$, as depicted in Fig. 11. The generative models, i.e. background intensities $\mu_1(\boldsymbol{x})$, $\mu_2(\boldsymbol{x})$ and $\mu_3(\boldsymbol{x})$ together with the offspring parameters $\boldsymbol{\theta}_\varphi$ can be recovered similar to the results in the main text. Here, we only report the average test-likelihood ℓ_{test} of 12 unseen data sets and ℓ_2 norm between true background intensity and predicted $\hat{\mu}$, see Table 9. GP-ETAS performs better in Experiment (1a) and (1b) compared with standard ETAS models.

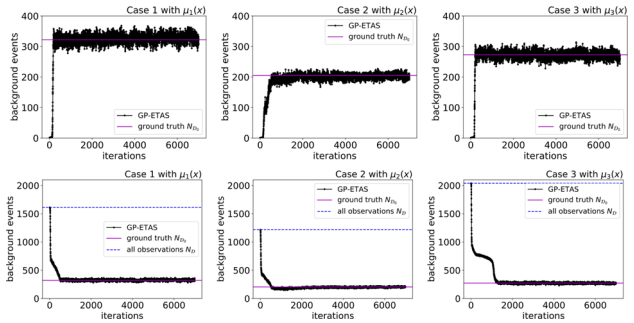

Fig. 11 Number of events assigned to the background process $\mu(\boldsymbol{x})$ in each iteration of the GP-ETAS Gibbs sampler. First to third column corresponds to Case 1, Case 2 and Case 3, as described in Sect. 5 of the main text. *First row* Experiment (1a), only one event is assigned as background event at the first iteration. *Second row* Experiment (1b), almost all events are assigned to the background process at the first iteration. After several iterations (< 2000), the Gibbs sampler converges to the true number of background events given by a horizontal, solid, magenta line

Table 9 Results of additional experiments (1a) and (1b): Comparison on ℓ_2 norm to the true background intensity (smaller is better), and averaged test-likelihood ℓ_{test} (higher = less negative is better) of twelve unseen data sets. C1, C2, C3 denote Case 1 to 3; ETAS stands for the ETAS–classical, ETAS-S for ETAS–Silverman, and GP-E stands for GP–ETAS

		ETAS	ETAS–S	GP-E (1a)	GP-E (1b)
C1	ℓ_2	9.614E−03	5.952E−03	3.304E−03	3.918E−03
C2	ℓ_2	3.901E−02	4.824E−02	2.283E−02	2.940E−02
C3	ℓ_2	8.893E−03	6.766E−03	6.361E−03	5.848E−03
C1	ℓ_{test}	−352.2	−349.3	−344.8	−345.4
C2	ℓ_{test}	−70.6	−93.4	−45.2	−49.7
C3	ℓ_{test}	−285.6	−284.4	−281.4	−282.0

Table 10 Results of additional experiment (2): performance measures ℓ_2 (lower is better) and ℓ_{test} (higher=less negative is better) of GP-ETAS compared to standard ETAS models; $(\nu_1, \nu_2)^{(0)}$ refers to initial values of the length scales

Model	$(\nu_1, \nu_2)^{(0)}$	ℓ_2	ℓ_{test}
Generative model		0	−2225.6
ETAS-classical		9.61E−05	−2235.8
ETAS-Silverman		5.95E−05	−2232.8
GP-ETAS	(5,5)	5.12E-05	−2231.3
	(20,20)	4.10E−05	−2229.9
	(50,50)	3.76E−05	−2228.7
	(100,100)	3.93E−05	−2228.9
	(500,500)	3.91E−05	−2228.9

Results of experiment 2. Inference results of Experiment (2) are summarized in Table 10. Here, synthetic data is simulated on a scaled spatial domain $\tilde{\mathcal{X}} = [0, 500] \times [0, 500]$. The background intensity model $\tilde{\mu}_1(\boldsymbol{x})$ underlying the simulated data can be retrieved in an acceptable way for all different initial values of the length scales ν_1, ν_2. Offspring parameters $\boldsymbol{\theta}_\varphi$ can also be retrieved in an acceptable way. We only report following metrics ℓ_2 (error in estimation the background intensity) and ℓ_{test} (test-likelihood of unseen data), which are described in the main text (Sect. 5), see Table 10. GP-ETAS has a lower error (ℓ_2) of the estimated background intensity and a higher predictive power (ℓ_{test}) compared with standard ETAS models.

Table 11 Mean and standard deviation of run time per sample iteration for all data sets

Data set	Run time per sample [s]
Case 1	4.96 ± 2.51
Case 2	47.81 ± 24.56
Case 3	2.22 ± 0.77
Aquila data	5.61 ± 4.40

E Run times

Run times of GP-ETAS are given in Table 11. All computations have been done on a iMac Pro 2017 with 32 GB RAM and a processor with 2.3 GHz.

References

Adams, R.P., Murray, I., MacKay, D.J.: Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp 9–16 (2009)

Adelfio, G., Chiodi, M.: Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. Stoch. Environ. Res. Risk Assess. **29**(2), 443–450 (2014)

Bacry, E., Muzy, J.F.: First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. IEEE Trans. Inf. Theory **62**, 2184–2202 (2016)

Bacry, E., Mastromatteo, I., Muzy, J.F.: Hawkes processes in finance. Mark. Microstruct. Liq. **1**(01), 1550005 (2015)

Console, R., Murru, M., Lombardi, A.M.: Refining earthquake clustering models. J. Geophys. Res. Solid Earth **108**(B10), 1–9 (2003)

Daley, D., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods, 2nd edn. Springer, New York (2003)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) **39**(1), 1–22 (1977)

Donner, C., Opper, M.: Efficient Bayesian inference of sigmoidal Gaussian cox processes. J. Mach. Learn. Res. **19**(1998), 1–34 (2018)

Donnet, S., Rivoirard, V., Rousseau, J.: Nonparametric Bayesian estimation of multivariate Hawkes processes. arXiv preprint arXiv:1802.05975 (2018)

Filimonov, V., Sornette, D.: Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to

high-frequency financial data. Quant. Finance **15**(8), 1293–1314 (2015)
Fox, E.W., Schoenberg, F.P., Gordon, J.S.: Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. Ann. Appl. Stat. **10**(3), 1725–1756 (2016)
Garbuno-Inigo, A., Nüsken, N., Reich, S.: Affine invariant interacting Langevin dynamics for Bayesian inference. Technical report. arXiv:1912.02859, SIAM J. Dyn. Syst. (**in press**) (2019)
Garbuno-Inigo, A., Hoffmann, F., Li, W., Stuart, A.: Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. SIAM J. Appl. Dyn. Syst. **19**, 412–441 (2020)
Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI **6**(6), 721–741 (1984)
Gerhard, F., Deger, M., Truccolo, W.: On the stability and dynamics of stochastic spiking neuron models: nonlinear Hawkes process and point process GLMs. PLoS Comput. Biol. **13**(2), 1–31 (2017)
Gianniotis, N., Schnörr, C., Molkenthin, C., Bora, S.S.: Approximate variational inference based on a finite sample of Gaussian latent variables. Pattern Anal. Appl. **19**(2), 475–485 (2015)
Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
Hawkes, A.G., Oakes, D.: A cluster process representation of a self-exciting process. J. Appl. Probab. **11**(3), 493–503 (1974)
Hu, W., Jin, P.J.: An adaptive Hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data. Transp. Res. Part C Emerg. Technol. **79**, 136–155 (2017)
Jordan, T.H., Chen, Y.T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., Zschau, J.: Operational earthquake forecasting. State of knowledge and guidelines for utilization. Ann. Geophys. 54(4), 315–391 (2011)
Kagan, Y.Y.: Aftershock zone scaling. Bull. Seismol. Soc. Am. **92**(2), 641–655 (2002)
Kingman, J.F.C.: Poisson Processes. Oxford University Press, Oxford (1993)
Kirchner, M.: An estimation procedure for the Hawkes process. Quant. Finance **17**(4), 571–595 (2017)
Kirichenko, A., Van Zanten, H.: Optimality of Poisson processes intensity learning with Gaussian processes. J. Mach. Learn. Res. **16**(1), 2909–2919 (2015)
Kolev, A.A., Ross, G.J.: Semiparametric Bayesian forecasting of spatial earthquake occurrences. arXiv preprint arXiv:2002.01706 (2020)
Lewis, P.A., Shedler, G.S.: Simulation of nonhomogeneous Poisson processes with log linear rate function. Biometrika **63**(3), 501–505 (1976)
Linderman, S.W., Adams, R.P.: Scalable Bayesian inference for excitatory point process networks. arXiv preprint arXiv:1507.03228 (2015)
Lippiello, E., Giacco, F., de Arcangelis, L., Marzocchi, W., Godano, C.: Parameter estimation in the ETAS model: approximations and novel methods. Bull. Seismol. Soc. Am. **104**(2), 985–994 (2014)
Lombardi, A.M.: Estimation of the parameters of ETAS models by simulated annealing. Sci. Rep. **5**(8417), 1–11 (2015)
Marsan, D., Lengliné, O.: Extending earthquakes' reach through cascading. Science **319**(5866), 1076–1079 (2008)
Marzocchi, W., Lombardi, A.M., Casarotti, E.: The establishment of an operational earthquake forecasting system in Italy. Seismol. Res. Lett. **85**(5), 961–969 (2014)
Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. J. Am. Stat. Assoc. **106**(493), 100–108 (2011)
Murray, I., Ghahramani, Z., MacKay, D.J.: MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006, pp. 359–366 (2006)
Ogata, Y.: Asymptotic behavoir of maximum likelihood. Ann. Inst. Stat. Math. **30**, 243–261 (1978)
Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Stat. Assoc. **83**, 9–27 (1988)
Ogata, Y.: Space-time point-process models for earthquake occurrences. Ann. Inst. Stat. Math. **50**(2), 379–402 (1998)
Ogata, Y., Zhuang, J.: Space-time ETAS models and an improved extension. Tectonophysics **413**(1–2), 13–23 (2006)
Omori, F.: On the after-shocks of earthquakes. J. Coll. Sci. **7**, 111–120 (1894)
Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pólya–Gamma latent variables. J. Am. Stat. Assoc. **108**(504), 1339–1349 (2013)
Porter, M.D., White, G.: Self-exciting hurdle models for terrorist activity. Ann. Appl. Stat. **4**(1), 106–124 (2010)
Ranganath, R., Gerrish, S., Blei, D.M.: Black box variational inference. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (2014)
Rasmussen, J.G.: Bayesian inference for Hawkes processes. Methodol. Comput. Appl. Probab. **15**(3), 623–642 (2013)
Rathbun, S.L.: Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. J. Stat. Plan. Inference **51**(1), 55–74 (1996)
Reich, S., Weissmann, S.: Fokker–Planck particle systems for Bayesian inference: computational approaches. Technical report. arXiv:1911.10832, University of Potsdam (2019)
Reynaud-Bouret, P., Schbath, S.: Adaptive estimation for Hawkes processes; application to genome analysis. Ann. Stat. **38**(5), 2781–2822 (2010)
Ross, G.: Bayesian estimation of the ETAS model for earthquake occurrences. Preprint (2018)
Schoenberg, F.P.: Facilitated estimation of ETAS. Bull. Seismol. Soc. Am. **103**(1), 601–605 (2013)
Silverman, B.W.: Density Estimation for Statistics and Data Analysis, vol. 26. CRC Press, Boca Raton (1986)
Titsias, M: Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics, pp 567–574 (2009)
Utsu, T.: A statistical study on the occurrence of aftershocks. Geophys. Mag. **30**, 521–605 (1961)
Utsu, T.: Aftershocks and earthquake statistics (1): some parameters which characterize an aftershock sequence and their interrelations. J. Fac. Sci. Hokkaido Univ. Ser. 7 Geophy. **3**(3), 129–195 (1970)
Veen, A., Schoenberg, F.P.: Estimation of space-time branching process models in seismology using an EM-type algorithm. J. Am. Stat. Assoc. **103**(482), 614–624 (2008)
Wang, Q., Schoenberg, F.P., Jackson, D.D.: Standard errors of parameter estimates in the ETAS model. Bull. Seismol. Soc. Am. **100**(5 A), 1989–2001 (2010)
Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., Opper, M.: Efficient gaussian process classification using Pòlya–Gamma data augmentation. Proc. AAAI Conf. Artif. Intell. **33**, 5417–5424 (2019)
Williams, C.K., Rasmussen, C.E.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA (2006)
Windle, J., Polson, N.G., Scott, J.G.: Sampling Pólya–Gamma random variates: alternate and approximate techniques. arXiv preprint arXiv:1405.0506 (2014)
Zhang, R., Walder, C., Rizoiu, M.A.: Variational inference for sparse Gaussian process modulated Hawkes process. arXiv preprint arXiv:1905.10496 (2019a)

Zhang, R., Walder, C., Rizoiu, M.A., Xie, L.: Efficient non-parametric Bayesian Hawkes processes. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 4299–4305 (2019b)

Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: SEISMIC: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1513–1522 (2015)

Zhou, K., Zha, H., Song, L.: Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 641–649 (2013)

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., Chen, F.: Scalable inference for nonparametric Hawkes process using Pólya–Gamma augmentation. arXiv:1910.13052v1 (2019)

Zhuang, J.: Next-day earthquake forecasts for the Japan region generated by the ETAS model. Earth Planets Space **63**(3), 207–216 (2011)

Zhuang, J., Ogata, Y., Vere-Jones, D.: Stochastic declustering of space-time earthquake occurrences. J. Am. Stat. Assoc. **97**(458), 369–380 (2002)