PGR_submission_Stockman_Sam_1503332.pdf

by Sam Stockman

Submission date: 01-Oct-2024 01:39PM (UTC+0100)
Submission ID: 240014320
File name: PGR_submission_Stockman_Sam_1503332.pdf (19.77M)
Word count: 46592
Character count: 254985

Enhancing Earthquake Forecasting:

Machine Learning Applications in Point Process Models

By

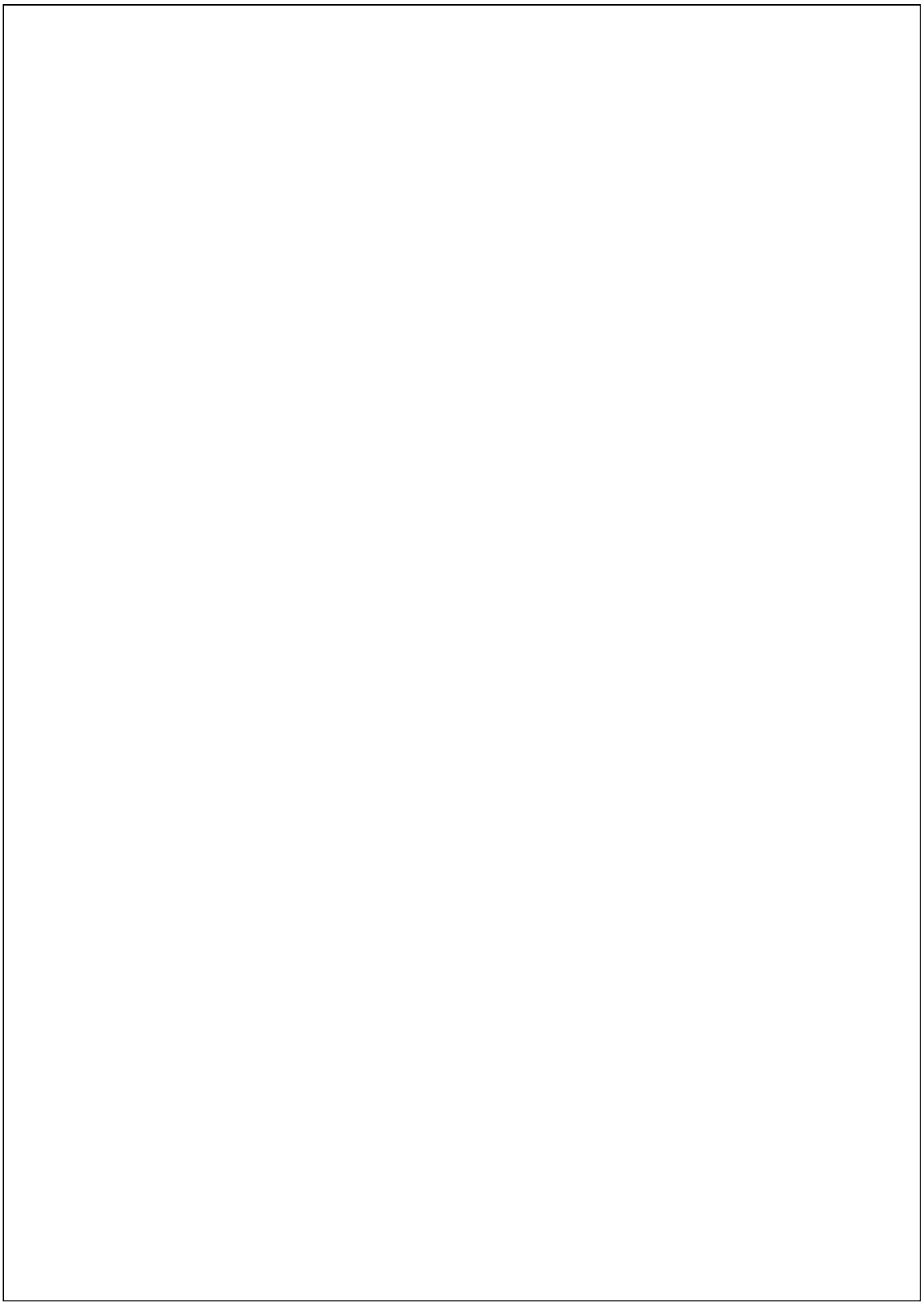
Samuel Stockman

School of Mathematics
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science.

October 2024

Word count: 26,852



Abstract

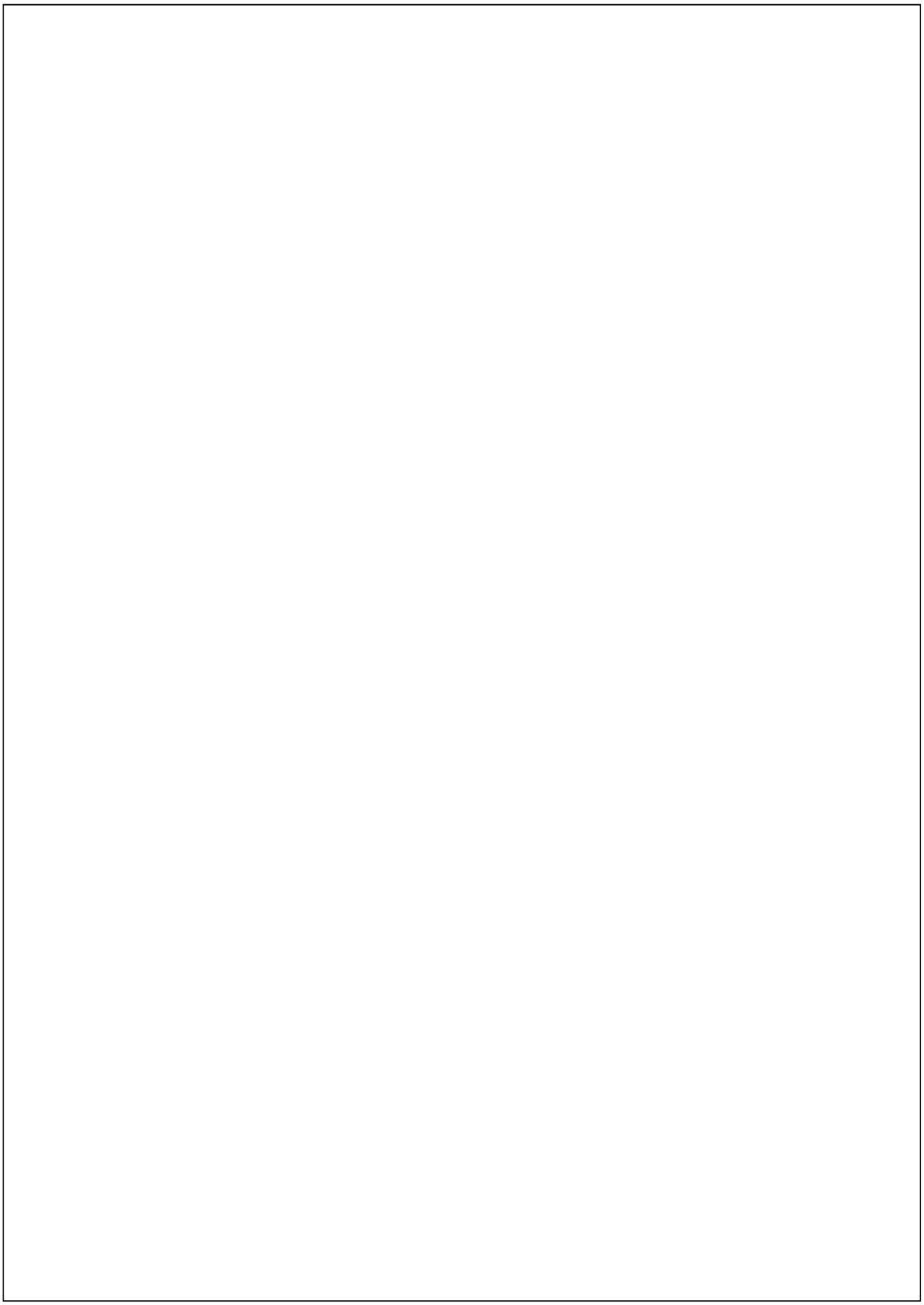
Short-term earthquake forecasting provides crucial hazard information during aftershock sequences. State-of-the-art models, such as the Epidemic-Type Aftershock Sequence (ETAS) model, are formulated as point processes - statistical models that represent earthquakes as points in time and space. Advances in earthquake detection now allow for the recording of much smaller magnitude events, leading to new earthquake catalogs containing a ten-fold increase in the number of recorded earthquakes. This volume of new data presents both opportunities and challenges for existing forecasting models, as well as calling for the exploration of innovative modeling approaches. This thesis explores the utility of Neural Point Processes (NPPs), a machine learning variant of point processes, to enhance earthquake forecasting capabilities.

I begin by extending an existing temporal NPP to the magnitude domain, adapting it to forecast earthquakes above a target magnitude threshold while depending on smaller magnitude events. I apply this model to a catalog of the Central Apennines earthquake sequence in Italy, demonstrating significant information gain over ETAS at the low magnitude thresholds of this enhanced catalog.

Next, I apply Simulation Based Inference (SBI) to Bayesian parameter estimation for the ETAS model, improving the scalability of inference from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$ with the number of earthquakes. By specifying a model through simulation rather than the likelihood, SBI broadens the scope of available models to encompass greater complexity. This would enable inference of earthquake models with intractable likelihoods, including data incompleteness and physics based simulators.

Finally, I develop EarthquakeNPP, a benchmarking platform for evaluating NPPs against state-of-the-art models from the seismology community. This platform provides benchmark datasets from California, along with a widely accepted implementation of the ETAS model, making these resources accessible to the machine learning community. The platform highlights the potential of NPPs and outlines a road-map for future implementations to provide more impact in earthquake forecasting.

By bridging the gap between statistical machine learning and seismology, this thesis provides a foundation for future interdisciplinary research, offering valuable insights for researchers in both fields.



Dedication and acknowledgements

Navigating the interdisciplinary boundary between Seismology and Statistics, I have at times felt like a "master of none." Yet, I have loved the interdisciplinary nature of my work. Successful interdisciplinary research relies on the insights of individuals from diverse backgrounds, and without their contributions, this work would not have been possible.

No more is this reflected than in the support I've received from my two supervisors, Max Werner and Dan Lawson, over the last four years. Whilst their experience and respective domain knowledge were an essential foundation for our work together, I am most grateful for their kindness, which made our meetings so enjoyable, and their mentoring which made me feel truly supported and gives me confidence well beyond my PhD.

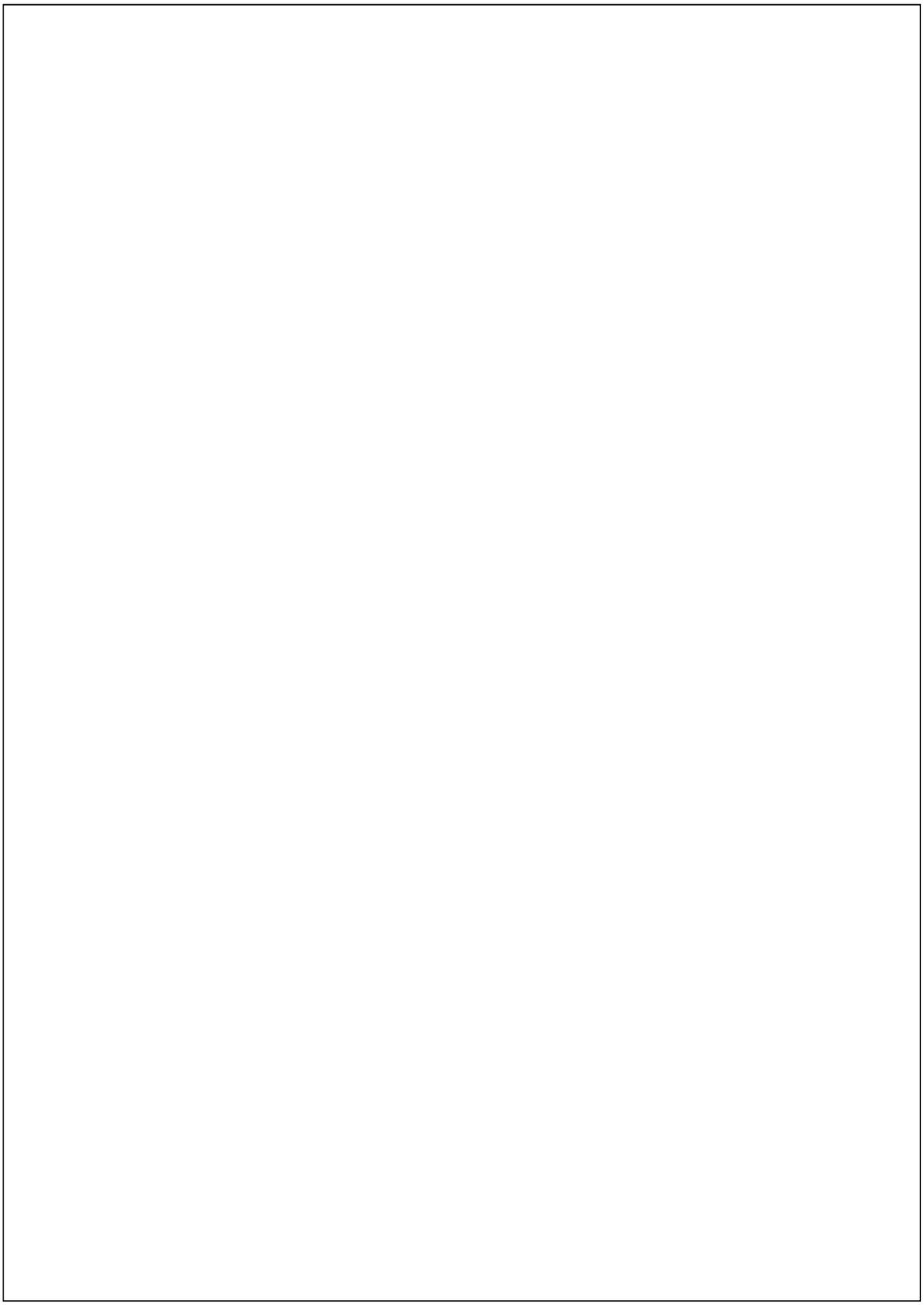
I have completed this thesis as part of the COMPASS center for doctoral training. This community of PhD students has provided me with laughter and friendship, giving me a reason to go to the office every day. The sense of community would not have been possible without Crina Radu and Harriet Lee, who worked tirelessly to support our welfare over the years.

I also feel incredibly lucky to have been welcomed into a warm, exciting and diverse community of seismologists over the course of my PhD. Either as part of Max's research group or at international conferences such as StatSei, I met brilliant fellow researchers, who inspired me and motivated me to work towards this thesis.

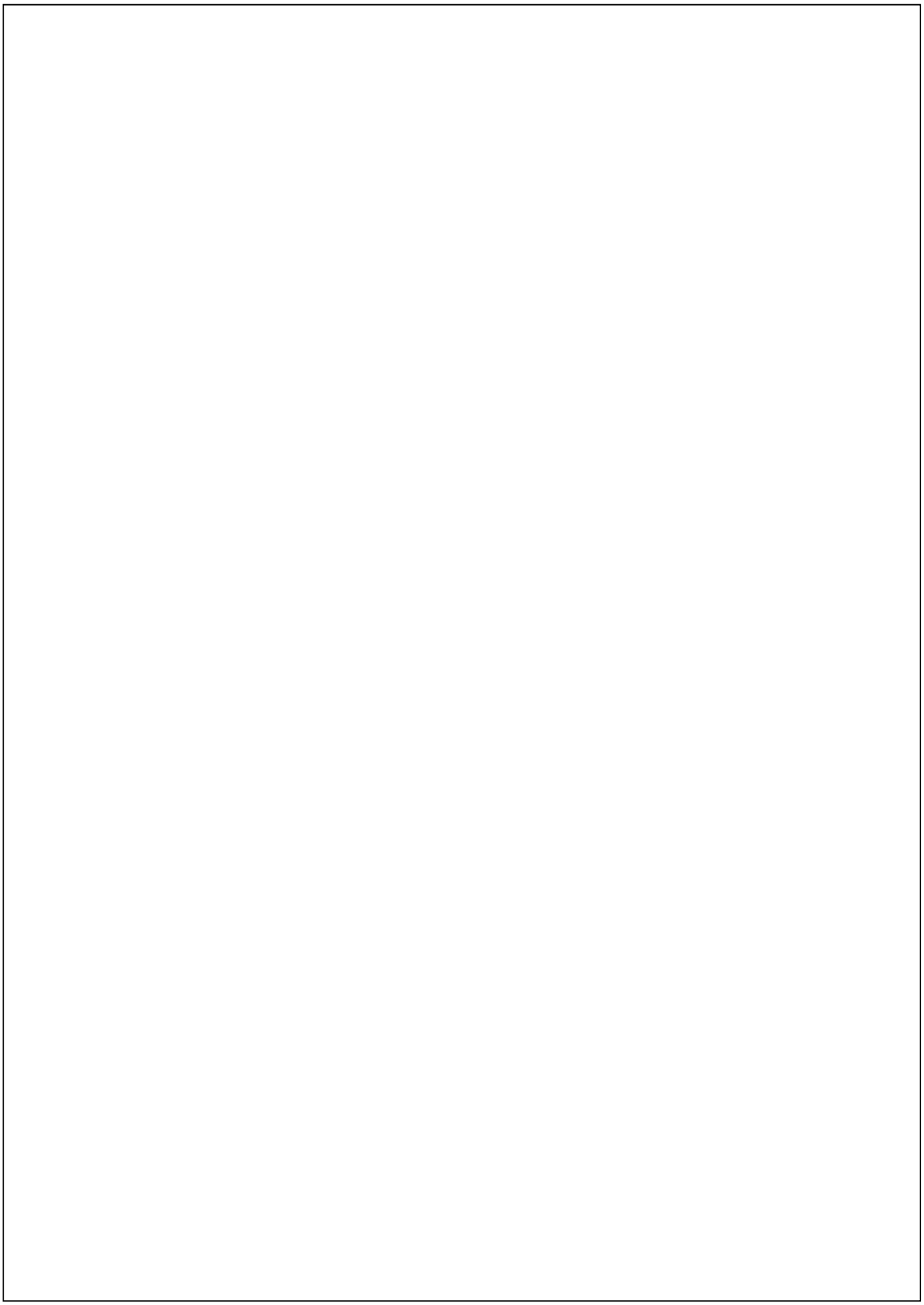
My journey to completing this thesis was also made possible by my dear friends: Charlie Barker, for his loyalty and all the crying with laughter we've done since Miss Sharkey's Maths lessons; Jack Simons, for embarking on this PhD journey with me, sharing all the highs and lows - I cannot wait to set off on our bike tour this weekend (05/10/24); Rebecca Vincent, Matt Clifford, Josh Cooper-Thorne and Joe Barker, with whom I have so many happy memories of living together in 21 Mogg Street; as well as the many other friends who continue to enrich my life.

I feel incredibly grateful to have a family who have supported me throughout my PhD. The greatest impact has come from my partner, Kayla Ellis. She has been with me through every moment, encouraging and believing in me every day - I'm so excited for whatever awaits us next. I am also incredibly grateful to James and Leonie Ellis for truly making Bristol feel like my home and to my Grandmother, who has believed in me since the day I was born.

I could never have begun a PhD without the uncountable sacrifices my parents have made for me. Their love and support has never stopped, and for this, I feel immensely grateful.



This thesis is dedicated to my sister, Sarah Stockman,
and to my friend, Sam Fitzsimmons, both of whom
passed away before I began my PhD.
I hope this thesis reflects their deep curiosity for the
world.



Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....SAMUEL STOCKMAN..... DATE: .....01/10/2024.....

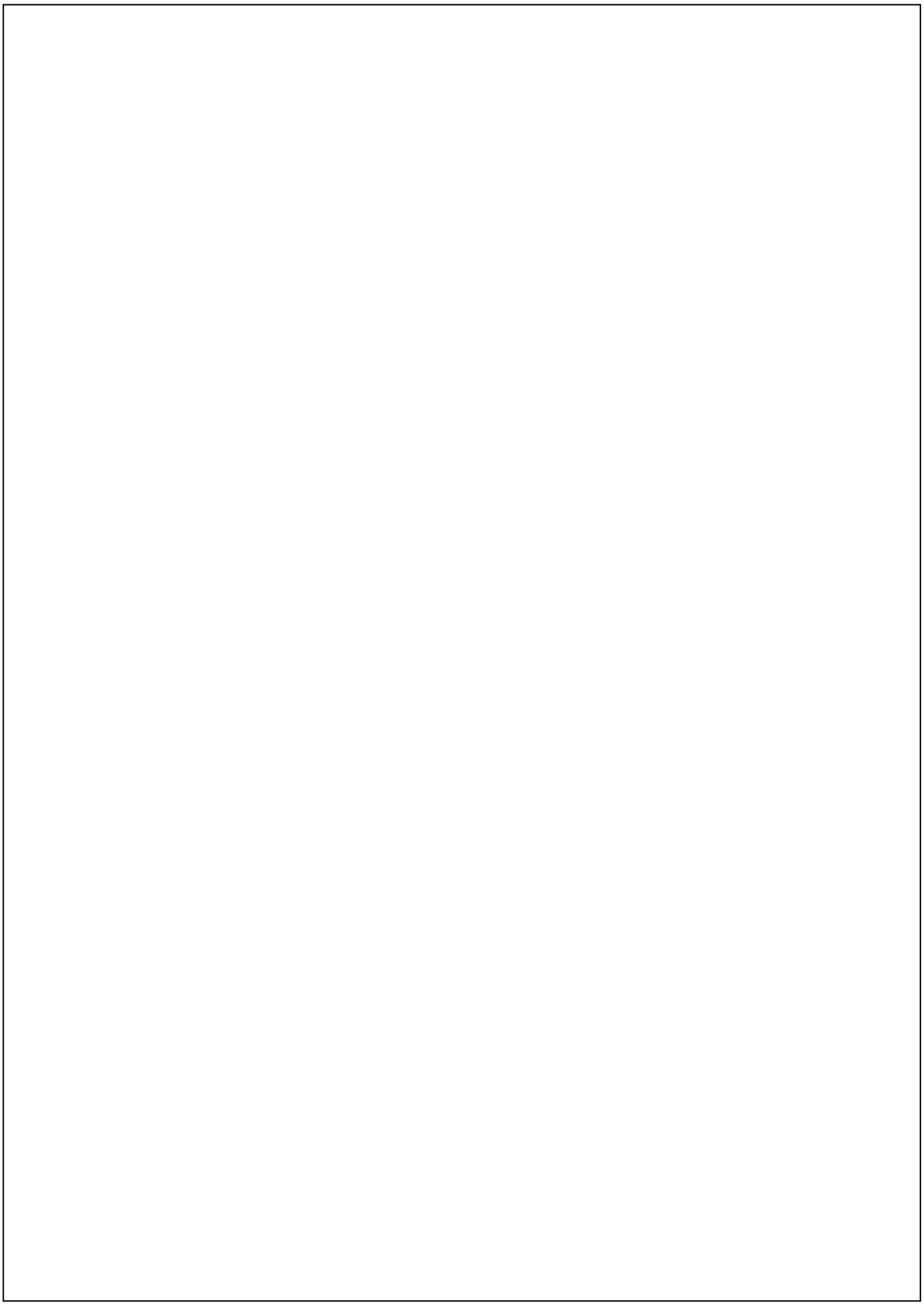


Table of Contents

List of Tables

List of Figures

Chapter 1

Background

Predicting earthquakes remains a significant scientific and societal challenge. Since we cannot make deterministic predictions, forecasts typically model uncertainty about future earthquakes. For very large and infrequent earthquakes, this uncertainty can be substantial, with confidence intervals spanning decades to hundreds of years [97]. Although major earthquakes themselves cannot be precisely forecasted, their aftershocks and the broader phenomenon of earthquake clustering exhibit more statistically predictable properties. These short-term forecasts, which are particularly critical during aftershock sequences, provide necessary hazard assessment to a now vulnerable region. For instance, in the aftermath of the 2010 Canterbury earthquake in New Zealand [164] and the 2016 Amatrice earthquake in Italy [125], subsequent aftershocks were more deadly than the initial earthquakes.

Understanding the physical mechanism of aftershock generation has been one key part of trying to make more reliable forecasts. Various mechanisms are often suggested retrospectively on a case-by-case basis. One such concept is dynamic triggering [14], where seismic waves from an earthquake provide the necessary energy to induce another rupture at a distance. Another is static triggering, where deformation in the elastic crust from an earthquake leads to stress changes, pushing other faults to rupture without the need for seismic waves [58, 119, 198]. Whilst performance of models based on such mechanisms have shown to be competitive in recent retrospective forecasting experiments [118], their use in real-time (operational) earthquake forecasting is limited. This limitation arises because detailed knowledge of earthquake sources and faults, necessary for these models, becomes available only well after the mainshock.

Alternatively, short-term earthquake forecasts based on empirical observations have proven to be a reliable and consistent approach. These forecasts rely on earthquake catalog data, which includes the times, magnitudes, and hypocenter locations of recorded earthquakes. Despite the

complex, volumetric nature of the earthquake rupture process, the point-like events documented in earthquake catalogs have proven a powerful way to represent data for constructing statistical forecasting models as well as providing more scientific insight.

In the remainder of this chapter, I will provide an overview of earthquake catalog data, including its generation, key challenges in its use, and its rapid growth in specific regions. I will then introduce the general family of statistical models known as point processes, which are used to describe this point-like type of spatio-temporal data. Following this, I will provide an overview of the most successful short-term earthquake forecasting model: the Epidemic-Type Aftershock Sequence (ETAS) model. Finally, I will introduce recent machine learning developments in point process modeling, known as neural point processes, before outlining the remainder of my thesis.

1.1 Earthquake Catalogs

Earthquake catalogs document the times, locations, magnitudes, and other characteristics of seismic events. While early 20th-century catalogs relied on rudimentary instruments and human interpretation of seismograms, today, better sensors, denser networks and advanced algorithms for automated seismic phase picking, phase association, location and magnitude calculation, provide more accurate and detailed earthquake catalogs that contain orders of magnitude more events. Despite these advancements, the basic framework of processing waveform data into earthquake catalogs remains unchanged (see [227] for an introduction to seismicity catalogs).

1.1.1 Catalog Generation

Seismometers and Seismic Networks. A seismometer is an instrument that detects and records the vibrations caused by seismic waves [193, 199]. It consists of a sensor to detect ground motion and a recording system to log three-dimensional ground motion over time, typically vertical and horizontal velocities. Seismic networks, comprising multiple seismometers, monitor seismic activity at regional, national or global scales (see, e.g., [229] and references therein). Key characteristics include spatial coverage, density, seismometer type and sensitivity, frequency range, and real-time data transmission capabilities. High-density networks with modern, sensitive equipment provide more detailed and accurate data, enhancing the ability to detect and analyse smaller and more distant earthquakes.

From Waveforms to Phase Picking. The process of converting raw continuous seismic waveforms into useful earthquake data begins with phase picking, which identifies the arrival times of the primary (P) and secondary (S) waves of an earthquake. Historically, this was done manually, but now automated algorithms, such as the Short Term Average/Long Term

Average (STA/LTA) algorithm, detect wave arrivals by analyzing signal amplitude changes [4]. Recent algorithms, such as machine learning classifiers for identifying seismic phases [e.g. 102, 242] and template-matching [e.g. 175], can process much higher volumes of data efficiently and are often able to detect events of much smaller magnitudes.

Earthquake Association and Location After phase picking, the next step is to associate phases from different seismometers with the same earthquake. Simple algorithms declare an event if at least four phase arrivals are detected on different stations within a short time interval. Once phases are associated, location estimation determines the earthquake's hypocenter and origin time by minimizing travel-time residuals using linearized or global inversion algorithms [112, 208]. Given the potential for misidentified or mis-associated phase arrivals due to low signal-to-noise of small events or the near-simultaneous occurrence during very active aftershock sequences, an automated system typically first picks arrival times and determines a preliminary location, which is subsequently reviewed by a seismologist [227]. Locations are typically reported as the geographical coordinates and depths where earthquakes first nucleated (hypocenters), although some catalogs report the centroid location, a central measure of the extended earthquake rupture.

Earthquake Magnitude Calculation The magnitude of an earthquake quantifies the energy released at the source and was originally defined in the seminal paper by Richter [172]. The original definition, now referred to as the local magnitude (ML), is calculated from the logarithm of the amplitude of waves recorded by seismometers. This scale, however, "saturates" at higher magnitudes, meaning it underestimates magnitudes for various reasons, as do several other popular magnitude scales often reported in local and regional catalogs. This led to introduction of the moment magnitude scale (Mw) [71], which computes the magnitude based on the estimated seismic moment M_0, which can be related to the physical rupture process via

$$M_0 = \text{rigidity} \times \text{rupture area} \times \text{slip}, \tag{1.1}$$

where rigidity is a mechanical property of the rock along the fault, rupture area is the area of the fault that slipped, and slip is the distance the fault moved. Mw is determined seismologically via a spectral fitting process to the earthquake waveforms. In practice, it can be challenging to use a single magnitude scale for a broad range of magnitudes, therefore a range of scales may be present within a single catalog, and approximate magnitude conversion equations may be used to homogenize the scales [e.g. 83, and references therein].

1.1.2 Catalog Incompleteness

Data missingness, referred to in seismology as catalog (in)completeness, is the primary challenge faced with earthquake catalogs. It is an important and unavoidable feature, and is a

result of how earthquakes are detected and characterised. Using "raw" earthquake catalog data often requires appropriate pre-processing to address this issue. This typically involves truncating the dataset above a magnitude threshold M_{cut} and within a target spatial region to address the incomplete data [e.g., 129, 130].

There are several reasons why an earthquake may not be detected by a seismic network. Small events may be indistinguishable from noise at a single station, or insufficiently corroborated across multiple stations. Another significant cause of missing events occurs during the aftershock sequence of large earthquakes, when the seismicty rate is high [67, 96]. Human or algorithmic detection abilities are hampered when numerous events occur in quick succession, e.g. when phase arrivals of different events overlap at different stations or the amplitudes of small events are swamped by those of large events. Since catalog incompleteness increases for lower magnitude events, typically the task is to find the value M_c above which there is approximately 100% detection probability. Choosing a truncation threshold M_{cut} that is too high removes usable data. Seismologists often investigate the biases of different magnitude thresholds by performing repeat forecasting experiments for different thresholds [e.g. 120, 201].

Typically M_c is determined by comparing the raw earthquake catalog to the Gutenberg-Richter law [64], which states that the distribution of earthquake magnitudes follows an exponential probability density function

$$f_{GR}(m) = \beta e^{\beta(m-M_c)} \quad : m \geq M_c. \tag{1.2}$$

where β is a rate parameter related to the b-value by $\beta = b \log 10$. Histogram-based approaches, such as the simple Maximum Curvature method [226] as well as many others [e.g. 83, and references therein], identify the magnitude at which the observed catalog deviates from this law, indicating incompleteness (Figure 1.1b).

In practice, catalog completeness varies in both time and space $M_c(t, \mathbf{x})$ [e.g. 179]. During aftershock sequences, $M_c(t)$ can be very high [e.g., 2, 66] (Figure 1.1a). Thresholding at the maximum value might remove too much data. Instead, modelers either omit particularly incomplete periods during training and testing [69, 94], model the incompleteness itself [65–67, 79, 133, 154, 224], or accept known biases from disregarding this issue [196]. Spatially, catalogs are less complete farther from the seismic network [129], so the spatial region can be constrained to remove outer, more incomplete areas (Figure 1.1c).

1.1.3 Catalog Growth

Earthquake catalogs are rapidly growing in size, largely due to the continued deployment of more seismic stations, which enhance the capabilities of regional and global seismic networks.

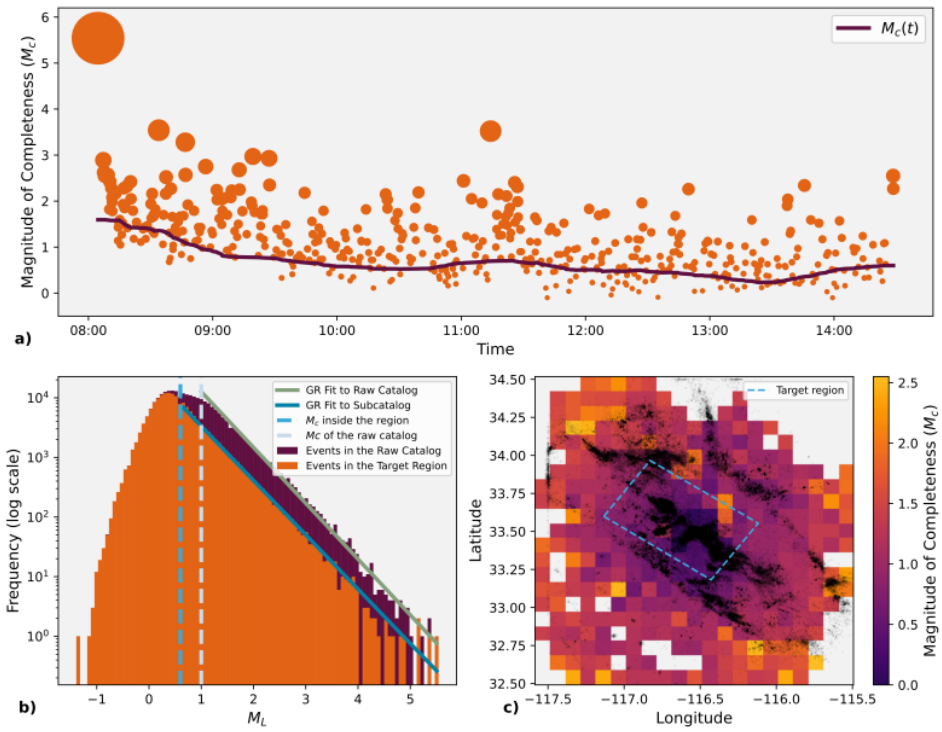


Figure 1.1: **a)** the June 10, 2016 Mw5.2 Borrego Springs earthquake and aftershocks, which occurred on the San Jacinto fault zone. An estimate of the magnitude of completeness $M_c(t)$ over time using the Maximum Curvature method reveals more incompleteness immediately following the large earthquake. **b)** magnitude-frequency histograms reveal that truncating the raw catalog [225] to inside the target region decreases M_c. Each histogram is fit to the Gutenberg-Richter (GR) law and an estimate of M_c for each catalog occurs where the histogram deviates from the (GR) line. **c)** An estimate of M_c for gridded regions of the San Jacinto fault zone, using the raw catalog [225].

For example, the Southern California Seismic Network grew from 7 stations in 1932 to around 400 by 2008 [88], and the INGV Italian Seismic Network expanded from approximately 75 stations in 1988 to around 550 by 2020 [5, 182]. This increased density not only broadens coverage of the network but also improves the accuracy in distinguishing small earthquakes from noise, thereby decreasing the magnitude of completeness of the catalog over time. Due to the logarithmic nature of the Gutenberg-Richter law (1.2), even a modest reduction in the magnitude of completeness by $1 M_w$ can result in a significant increase in the number of detected events.

The reduction in the smallest magnitude earthquakes that seismic networks can detect has been dramatically enhanced by machine learning-based algorithms. These advances, introduced in section 1.1.1, allow for historical waveform data stored in data centers to be reanalyzed, generating enhanced catalogs alongside routine ones. For instance, during the Central Apennines earthquake sequence, which began on August 24, 2016, with the M_w 6.0 Amatrice earthquake, the INGV produced a routine catalog covering the subsequent year [20]. This catalog contained roughly 82,000 events with a completeness magnitude of $M_c = 2.3$. Utilizing PhaseNet, a neural network-based phase picker, to determine P and S-wave arrival times, an enhanced catalog for the same sequence was created, containing around 900,000 events with a reported completeness magnitude of $M_c = 0.2$ [206]. Similarly, the QTM catalog in Southern California was produced using template matching to identify seismic phases, resulting in 1.81 million events between 2008-2017 and a reduced magnitude of completeness from $M_c = 1.7$ to $M_c = 0.3$ [175].

Key Question:
To what extent, if any, does this newly revealed low magnitude data enhance earthquake predictability?

Key Question:
What challenges does this new quantity of data pose to our current forecasting frameworks, and what are the potential solutions?

1.2 Point Processes

To model and ultimately forecast using the point-like representation of earthquakes that form earthquake catalogs requires the use of the family of statistical models: point processes. These models describe the random locations of points in space. The term 'space' is used in the mathematical sense here, and for the purpose of modeling earthquakes this spans time, geographical space and magnitude. Point process models for earthquakes can be comprised of

any combination of these three components.

For forecasting during aftershock sequences, typically all three components are considered. However, often the spatial-component can be dropped either to simplify the problem (at the expense performance) or if the region of interest is small enough such that contributions from the spatial components are relatively small. I will introduce the family of models successively introducing time (1.2.1), magnitudes (referred to as marks when generalising beyond earthquakes) (1.2.2) and finally space (1.2.3).

1.2.1 Temporal Point Process

A temporal point process is a continuous-time stochastic process that models the random number of events $N((t_a, t_b])$ which occur in a time interval $(t_a, t_b] \in \mathbb{R}^+$. The process is typically defined by a non-negative *conditional intensity function*

$$\lambda(t, \mathbf{x}|\mathcal{H}_t) := \lim_{\Delta t \to 0} \frac{\mathbb{E}\left[N([t, t+\Delta t)|\mathcal{H}_t\right]}{\Delta t}, \tag{1.3}$$

where $\mathcal{H}_t = \{t_i | t_i < t\}$ denotes the history of events preceding time t. The intensity function [168] completely defines the point process and can take a variety of functional forms. The most basic form is the stationary Poisson process [27] which assumes that all events are independent of each other, and the conditional intensity function is constant, $\lambda(t|\mathcal{H}_t) = \mu$. Self-exciting point processes assume that events increase the likelihood of subsequent events, with a popular class of these processes being the Hawkes process [76]. The Hawkes process is defined by its conditional intensity $\lambda(t|H_t) = \mu + \sum_{t_i < t} g(t - t_i)$, where $g(s)$ is a non-negative kernel function defining how past events trigger subsequent events.

Parameters of the intensity function, θ, are typically estimated from data through maximising the log-likelihood function. Due to the natural ordering of events in time, the log-likelihood of observing a sequence of events $\{t_i\}_{i=1}^n$ in the interval $[0, T]$ can be written as the sum of next-event probability density functions (pdf),

$$\log p(t_1, \ldots, t_n; \theta) = \sum_{i=1}^{n} \log p(t_i|\mathcal{H}_{t_i}; \theta) + \log \mathbb{P}(t_{n+1} \in [T, \infty)) \tag{1.4}$$

$$= \sum_{i=1}^{n} \left[\log \lambda(t_i|H_{t_i}; \theta) \; - \int_{t_{i-1}}^{t_i} \lambda(t|H_t; \theta) dt \right] - \int_{t_n}^{T} \lambda(t|H_t; \theta) dt \tag{1.5}$$

$$= \sum_{i=1}^{n} \log \lambda(t_i|H_{t_i}; \theta) \; - \int_{0}^{T} \lambda(t|H_t; \theta) dt, \tag{1.6}$$

where the final term in 1.4 and 1.5 states that no events happen in the interval $[t_n, T]$.

1.2.2 Marked Temporal Point Process

Temporal point processes can be extended to incorporate marks. A marked point process is stochastic process that generates events paired with a mark, $\{t_i, m_i\}_{i=1}^n \in (\mathbb{R}_{>0} \times \mathcal{M})$. In the context of earthquakes, this represents the occurrence times of earthquakes along with their magnitudes. A marked point process is defined by its conditional intensity function,

$$\lambda(t, m|\mathcal{H}_t) := \lim_{\Delta t, \Delta m \to 0} \frac{\mathbb{E}\left[N([t, t+\Delta t) \times B(m, \Delta m)|\mathcal{H}_t\right]}{\Delta t |B(m, \Delta m)|}.$$

where $|B(m, \Delta m)|$ is the Lebesgue measure of the ball $B(m, \Delta m)$ with radius Δm. The log-likelihood of observing a marked sequence of events in $[0, T]$ is given by

$$\log L(\{t_i, m_i\}) = \sum_i \log \lambda(t_i, m_i|H_{t_i}; \theta) \; - \int_0^T \int_{\mathcal{M}} \lambda(t, m|H_t; \theta) dm dt. \tag{1.7}$$

1.2.3 Spatio-temporal Point Process

One final extension enables the full modelling of events recorded in earthquake catalogs. A marked spatio-temporal point process generates event-times paired with a mark and a location, $\{t_i, m_i, \mathbf{x}_i\}_{i=1}^n \in \left(\mathbb{R}_{>0} \times \mathcal{M} \times \mathbb{R}^2\right)$. Locations can be either 2-dimensional coordinates or include a third depth dimension. This process is again defined by the conditional intensity function,

$$\lambda(t, m, \mathbf{x}|\mathcal{H}_t) := \lim_{\Delta t, \Delta m, \Delta \mathbf{x} \to 0} \frac{\mathbb{E}\left[N([t, t+\Delta t) \times B(m, \Delta m) \times B(\mathbf{x}, \Delta \mathbf{x})|\mathcal{H}_t\right]}{\Delta t |B(m, \Delta m)||B(\mathbf{x}, \Delta \mathbf{x})|}.$$

The log-likelihood of observing the sequence $\{t_i, m_i, \mathbf{x}_i\}_{i=1}^n$ in the space-time region $[0, T] \times \mathcal{S}$ given by,

$$\log L(\{t_i, m_i\}) = \sum_i \log \lambda(t_i, m_i, \mathbf{x}_i|H_{t_i}; \theta) \; - \int_0^T \int_{\mathcal{M}} \int_{\mathcal{S}} \lambda(t, m, \mathbf{x}|H_t; \theta) d\mathbf{x} dm dt. \tag{1.8}$$

For brevity, the notation $\lambda^*(t, m, \mathbf{x}) := \lambda(t, m, \mathbf{x}|\mathcal{H}_t)$ is often used to indicate the dependence of the function on the history $\mathcal{H}_t = \{(t_i, m_i, \mathbf{x}_i)|t_i < t\}$. Additionally, the conditional intensity function is often factorised $\lambda^*(t, m, \mathbf{x}) = \lambda_g^*(t) p^*(m, \mathbf{x}|t)$ without loss of generality, where $\lambda_g^*(t) = \int_{\mathcal{M}} \int_{\mathcal{S}} \lambda(t, m, \mathbf{x}|H_t) d\mathbf{x} dm$, is referred to as the ground intensity, a quantity that describes the temporal-only part of the process (for temporal-only point processes $\lambda^*(t) = \lambda_g^*(t)$), and $p^*(m, \mathbf{x}|t)$ is the conditional density function of the magnitude m, and location $\mathbf{x}$, at time t.

1.2.4 Auto-regressive Forecasting

The conditional intensity function allows us to define the (log) pdf of the next event $(t, m, \mathbf{x})$ given an observed history of events $\mathcal{H}_{t_i}$,

$$\log p(t, m, \mathbf{x}|\mathcal{H}_{t_i}) = \log \lambda(t, m, \mathbf{x}|H_t) \; - \int_{t_i}^{t} \lambda_g^*(s)ds. \tag{1.9}$$

This allows us to define an auto-regressive earthquake forecasting procedure, whereby at the occurrence of an earthquake $(t_i, m_i, \mathbf{x}_i)$, a pdf for the next time, magnitude and location is generated.

When the next event $(t_{i+1}, m_{i+1}, \mathbf{x}_{i+1})$ is observed, we can compare the performance of one model $p_1(\cdot)$ against another model $p_1(\cdot)$ using the difference of the log pdfs,

$$\log p_1(t_{i+1}, m_{i+1}, \mathbf{x}_{i+1}|\mathcal{H}_{t_i}) - \log p_2(t_{i+1}, m_{i+1}, \mathbf{x}_{i+1}|\mathcal{H}_{t_i}). \tag{1.10}$$

Taking the mean of this difference over all observed earthquakes gives the *information gain* of one model over another.

1.2.5 Catalog-Based Forecasting

Although auto-regressive forecasting is a natural consequence of the formulation of (spatio-)temporal point processes, it is typically not the desired forecasting object. Probabilistic seismic hazard analysis (PSHA) requires more long-term prediction beyond the next-event [41, 53], so daily or weekly forecasts are generally preferred. Unfortunately in general for point process models, the full pdf over events in the future time horizon is an intractable quantity. Instead probabilities over the forecast horizon can be approximated with a Poisson assumption [224] or can be generated empirically using simulations.

There is currently a very active effort to evaluate forecasting models that provide this type of simulated forecast. The Collaboratory for the Study of Earthquake Predictability (CSEP) [89, 128, 177, 181] [`https://cseptesting.org/`] aims to unify the framework for earthquake model testing and evaluation, hosting retrospective (on past earthquakes) and fully prospective (on future earthquakes yet to happen) forecasting experiments globally. CSEP benchmarks short-term models using performance metrics that require forecasts to be generated by simulating many repeat sequences over a specified time horizon (typically one day). These simulated forecasts are compared by discretizing time and space intervals, with test statistics calculated for event counts, magnitudes, locations, and times. This simulation-based approach also allows the comparison of models that don't output explicit next-event probabilities .

Since generating repeated sequences over forecast horizons is computationally costly, log-likelihood metrics offer a more streamlined metric during model development [75, 150]. All

models evaluated in this thesis have an auto-regressive likelihood, allowing comparison through mean log-likelihood or information gain. However, Chapter 4, directs future machine learning model development towards more impactful catalog-based CSEP evaluations.

1.2.6 Random Time Change Theorem

Another characteristic quantity of point processes is the compensator function,

$$\Lambda^*(t) := \int_0^t \lambda_g^*(s) ds, \tag{1.11}$$

which can be used either to define point process models (see Chapter 2), to evaluate the performance of models, or to simulate events from models (see Section 1.2.7). Its uses are the result of a fundamental theorem in point process theory.

Theorem 1.1. *(Random time change theorem [168])*
Suppose $\{(t_i, m_i, \mathbf{x}_i)\}_{i=1}^n$ *is a realisation of a point process with compensator function* $\Lambda^*(t)$ *on the interval* $[0, T]$ *and that* $\Lambda^*(T) < \infty$. *Then the sequence* $\{\Lambda^*(t_i)\}_{i=1}^n$ *is distributed according to a stationary Poisson process with unit rate on the interval* $[0, \Lambda^*(T)]$.

A visual goodness-of-fit test known as a residual plot, compares the points $\{\Lambda^*(t_i), N(t_i)\}_{i=1}^n$ against the line $y = x$. As a consequence of Theorem 1.1, if $\Lambda^*(t)$ is the data generating model then the points should provide a good fit to the line. Another equivalent visual goodness-of-fit test compares $\{t_i, \Lambda^*(t_i)\}_{i=1}^n$ with $\{t_i, N(t_i)\}_{i=1}^n$, since $\Lambda^*(t) = \mathbb{E}\left(N((0, t])\right)$.

1.2.7 Simulation

We now describe two methods for simulating events from general marked spatio-temporal point processes. The two approaches describe how to simulate from the ground process $\lambda_g^*(t)$, and assume that one can simulate the magnitude and location given the event time, $m, \mathbf{x} \sim p^*(m, \mathbf{x}|t)$.

The first uses the random time change theorem (Theorem 1.1) and assumes that the user has access to $\Lambda^{*-1}(\cdot)$.

Algorithm 1 Inverse transform sampling [189]

Parameters: Interval $[0, T]$, compensator $\Lambda^*(t)$
$t = 0$, $s = 0$ and $i = 1$
while $t \leq T$ **do**
  $\tau_i \sim \text{Exp}(1)$ ▷ Simulate the next event from Poisson process.
  $s = s + \tau_i$
  $t = \Lambda^{*-1}(z)$ ▷ Transform back into the observed domain.
  $m, \mathbf{x} \sim p^*(m, \mathbf{x}|t)$ ▷ Sample the magnitude and location.
  if $t < T$ **then** ▷ Only keep event if it is within the time interval.
    Set $t_i = t$
    $i = i + 1$
  end if
end while
Output: Sequence of events $\{t_i, m_i, \mathbf{x}_i\}$

While this algorithm is computationally efficient, typical point process models do not have access to an analytical inverse of $\Lambda^*(t)$ and require numerically solving for $t = \Lambda^{*-1}(z)$. Instead [149] developed a thinning based algorithm which proposes new events from a proposal process and accepts or rejects them proportionally to their intensity $\lambda_g^*(t)$. The procedure requires simulation from a Poisson process that upper bounds the target intensity function on $[t, T]$,

$$m^*(t) \geq \sup_{s \in [t,T]} \lambda_g^*(t) \tag{1.12}$$

Algorithm 2 Ogata's Modified Thinning Algorithm [149, 189, 190]

Parameters: Interval length T, conditional intensity $\lambda^*(t)$, upper bound $m^*(t)$
$t = 0$ and $i = 1$
while $t \leq T$ **do**
  $\mu_0 = m^*(t)$ ▷ Compute the upper bound
  $\tau_i \sim \text{Exp}(\mu_0)$ ▷ Simulate from the proposal Poisson process
  $t = t + \tau_i$
  $u \sim \text{Uniform}([0, 1])$
  if $t < T$ **and** $u < \lambda^*(t)/\mu_0$ **then** ▷ Check acceptance
    $t_i = t$
    $m_i, \mathbf{x}_i \sim p^*(m, \mathbf{x}|t_i)$ ▷ Sample the magnitude and location.
    $i = i + 1$
  end if
end while
Output: Sequence of events $\{t_i, m_i, \mathbf{x}_i\}$

1.3 ETAS

The Epidemic Type Aftershock Sequence (ETAS) model has been the most dominant way of forecasting seismicity since its introduction by [150]. It has been adopted for operational earthquake forecasting by government agencies in California [131], New-Zealand [23], Italy [197], Japan [156] and Switzerland [135], and performs consistently well in CSEP's retrospective and fully prospective forecasting experiments [e.g. 16, 118–120, 171, 207, 228].

1.3.1 Intensity Function Formulation

It's success is due to modeling the self-exciting nature of seismicity through its formulation as a Hawkes process. The temporal version [150] has the general formulation,

$$\lambda(t, m|\mathcal{H}_t) = \left(\mu + \sum_{i:t_i<t} g(t - t_i, m_i)\right) f_{GR}(m), \tag{1.13}$$

where μ is a constant background rate of events, $g(t, m)$ is a non-negative excitation kernel which describes how past events contribute to the likelihood of future events and $f_{GR}(m)$ is the probability density of observing magnitude m. The magnitudes are said to be "unpredictable" since they do not depend on previous events and are distributed according to the most significant empirical law in statistical seismology: the Gutenberg-Richter law for magnitudes [64] (equation 1.2).

The triggering kernel $g(t, m)$ factorises the contribution from the magnitude and the time using two other empirical laws from statistical seismology,

$$g(t, m) = k(m)h(t) \tag{1.14}$$

$$k(m) = Ke^{\alpha(m-M_c)} \ : \ m \geq M_c \tag{1.15}$$

$$h(t) = c^{p-1}(p-1)(t+c)^{-p} : \ t \geq 0 \tag{1.16}$$

where the $k(m)$ is known as the Utsu law of productivity [212] and $h(t)$ is a power law known as the Omori-Utsu decay [215].

The spatio-temporal version of ETAS [151] extends the triggering kernel to include a spatial component based on the squared distance between events,

$$\lambda(t, m, \mathbf{x}|\mathcal{H}_t; \theta) = \left(\mu + \sum_{i:t_i<t} g(t - t_i, m_i, ||\mathbf{x} - \mathbf{x}_i||_2^2)\right) f_{GR}(m), \tag{1.17}$$

where,

$$g(t, m, r^2) = k(m)h(t)d(r^2, m) \tag{1.18}$$

$$d(r^2, m) = \left(r^2 + d \cdot e^{\gamma(m - M_0)}\right)^{-1-\rho}. \tag{1.19}$$

The additional spatial component to the kernel, $d(r^2, m)$, is an isotropic power law decay kernel that is based on another empirical law from statistical seismology known as the Utsu-Seki formula [211], which relates the magnitude of a triggering earthquake m with the area of the aftershock region. Other formulations of the spatial kernel are often used [153], however these don't vary substantially from equation (1.19).

1.3.2 Branching Process Formulation

An equivalent way of formulating the ETAS model is as a Poisson cluster process [167]. In this formulation a set of immigrants I are realisations of a Poisson process with rate μ. Each immigrant $t_i \in I$ has a magnitude m_i with probability density f_{GR} and generates offspring S_i from an independent non-homogeneous Poisson process, with rate $g(t - t_i, m_i)$. Each offspring $t_j \in S_i$ also has magnitude m_j with probability density f_{GR} and generate offspring S_j of their own. This process is repeated over generations until a generation with no offspring in time interval $[0, T]$ is produced. If the average number of offspring for a given event is greater than one, the process is called super-critical and there is a non-zero probability that infinitely many offspring are created within the finite time interval $[0, T]$. Although it is not observed in the data, this process is accompanied by latent branching variables $B = \{B_1, \ldots, B_n\}$ which define the branching structure of the process,

$$B_i = \begin{cases} 0 \text{ if } t_i \in I \text{ (i.e. } i \text{ is a background event)} \\ j \text{ if } t_i \in S_j \text{ (i.e. } i \text{ is an offspring of } j) \end{cases} \tag{1.20}$$

This causal model for earthquake interactions broadly aligns with the understanding of the physics of earthquake triggering introduced earlier. Either with dynamic wave triggering [14] or by static stress triggering [58, 119].

The branching process formulation of ETAS defines another method for simulating from this point process model and as we will see in section 1.3.3, defines an additional method of inferring parameters to standard maximum likelihood inference. Algorithm 3 describes how to simulate a temporal only earthquake sequence using the branching process formulation. To extend it to include spatial coordinates, step 3 would include simulating each offspring's location from $d(r^2, m)$.

Algorithm 3 ETAS Branching Process Simulation in the interval $[0, T]$

1. Generate events $(t_i, m_i) \in G^{(0)}$ from a stationary Poisson process with intensity μ in $[0, T]$.
2. $l = 0$.
3. For each $(t_i, m_i) \in G^{(l)}$ simulate its $N^{(i)}$ offspring, where $N^{(i)} \sim \text{Poisson}(k(m_i))$. Each offspring's time is generated from $h(t)$ and magnitude from f_{GR} and are labelled $S_i^{(l)}$.
4. $G^{(l+1)} = \bigcup_{i \in G^{(l)}} S_i^{(l)}$
5. If $G^{(l)}$ is not empty, $l = l + 1$ and return to step 3.
6. Sort and return $S = \bigcup_{j=0}^{l} G^{(j)}$ as the set of all simulated events.

This procedure has time complexity $\mathcal{O}(n)$ for steps 1-5, since there is only a single pass over all events. An additional time constraint is added in step 6, where the whole set of events are sorted chronologically, which is at best $\mathcal{O}(n \log n)$. This scales much better than using inverse transform sampling (Algorithm 1) and Ogata's modified thinning algorithm (Algorithm 2) which requires evaluating the intensity function $\lambda(t|\mathcal{H}(t))$ at least once for each one of n events that are simulated. Evaluating the intensity function requires a summation over all events before time t, thus giving these simulation procedures time complexity $\mathcal{O}(n^2)$.

1.3.3 Parameter Estimation

Most commonly, point estimates of ETAS parameters are found through maximum likelihood estimation (MLE). No analytical solution exists due to the form of the likelihood function, and so numerical optimisation is required. Conventional methods such as Nelder–Mead [146], BFGS [15, 49, 57, 185] and the conjugate-gradient method [48] are typically employed for the optimisation [222].

For spatio-temporal ETAS, the second term in the likelihood (1.8) involves solving a spatio-temporal integral over the target region $\mathcal{S}$. This also typically doesn't have an analytical solution and so is either numerically solved for each call of the likelihood [109, 151], slowing the procedure, or the approximation $\mathcal{S} \to \infty$ is used to simplify the integral resulting in marginal errors [109].

Likelihood evaluation has time complexity $\mathcal{O}(n^2)$ due to the double summation and so this procedure is becoming increasingly slow with the growing size of earthquake catalogs and in many cases infeasible. Furthermore, the likelihood function can be very flat in large regions of

the parameter space, slowing the convergence of optimisation techniques [221].

Veen and Schoenberg [221] developed an Expectation Maximisation (EM) procedure based on the branching formulation of ETAS. For the temporal ETAS, the joint likelihood $p(\mathcal{H}_T, B|\theta)$ can be expressed in terms of the conditional densities,

$$\mathbb{P}(B_i = j|\mathcal{H}_T, \theta) = \begin{cases} \frac{\mu}{\mu+\sum_{j=1}^{i-1} k(m_j)h(t_i-t_j)} & : j = 0 \\ \frac{k(m_j)h(t_i-t_j)}{\mu+\sum_{j=1}^{i-1} k(m_j)h(t_i-t_j)} & : j = 1, 2, \ldots, i-1 \end{cases} \tag{1.21}$$

$$\begin{aligned} \log p(\mathcal{H}_T|\theta, B) = |S_0| \log \mu - \mu T \\ + \sum_{j=1}^{n} \left(-k(m_j)H(T-t_j) + |S_j| \log k(m_j) + \sum_{t_i \in S_j} \log h(t_i - t_j) \right), \end{aligned} \tag{1.22}$$

where $|S_j|$ denotes the number of events that were triggered by the event at t_j and $H(t) = \int_0^t h(s)ds$ denotes the integral of the Omori decay kernel. Conditional densities for spatio-temporal ETAS can be found in Nandan et al. [142].

The EM procedure maximises the marginal likelihood over the unobserved branching structure,

$$\log \int p(\mathcal{H}_T|B, \theta)p(B|\theta)dB, \tag{1.23}$$

through the iteration,

$$\theta^{(k+1)} = \arg\max_{\theta} \mathbb{E}_{B \sim p(\cdot|\mathcal{H}_T, \theta^{(k)})} \left[\log p(\mathcal{H}_T, B|\theta) \right]. \tag{1.24}$$

This avoids the need to numerically approximate the integral term in the likelihood, provides more stability during estimation and simultaneously estimates the causal structure.

From these point estimates, forecasts can be issued by simulating multiple catalogs over the forecasting horizon. Forecast uncertainty is quantified by the distribution of simulations from MLE parameter values, however, this approach fails to quantify uncertainty contained in estimating the parameters themselves. Parameter uncertainty for MLE can be estimated using the Hessian of the likelihood [148, 169, 222], which requires a very large sample size to be effective, and is only asymptotically unbiased (i.e. when the time horizon is infinite). Multiple runs of the MLE procedure with different initial conditions [113] can also be used to express parameter uncertainty.

1.3.4 Bayesian Inference

Full characterisation of the parameter uncertainty is achieved with Bayesian inference, a procedure which returns the entire probability distribution over parameters conditioned on the

observed data and updated from prior knowledge. The procedure is most commonly used for parameter estimation during the early part of an aftershock sequence, either to account for parameter uncertainty due to data incompleteness [155], or to account for inter-sequence ETAS parameter variability [159, 217], allowing for real-time parameter updating as an aftershock sequence progresses and new data is collected.

In what follows, we will describe methods for learning posterior distributions for the temporal ETAS model formulated as,

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{i:t_i<t} Ke^{\alpha(m-M_c)}c^{p-1}(p-1)(t+c)^{-p}. \tag{1.25}$$

This is the focus of Chapter 3, where we construct an approximate inference method for temporal ETAS.

Given we observe the sequence $\mathbf{Y}_{\text{obs}} = \{(t_1, m_1), (t_2, m_2), \ldots, (t_n, m_n)\}$ in the interval $[0, T]$, we are interested in the posterior probability $p(\theta|\mathbf{Y}_{\text{obs}})$ for the parameters $\theta = (\mu, K, \alpha, c, p)$ of the ETAS model defined in (1.13)-(1.16), updated from some prior probability $p(\theta)$. The posterior distribution, expressed in Bayes' rule,

$$p(\theta|\mathbf{Y}_{\text{obs}}) \propto p(\mathbf{Y}_{\text{obs}}|\theta)p(\theta), \tag{1.26}$$

is known up to a constant of proportionality through the product of the prior $p(\theta)$ and the likelihood $p(\mathbf{Y}_{\text{obs}}|\theta)$.

The posterior does not have a closed form expression for this temporal ETAS model (and other temporal and spatio-temporal formulations) and so several approaches have used Markov Chain Monte Carlo (MCMC) to obtain samples from this distribution [136, 155, 174, 219]. Vargas and Gneiting [219] draw samples from the posterior $p(\theta|\mathbf{Y}_{\text{obs}})$ through independent random walk Markov Chain Monte Carlo (MCMC) with Metropolis-Hastings rejection of proposed samples. This approach, however, can suffer from slow convergence due to parameters of the ETAS model having high correlation [174].

BayesianETAS

In light of this, Ross [174] developed an MCMC sampling scheme, `bayesianETAS`, which uses the latent branching structure $B = \{B_1, \ldots, B_n\}$. The scheme iteratively samples the branching structure,

$$\mathbb{P}(B_i^{(k)} = j|\mathbf{Y}_{\text{obs}}, \theta^{(k-1)}) = \begin{cases} \frac{\mu^{(k-1)}}{\mu^{(k-1)}+\sum_{j=1}^{i-1} k(m_j)h(t_i-t_j)} & : j = 0 \\ \frac{k(m_j)h(t_i-t_j)}{\mu^{(k-1)}+\sum_{j=1}^{i-1} k(m_j)h(t_i-t_j)} & : j = 1, 2, \ldots, i-1 \end{cases} \tag{1.27}$$

where each B_i is drawn independently from $\{0, \dots, i-1\}$ with weights from 1.27.

$\theta^{(k)} = (\mu^{(k)}, K^{(k)}, \alpha^{(k)}, c^{(k)}, p^{(k)})$, can then be sampled using the conditional likelihood (1.22).

$$\mu^{(k+1)} \sim p(\mu|\mathbf{Y}, \theta, B) \propto p(\mu) e^{-\mu T} \mu^{|S_0|}, \tag{1.28}$$

where a conjugate prior $p(\mu) = \text{Gamma}(\alpha_\mu, \beta_\mu)$ allows direct sampling from $p(\mu|\mathbf{Y}, \theta, B) = \text{Gamma}(\alpha_\mu + |S_0|, \beta_\mu + T)$.

$$K^{(k+1)}, \alpha^{(k+1)} \sim p(K, \alpha|\mathbf{Y}, \theta, B) \propto p(K, \alpha) \prod_{j=1}^{n} e^{-k(m_j)H(T-t_j)} k(m_j)^{|S_j|}, \tag{1.29}$$

are sampled using random walk MCMC.

$$c^{(k+1)}, p^{(k+1)} \sim p(c, p|\mathbf{Y}, \theta, B) \propto p(c, p) \prod_{j=1}^{n} e^{-k(m_j)H(T-t_j)} \prod_{t_i \in S_j} h(t_i - t_j) \tag{1.30}$$

are again sampled using random walk MCMC.

By conditioning on the branching structure, the dependence between parameters (K, α) and (c, p) is reduced, decreasing the time it takes for the sampling scheme to converge. We can see from equation (1.27) that estimating the branching structure from the data is a procedure that is $\mathcal{O}(n^2)$. Since for every event $i = 1, \dots, n$, to estimate its parent we must sum over $j = 1, \dots, i-1$. For truncated version of the time kernel $h(t)$, this operation can be streamlined to $\mathcal{O}(n)$. However, due to the heavy-tailed power-law kernel typically used, the complexity scaling remains high as significant truncation of the kernel is unfeasible.

Other MCMC approaches which evaluate the likelihood (not conditional on branching structure) have quadratic complexity $\mathcal{O}(n^2)$, and therefore are only suitable for catalogs up to 10,000 events. In fact, the GP-ETAS model by Molkenthin et al. [136] has cubic complexity $\mathcal{O}(n^3)$, since their spatially varying background rate uses a Gaussian-Process (GP) prior.

INLABRU

More recently Serafini et al. [184] have constructed an approximate method of Bayesian inference for the ETAS model based on an Integrated Nested Laplace Approximation (INLA) implemented in the R-package `inlabru` as well as a linear approximation of the likelihood. This approach expresses the log-likelihood as 3 terms,

$$\log p(\mathbf{Y}_{obs}|\theta) = -\Lambda_0(\mathbf{Y}_{\text{obs}}, \theta) - \sum_{i=1}^{n} \Lambda_i(\mathbf{Y}_{\text{obs}}, \theta) + \sum_{i=1}^{n} \log \lambda(t_i|\mathcal{H}_{t_i}), \tag{1.31}$$

where,

$$\begin{aligned}\Lambda_0(\mathbf{Y}_{\text{obs}}, \theta) &= \int_0^T \mu dt, \\ \Lambda_i(\mathbf{Y}_{\text{obs}}, \theta) &= \sum_{h=1}^{C_i} \int_{b_{h,i}} g(t - t_i, m_i) dt \\ &= \sum_{h=1}^{C_i} \Lambda_i(\mathbf{Y}_{\text{obs}}, \theta, b_{h,i}).\end{aligned}$$

where $b_{1,i}, \ldots, b_{C_i,i}$ are chosen to partition the interval $[t_{i-1}, t_i]$. The log-likelihood is then linearly approximated with a first order Taylor expansion with respect to the posterior mode θ^*,

$$\begin{aligned}&\widehat{\log p}(\mathbf{Y}_{\text{obs}}|\theta; \theta^*) = \\ &= -\widehat{\Lambda}_0(\mathbf{Y}_{\text{obs}}, \theta, \theta^*) - \sum_{i=1}^{n} \sum_{h=1}^{C_i} \widehat{\Lambda}_i(\mathbf{Y}_{\text{obs}}, \theta, b_{h,i}; \theta^*) + \sum_{i=1}^{n} \widehat{\log \lambda}(\mathbf{Y}_{\text{obs}}, \theta; \theta^*) \\ &= -\exp\{\overline{\log \Lambda_0}(\mathbf{Y}_{\text{obs}}, \theta, \theta^*)\} - \sum_{i=1}^{n} \sum_{h=1}^{C_i} \exp\{\overline{\log \Lambda}_i(\mathbf{Y}_{\text{obs}}, \theta, b_{h,i}; \theta^*)\} + \sum_{i=1}^{n} \overline{\log \lambda}(\mathbf{Y}_{\text{obs}}, \theta; \theta^*),\end{aligned}$$

where the notation, $\widehat{\Lambda}$, denotes the approximation of Λ and $\overline{\log \Lambda(\ ; \theta^*)}$ denotes the first order Taylor expansion of $\log \Lambda$ about the point θ^*.

The posterior mode θ^* is found through a Quasi-Newton optimisation method and the final posterior densities are found using INLA, which approximates the posteriors $p(\theta|\mathbf{Y}_{\text{obs}})$ using a latent Gaussian model.

This approach speeds up computation of the posterior densities, since it only requires evaluation of the likelihood function during the search for the posterior mode. However, the approximation of the likelihood requires partitioning the space into a number of bins, which the authors recommend choosing as greater than 3 per observation. This results in the approximate likelihood having complexity $\mathcal{O}(n^2)$. Serafini et al. [184] demonstrated a factor 10 speed-up of `inlabru` over `bayesianETAS` for a catalog of 3,500 events but they do not provide results on larger catalogs.

Modern earthquake catalogs, now comprising up to 10^6 events, have outgrown the computational capacity of these traditional methods for fitting ETAS models. While larger datasets often reduce uncertainty, Bayesian inference enables seismologists to rigorously quantify and express model uncertainty, particularly when dealing with non-stationary data observed within finite time windows.

Key Question:
How can Bayesian inference for the ETAS model scale to the size of modern earthquake catalogs?

1.3.5 Limitations

It is commonly accepted that ETAS is misspecified. As ETAS only describes the self-exciting nature of seismicity, it cannot capture any kind of inhibition or release of stress such as captured by stress-release models [10, 232, 238] or models based on elastostatic stress transfer and Coulomb Rate-and-State (CRS) friction [35]. Furthermore, foreshock activity that differs from ETAS has also been observed [13, 108, 126, 152].

A significant portion of the discrepancies between observed earthquakes and ETAS forecasts is often attributed to non-stationarities in underlying aftershock parameters. These include considerable inter-sequence variations in aftershock productivity [159], shifts in the size distribution of earthquake magnitudes [60, 216], changes in Omori law parameters governing aftershock decay [147], and transient increases in background seismicity rates during earthquake swarms [68, 84, 111, 114]. While stationary ETAS models can use Bayesian parameter updating to account for such variability [159, 217], this approach propagates substantial uncertainty. In contrast, non-stationary ETAS models [70, 101, 110] attempt to capture parameter variability, but they make linear assumptions about non-stationarity, which limits their expressiveness and restricts them to only 1-2 varying parameters.

Beyond the understanding that ETAS is misspecified, there are also difficulties and inefficiencies with fitting and forecasting. To estimate the intensity or branching structure, ETAS sums over all previous earthquakes, which requires substantial memory and slows the fitting process and forecasting simulations. For large earthquakes in the past this is important, because their contribution can last more than 100 years [215]. However, for smaller earthquakes particularly found in enhanced catalogs, one expects the contribution to be close to zero after a far shorter amount of time, making summing over these terms inefficient [77, 123]. Nonetheless, the time complexity $\mathcal{O}(n^2)$ remains increasingly challenging for parameter inference, particularly for Bayesian inference where recently earthquake catalogs have grown to a size that has overtaken what is computationally feasible to fit an ETAS model to.

Furthermore, a particular difficulty with fitting the ETAS model is that there needs to be a reliable estimate of the completeness across the time of the catalog and this needs to be incorporated into the model. Failing to do so will result in biases [66, 183, 245]. Methods that attempt to do this either truncate the periods of time where the catalog is most incomplete [69, 94], leading to parameters that can be dominated by a few aftershock sequences, or

attempt to model the data incompleteness itself [65–67, 133, 154], using a an additional probabilistic detection model,

$$v(t, m|\mathcal{H}_t) = \lambda(t, m|\mathcal{H}_t)\ r(t, m), \tag{1.32}$$

where $v(t, m|\mathcal{H}_t)$ is the intensity function of the observed process and $r(t, m)$ is the probability of detecting an earthquake of magnitude m at time t. Whilst these approaches offer a good estimate of the true earthquake rate from the observed earthquake rate, they assume that there is no triggering contribution from undetected events. Although there are biases from ignoring this type of triggering [196], estimating the parameters of such a model is challenging since it involves integrating over a large space of unobserved variables, rendering the likelihood intractable.

Key Question:
How do we define forecasting models that are robust to catalog incompleteness?

1.4 Neural Point Processes

Neural Point Processes (NPPs) represent the integration of machine learning and deep learning techniques into traditional parameterized point process models. By relaxing the parameterization constraints, NPPs aim to enhance the modeling and forecasting capabilities of point process models in combination with the increased amount of available data.

1.4.1 Temporal Neural Point Processes

The major step in the development of neural point processes was the use of a recurrent neural network (RNN) to learn a compact representation of the history of events, first introduced by Du et al. [39] for temporal point processes. The sequential nature of the way data pass through RNNs makes them an ideal modeling tool for temporal data. Instead of directly summing over all past events, as in models based on the Hawkes process [76], a fixed length vector representation of the past is learnt and updated at each new time step. In this approach, an input representing the inter-event times $\tau_i = t_{i+1} - t_i$ is first fed into the RNN. A hidden state $\mathbf{h}_i$ of the RNN is updated

$$\mathbf{h}_i = \sigma(W^h \mathbf{h}_{i-1} + \mathbf{w}^\tau \tau_i + \mathbf{b}^h) \tag{1.33}$$

where $\{W^h, \mathbf{w}^\tau, \mathbf{b}^h\}$ are learnable parameters, and σ is an activation function. Other types of sequence encoding were later adopted such as Long Short-Term Memory (LSTM) [127], Gated

Recurrent Unit (GRU) [32], and Transformer [247], with improved empirical performance. The conditional intensity function is then formulated as a function of the elapsed time from the most recent event and is dependent on the hidden state of the RNN,

$$\lambda(t|H_t) = \phi(t - t_i|\mathbf{h}_i),$$

where ϕ is a non-negative function referred to as the hazard function and t_i is the time of the most recent event. Parameter estimation and next-event forecasting can then be performed using the auto-regressive log-likelihood function

$$\log p(\mathcal{H}_T) = \sum_i \left[\log \phi(t_{i+1} - t_i|\mathbf{h}_i) - \int_0^{t_{i+1}-t_i} \phi(\tau|\mathbf{h}_i) d\tau \right], \tag{1.34}$$

where either $\phi(\cdot)$ is chosen such that it can be integrated analytically [39], or left unrestricted and integrated numerically [87, 105, 210, 231].

To avoid numerical integration whilst still retaining expresivity, Omi et al. [157] model the integral of the hazard function with a fully connected neural network,

$$\Phi(\tau|\mathbf{h}_i) = \int_0^{\tau} \phi(s|\mathbf{h}_i) ds.$$

With the construction of the model in this way, the log-likelihood of observing a sequence of event times no longer requires integration,

$$\log L(\{t_i\}) = \sum_i \left[\log \frac{\partial}{\partial \tau} \Phi(\tau_i|\mathbf{h}_i) - \Phi(\tau_i|\mathbf{h}_i) \right].$$

Instead, $\frac{\partial}{\partial \tau}\Phi(\cdot)$ can easily computed through neural network back-propagation [218]. Sampling from this neural point process formulation can be achieved through inverse transform sampling (1) using numerical root finding to find the inverse $\Phi^{-1}(\cdot|\mathbf{h}_i)$.

Rather than defining a flexible intensity (or cumulative intensity) function, instead Shchur et al. [190] directly model the pdf of the next event with a mixture of log-normal distributions,

$$p(\tau|\mathbf{w}, \mu, \mathbf{s}) = \sum_{k=1}^{K} w_k \frac{1}{\tau s_k \sqrt{2\pi}} \exp\left(-\frac{(\log \tau - \mu_k)^2}{2s_k^2} \right), \tag{1.35}$$

where the parameters of the pdf encode the dependence on the history vector $\mathbf{h}_i$,

$$\mathbf{w}_i = \text{softmax}(\mathbf{V}_\mathbf{w}\mathbf{h}_i + \mathbf{b}_\mathbf{w}), \quad \mathbf{s}_i = \exp(\mathbf{V}_\mathbf{s}\mathbf{h}_i + \mathbf{b}_\mathbf{s}), \quad \mu_i = \mathbf{V}_\mu \mathbf{c}_i + \mathbf{b}_\mu \tag{1.36}$$

where the softmax and exp transformations are applied to enforce the constraints on the distribution parameters, and $\{\mathbf{V_w}, \mathbf{V_s}, \mathbf{V}_\mu, \mathbf{b_w}, \mathbf{b_s}, \mathbf{b}_\mu, \}$ are learnable parameters. This approach allows fast auto-regressive sampling from the mixture model,

$$\mathbf{z} \sim \text{Categorical}(\mathbf{w}), \quad \epsilon \sim \text{Normal}(0, 1), \quad \tau = \exp(\mathbf{s}^T\mathbf{z} \cdot \epsilon + \mu^T\mathbf{z}) \tag{1.37}$$

None of the aforementioned temporal NPPs are directly suitable for earthquake forecasting. They either do not consider marked point processes or they consider only catagorical marks rather than continuous ones.

1.4.2 Spatio-temporal Neural Point Processes

Several efforts have extended temporal neural point processes to the spatial domain. Whilst the encoding of spatio-temporal sequences is done through including spatial locations as additional features in sequential neural networks (RNN,GRU,LSTM,Transformer), approaches have focused on how to accurately "decode" the representation of the history and model the time-dependent spatial distribution $p(\mathbf{x}|t, \mathcal{H}_t)$.

NSTPP [19]

Chen et al. [19] use a continuous time representation of the hidden state repressenting the history,

$$\lambda_g^*(t) = g_\lambda(\mathbf{h}_t), \tag{1.38}$$

where the hidden state is mapped to the intensity through a neural network $g_\lambda(\cdot)$ (with enforced positivity). The hidden state is governed by an Ordinary Differential Equation (ODE),

$$\mathbf{h}_{t_0} = \mathbf{h}_0 \tag{1.39}$$

$$\frac{d\mathbf{h}_t}{dt} = f_h(t, \mathbf{h}_t) \tag{1.40}$$

$$\lim_{\epsilon \to 0} \mathbf{h}_{t_i+\epsilon} = g_h\left(t_i, \mathbf{h}_{t_i}, \mathbf{x}_{t_i}^{(i)}\right), \tag{1.41}$$

where $f_h(\cdot)$ is a standard multi-layer fully connected neural network and $g_h(\cdot)$ uses the GRU update (similar to 1.33).
The spatial component is also modelled with continuous dynamics,

$$\log p(\mathbf{x}_{t_i}^{(i)}|t_i, \mathcal{H}_{t_i}) = \log p(\mathbf{x}_0^{(i)}) - \int_0^{t_i} \text{tr}\left(\frac{\partial f_\mathbf{x}}{\partial x}(\tau, \mathbf{x}_\tau^{(i)}, \mathbf{h}_{t_i})\right) d\tau, \tag{1.42}$$

where a Transformer architechture is used for $f_\mathbf{x}(\cdot)$.
Training the model requires backpropagating through an ODE solver, details of which can be found in Chen et al. [18].

DeepSTPP [240]

Zhou et al. [240] encode the history sequence to a latent multivariate Gaussian process,

$$z_i \sim q_\phi(z_i|\mathbf{h}_i) = \mathcal{N}(\mu, \text{Diag}(\sigma)), \tag{1.43}$$

where the mean μ and covariance $\text{Diag}(\sigma)$ are the outputs of a transformer encoding, $\mathbf{h}$. The intensity function is then constructed,

$$\lambda^*(t, \mathbf{x}|z) = \sum_{t_i<t} w_i k_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i; \gamma_i) k_t(t, t_i; \beta_i), \tag{1.44}$$

where $w_i(z), \gamma_i(z), \beta_i(z)$ are parameters conditioned on the latent process and

$$k_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_i) = \alpha^{-1} \exp(-\gamma_i||\mathbf{x} - \mathbf{x}_i||), \quad k_t(t, t_i) = \exp(-\beta_i||t - t_i||). \tag{1.45}$$

AutoSTPP [239]

Zhou and Yu [239] use the spatio-temporal Hawkes process formulation

$$\lambda^*(t, \mathbf{x}) = \mu + \sum_{t_i<t} f_\phi^+(t - t_i, \mathbf{x} - \mathbf{x}_i, \mathcal{H}_{t_i}), \tag{1.46}$$

where f_ϕ^+ is a flexible positive scalar function. In a similar approach to Omi et al. [157], Zhou and Yu [239] jointly model the integral,

$$F_\phi(t, \mathbf{x}, \mathbf{h}) = \int_{\mathcal{S}} \int_0^t f_\phi^+(\tau_i, \mathbf{x} - \mathbf{x}_i, \mathbf{h}) d\tau d\mathbf{x} \tag{1.47}$$

and the influence function $f_\phi = \frac{\partial F_\phi}{\partial t \partial \mathbf{x}}$ with neural networks with shared parameters. In order to allow for 3 dimensional space-time integration of the influence function network, the spatial domain has to be be rectangular.

The pace of NPP development is fast in the machine learning community, and while some initial benchmarking of these spatio-temporal NPP models has been conducted on an earthquake dataset in Japan, these experiments lack relevance for stakeholders in the seismology community. The benchmark in use lacks a key earthquake sequence from the region, fails to recreate an operational setting with proper train-test splits, and doesn't compare against state-of-the-art models like ETAS.

Key Question:
How do we ensure that future NPP development is directly relevant for application in seismology?

1.5 Outline

The remainder of this thesis provides some answers to all the key questions we have encountered.

In Chapter 2, I extend an existing temporal NPP [157] to the magnitude domain and show how this model can forecast earthquakes above a target magnitude threshold whilst being dependent on smaller magnitude earthquakes. This allows for forecasts to be made using new low magnitude data from an enhanced catalog from the Central Apennines earthquake sequence [206]. This includes lowering the threshold down below $M_c(t)$ for some $t \in [0, T]$, exploring the model's robustness to incomplete parts of the earthquake sequence. I compare the NPP model's forecasting performance against ETAS as well as the models' capacity for the large volume of data.

In Chapter 3, I address Bayesian inference for the temporal ETAS model. I construct an approach (SB-ETAS) for learning the posterior distributions of ETAS parameters using repeated simulations instead of the computationally expensive ETAS likelihood function. By specifying a model through simulation rather than the likelihood, this approach broadens the scope of available models to encompass greater complexity. This could include more complicated extensions of ETAS and models of earthquakes that include data incompleteness and physics based simulators. The approach scales $\mathcal{O}(n \log n)$ with the number of earthquakes, compared to the $\mathcal{O}(n^2)$ achieved by previous methods.

In Chapter 4, I create a platform for benchmarking existing and future spatio-temporal NPPs against state-of-the-art models from seismology. The platform hosts datasets covering various regions of California, representing typical forecasting zones and encompassing data commonly utilized by forecast issuers. Moreover, employing modern techniques, some datasets include smaller magnitude earthquakes, exploring the potential of numerous small events to enhance forecasting performance. Although the datasets are derived from publicly available raw data, I pre-process and configure them within the platform to facilitate future benchmarking that is directly relevant to seismology.

Chapter 2

Forecasting the 2016–2017 Central Apennines Earthquake Sequence With a Neural Point Process

Declaration

The methodology, experimentation and writing of this chapter was undertaken by me, Samuel Stockman, with guidance from my two supervisors: Maximilian Werner and Daniel Lawson. This work benefited from external reviews from Leila Mizrahi and an anonymous reviewer, with all changes made by me.

The following chapter was published in Earth's Future on September 11th 2023 [201]: Stockman, Samuel, Daniel J. Lawson, and Maximilian J. Werner. "Forecasting the 2016–2017 Central Apennines earthquake sequence with a Neural Point Process." Earth's Future 11.9 (2023): e2023EF003777.

This work appears in this thesis in near-identical format to the original publication. The Supplementary Material has been included as Appendix A.

Abstract

Point processes have been dominant in modeling the evolution of seismicity for decades, with the Epidemic-Type Aftershock Sequence (ETAS) model being most popular. Recent advances in machine learning have constructed highly flexible point process models using neural networks to improve upon existing parametric models. We investigate whether these flexible point process models can be applied to short-term seismicity forecasting by extending an existing temporal neural model to the magnitude domain and we show how this model can forecast earthquakes above a target magnitude threshold. We first demonstrate that the neural model can fit synthetic ETAS data, however, requiring less computational time because it is not dependent on the full history of the sequence. By artificially emulating short-term aftershock incompleteness in the synthetic dataset, we find that the neural model outperforms ETAS. Using a new enhanced catalog from the 2016-2017 Central Apennines earthquake sequence, we investigate the predictive skill of ETAS and the neural model with respect to the lowest input magnitude. Constructing multiple forecasting experiments using the Visso, Norcia and Campotosto earthquakes to partition training and testing data, we target M3+ events. We find both models perform similarly at previously explored thresholds (e.g., above M3), but lowering the threshold to M1.2 reduces the performance of ETAS unlike the neural model. We argue that some of these gains are due to the neural model's ability to handle incomplete data. The robustness to missing data and speed to train the neural model present it as an encouraging competitor in earthquake forecasting.

Plain Language Summary

For decades, the Epidemic-Type Aftershock Sequence (ETAS) model has been the most popular way of forecasting earthquakes over short time spans (days/weeks). It is formulated mathematically as a point process, a general class of statistical model describing the random occurrence of points in time. Recently the machine learning community have used neural networks to make point processes more expressive and titled them neural point processes. In this study we investigate whether a neural point process can compete with the ETAS model. We find that the two models perform similarly on computer simulated data; however, the neural model is much faster with large datasets and is not hindered if there is missing data for smaller earthquakes. Most earthquake catalogs contain missing data due to varying capability in our detection methods, therefore we need models that are robust to this missingness. We then find that the neural model outperforms ETAS on a new catalog for the 2016-2017 Central Apennines earthquake sequence, which through machine learning detection contains thousands of previously undetected small magnitude events. We argue that some of this improvement can in fact be explained by missing data. These results present neural point processes as an encouraging competitor in earthquake forecasting.

2.1 Introduction

The construction of machine learning algorithms for detecting the arrival times of earthquake phases (eg. [242]) combined with an accelerated growth in the number of seismic sensors, has meant that sizes of earthquake catalogs have grown substantially [99]. With the amount of available seismicity data increasing, current forecasting methods that include the full history of the catalog in the form of all event pairs are increasingly inefficient and might not be flexible enough to incorporate this additional data, thus the need for the application of methods developed in the machine learning community is becoming more apparent. However, there exists a disconnect between the tools used by statistical seismologists and those in the machine learning community that apply their methods to seismic data. This work attempts to bridge that gap (to some extent) by considering a machine learning variant of point processes. Point processes are a class of models that contain the Epidemic-Type Aftershock Sequence (ETAS) model, a widely accepted and used point process model for earthquakes [118, 124, 150, 151]. In working with a machine learning method with a conditional intensity function (the function that explicitly defines a point process) [244], we present a model that is directly comparable to ETAS models, the current benchmark for short-term earthquake forecasting, yet now with desirable properties such as flexibility and scalability. The machine learning variant of point processes we introduce are known as neural point processes.

In this work we extend the architecture introduced by Omi et al. [157] so that it may also deal with earthquake magnitudes. For this we require a model that is dependent on previous event magnitudes, can forecast subsequent magnitudes as well as forecast earthquakes above a threshold magnitude. Distinguishing between a threshold for the input magnitude and a threshold for the target earthquakes is a problem specific requirement for earthquake forecasting, so does not exist in other works on temporal point processes. We choose to extend Omi et al. [157] to give the most flexible representation of the intensity, since they use a fully non-parametric approach compared to other intensity based methods that use a semi-parametric approach. Working with the intensity function rather than directly modelling the likelihood of the next event provides a model that is closer in interpretation to ETAS and provides a natural way to forecast earthquakes above some target threshold magnitude, detailed in section 2.3.2.

To benchmark our proposed neural point process with ETAS, we design forecasting experiments on both synthetic data as well as a new enhanced catalog for the 2016–2017 Amatrice–Visso–Norcia (AVN) seismic sequence. The catalog generated by Tan et al. [206], containing roughly 900,000 earthquakes, was generated using a deep neural network based phase picker for earthquake arrival times [242]. As a result, the size of the catalog increased 10 fold from the routine catalog generated by the Italian National Institute of Geophysics and

Volcanology (INGV). The INGV catalog has been used in several retrospective forecasting experiments [40, 118, 120, 125], but there has yet to be much development of forecasting models using enhanced catalogs such as this one and, given that they contain considerably more earthquakes, investigations into how we can harness these machine learning generated enhanced catalogs are essential. The AVN sequence contains ten $M5+$ events during a five month period over an 80 km long normal-fault system [118]. The number of large earthquakes as well as the compactness of the region on which they occur make this sequence preferable for testing purely temporal forecasting models which contain no spatial covariates. We do still expect some loss of information by ignoring the spatial covariates, particularly for smaller earthquakes where there is a spatial extent across which earthquakes won't interact.

We seek to understand how taking different magnitude thresholds and temporal partitions of our datasets affects the performance of the two models. Through altering these two aspects, we naturally change the amount of data shown to each model so that we may see how sensitive their forecasting performance is to training sample size [222]. Through partitioning in time we can see how the performance is affected by the number of major earthquakes that each model is trained on. By altering the magnitude threshold of the input catalog, we seek to improve the predictive skill of forecasts by using the hypothesis that small earthquakes should help to forecast the moderate-to-large earthquakes. Either from a time-independent perspective where large earthquakes are found to nucleate in areas that have a large density of small events [92, 93], or in time-dependent forecasting (eg. ETAS and CRS) where earthquake triggering is believed to exist at all scales [77, 122, 141], reducing the threshold of the input catalog generally leads to improved forecasting performance of moderate-to-large events [78, 80, 120, 224].

Particularly, Mancini et al. [120] consider the same sequence as this study, and compare forecasting results from models trained on several different enhanced catalogs (including the Tan et al. [206] catalog used in this study). They find that the forecasting of M3+ events by CRS and ETAS models is not improved by training on the enhanced catalogs. When using the same catalog as this study, however, they see the models increase in performance as the input magnitude threshold is lowered from M5 to M3, but at the lowest two thresholds M1 and M2, the performance of ETAS is worse than for M3+. Direct comparison of results, however, isn't possible as they report information gains that are for spatio-temporal forecasts using the Poisson assumption of earthquake rates in gridded space-time windows. This study hopes to provide some further insight into the performance of forecasting models using the low magnitude earthquakes found in this catalog and presents neural point processes as a competitive model using such events.

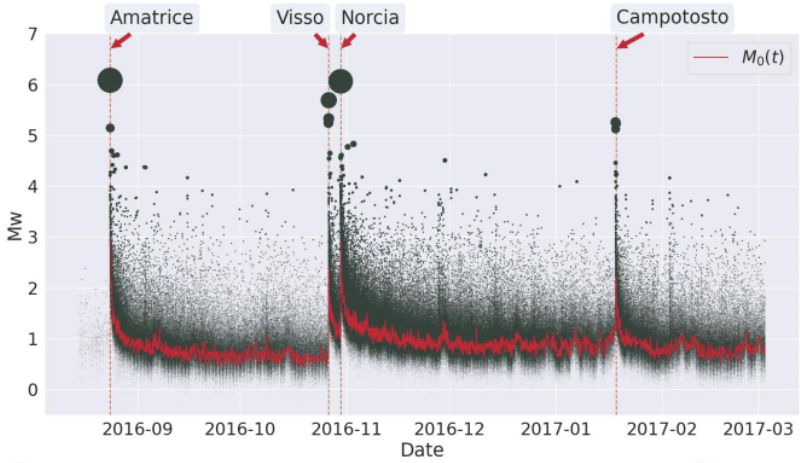


Figure 2.1: The magnitudes and times of the AVN sequence 2016-2017 [206] used to evaluate the performance of the neural and ETAS model. Marked with a dashed red line are the times of the 4 major events of the sequence. The size of the points are plotted on a log scale corresponding to M_{w}. An estimate of the temporal completeness $M_c(t)$ is plotted using the maximum curvature method [226].

2.2 Data

To benchmark our neural point process against ETAS we conduct forecasting experiments on both real data from the Amatrice-Visso-Norcia sequence as well as synthetic data generated by ETAS. Since we are making comparisons about temporal models, in both catalogs, we remove all spatial covariates.

2.2.1 Amatrice-Visso-Norcia High Resolution Catalog

On the 24th of August 2016 a $\text{M}_\text{w}6.0$ earthquake was recorded near the town of Amatrice in northern Lazio, central Italy. It was followed by a $\text{M}_\text{w}5.9$ near the town of Visso on the 26th of October and a $\text{M}_\text{w}6.5$ near the town of Norcia four days later. Finally, in January 2017, four events between $\text{M}_\text{w}5.0$ and $\text{M}_\text{w}5.5$ struck the Campotosto area. Figure 2.1 depicts the evolution of this seismic sequence over time.

The INGV produced a routine catalog for the 1 year period containing this sequence [20]. Their catalog contains roughly 82,000 events with a completeness of $Mc = 2.3$ [118]. With the use of a neural network based phase picker to determine P and S-wave arrival times, an enhanced catalog has been created for the same earthquake sequence [206]. This catalog contains around 900,000 events and has an overall value of completeness of the catalog $Mc = 0.2$. We estimated

the time varying completeness of this catalog using the maximum curvature method [226] with samples of 300 events and can see clear variation in completeness particularly following large magnitude earthquakes. This approximate method for the completeness is only used to show the variability across the catalog and is not used directly in any modelling.

2.2.2 Synthetic Catalog

For the forecasting experiment of synthetic ETAS data, we generate a dataset using the simulator by [132], with uniform background intensity μ and triggering function,

$$g(t,x,y,m) = \frac{k_0 e^{a(m-M_c)}}{\frac{(t+c)^{1+\omega}}{e^{t/\tau}}((x^2+y^2)+de^{\gamma(m-M_c)})^{1+\rho}}.$$

The ETAS parameters ($\log_{10}\mu = -6.6,\ \log_{10}k_0 = -3.15,\ a = 2.85,\ \log_{10}c = -2.95,\ \omega = -0.03,\ \log_{10}\tau = 3.99,\ \log_{10}d = -0.35,\ \gamma = 1.22,\ \rho = 0.51,\ M_c = 1\ M_w$) are taken close to Mizrahi et al. [132] with higher background rate to account for the lower M_c. The resulting dataset of roughly 250,000 events comes from removing all the spatial covariates.

We also generate a second synthetic dataset from the first by emulating short-term aftershock incompleteness using the time-dependent formula from Helmstetter et al. [79],

$$M_c(M,t) = M - 4.5 - 0.75\log(t),$$

where M is the mainshock magnitude. Events below the function are removed using the six largest events as mainshocks in this synthetic catalog.

2.3 Methods

Since the neural point process introduced by Omi et al. [157] is purely temporal, to model seismicity we must extend their model to our requirements. Particularly we require that forecasts be dependent on the history of both times and magnitudes, as magnitudes are an important predictor of seismicity [212, 213]. We also require a forecast over the next magnitude where, unlike the Gutenberg-Richter law [64] routinely assumed in the ETAS framework, this is also dependent on the history of events. Finally, we require that we may make forecasts of earthquakes above some target magnitude threshold despite a dependence on earthquakes below that target threshold. In Section 2.3.1, we first extend the structure of the neural network by Omi et al. [157] to maximise the likelihood of observing a marked sequence of events, including constructing a time-history dependent magnitude distribution. In Section 2.3.2, we show how we adjust this new structure to target events above a magnitude threshold.

To aid in the development of more flexible forecasting models, we will make both the dataset and models used in this study available after publication on `https://github.com/ss15859/Neural-Point-Process`.

2.3.1 Continuously Marked Neural Point Process

We begin with the factorisation of the joint conditional intensity function into its marginal intensity and conditional density function, following Daley et al. [29],

$$\lambda^*(t, m) = \lambda^*(t) f^*(m|t),$$

where $\lambda^*(t)$ is the marginal conditional intensity function of t, and $f^*(m|t)$ is the conditional density function of the mark at time t. Both of these functions are dependent on the history H_t, here denoted by the asterisk *. To construct the likelihood for the marked sequence we model these two functions separately.

With the factorisation, the expression for the log-likelihood of observing the marked sequence of events (1.7) becomes,

$$\log L(\{t_i, m_i\}) = \sum_i \left[\log \lambda^*(t_i) + \log f^*(m_i|t_i) - \int_{t_{i-1}}^{t_i} \lambda^*(t) dt \right]. \tag{2.1}$$

Now with a two dimensional input, the hidden state of the RNN is updated as a linear combination of the inter-event times and magnitudes. This is the continuous mark extension to the RNN update from Du et al. [39],

$$\mathbf{h}_i = \sigma(W^h \mathbf{h}_{i-1} + \mathbf{w}^\tau \tau_{i-1} + \mathbf{w}^m m_{i-1} + \mathbf{b}^h),$$

where $\mathbf{w}^m$ is an additional learnable parameter.

The marginal intensity function is formulated as a function of the elapsed time from the most recent event and is dependent on the hidden state of the RNN [39],

$$\lambda(t|H_t) = \phi(\tau = t - t_i|\mathbf{h}_i), \tag{2.2}$$

where ϕ is a non-negative function referred to as the hazard function and t_i is the time of the most recent event. Following Omi et al. [157], we model the cumulative hazard function using a fully connected neural network,

$$\Phi(\tau|\mathbf{h}_i) = \int_0^\tau \phi(s|\mathbf{h}_i) ds.$$

which allows us to differentiate this with respect to τ to extract the hazard function. The derivative is easily obtained through automatic differentiation [218], which is available in all neural network packages.

We now also formulate the conditional density function of the mark at time t as a function of the current mark. This is dependent on the time since the most recent event and the hidden state of the RNN,

$$f(m|t, H_{t_i}) = \psi(m|\tau, \mathbf{h}_i).$$

We again model its cumulative distribution with a fully connected neural network,

$$\Psi(m|\tau, \mathbf{h}_i) = \int_0^m \psi(\mu|\tau, \mathbf{h}_i) d\mu.$$

Although this integral does not feature in the expression for the log-likelihood, we still opt for this approach over directly modelling the density function with a neural network. This follows from work on neural density estimation where positive weights can be enforced in the network to capture the positivity and monotonicity of cumulative distribution functions [22]. We can then obtain the density function again through automatic differentiation.

We can now write the log-likelihood as:

$$\log L(\{t_i, m_i\}) = \sum_i \left[\log \lambda^*(t_i) + \log f^*(m_i|t_i) - \int_{t_{i-1}}^{t_i} \lambda^*(t) dt \right] \tag{2.3}$$

$$= \sum_i \left[\log \phi(\tau_i|\mathbf{h}_i) + \log \psi(m_i|\tau_i, \mathbf{h}_i) - \int_0^{t_i - t_{i-1}} \phi(t|\mathbf{h}_i) dt \right] \tag{2.4}$$

$$= \sum_i \left[\log \frac{\partial}{\partial \tau} \Phi(\tau_i|\mathbf{h}_i) + \log \frac{\partial}{\partial m} \Psi(m_i|\tau_i, \mathbf{h}_i) - \Phi(\tau_i|\mathbf{h}_i) \right]. \tag{2.5}$$

We model both the cumulative hazard function Φ, and the conditional distribution function of the marks Ψ, using a feed-forward neural network. The network depicted in Figure 2.2 consists of four component parts. The first part is the recurrent section, which finds an encoding of the history of the point process. The output of the recurrent section $\mathbf{h}_i$ passes into two fully connected components. One models the integral of the intensity function, this is a function of the time of the next event τ. The other component models the integral of the conditional density function of the next mark, this is dependent on the next time τ, but is a function of the next mark m. The structure of the network describes the dependence relationship of the point process, represented by a connection between sections in the diagram. For example, the magnitude network models the conditional cumulative magnitude distribution: a function that depends on the observed magnitude m_i, time since the last event τ_i and the history $\mathbf{h}_i$.

Since passing long sequences into recurrent neural networks can often lead to exploding or vanishing gradient problems [85], we do not pass the whole history of the point process into the

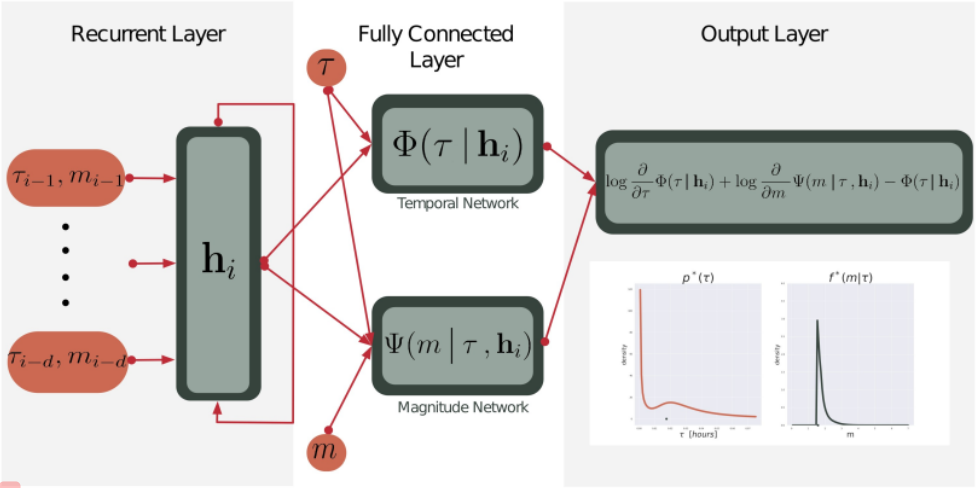


Figure 2.2: The proposed network comprises four sections. First, the inter-event times and magnitudes of the last d events are fed into a recurrent section consisting of 64 recurrent units. The output of this section is fed into two fully connected sections where it is combined with the next inter-event time τ for the temporal network and additionally with the next magnitude m for the magnitude network. The outputs of both these sections are combined to formulate the log-likelihood of the next inter-event time and magnitude $\{\tau, m\}$. We can separate the temporal and magnitude terms in this likelihood to give point evaluations of the density of the next inter-event time and conditional density of the next magnitude. The dependence structure of the point process is expressed by the connections between sections in the network.

recurrent section, but the past d events. This implies that we use only the past d events to forecast the next event. Thus, this model is learning to estimate the intensity given a recent history of d events. This hyperparameter is kept the same as Omi et al. [157] at $d = 20$. A naive tuning search found no significant improvement at larger values of $\{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$ (Figure A.4). This difference to the full-history ETAS model is discussed in Section 2.5.1.

We enforce positive weights in both fully connected sections of the network to capture the positivity and monotonicity that is required by both cumulative functions. A sigmoid function is applied to the final output of the magnitude network to map to the $[0, 1]$ interval. In the final component of the network we formulate the output as the log-likelihood of observing the pair $\{\tau, m\}$. To construct this output, one backward pass is performed to calculate the derivatives with respect to the next time and the next magnitude found in eq (2.5). This output is exactly what is maximised to learn the weight parameters, during a second backwards pass.

2.3.2 Target Events

The growth in machine learning generated catalogs from their predecessors is found through detecting events of magnitude much lower than previously possible. However, operationally, we may not care about the forecasting of these smaller events if it is the larger ones that are the most hazardous. The aim is therefore to use the smaller events to forecast earthquakes above some target threshold. In this section we outline how this is done for both ETAS and the neural model. We hereafter call events above the target magnitude threshold M_d target events.

Let $\{(t_i, m_i)\}_{i=1}^n \in [0, T] \times \mathcal{M}$ be the entire sequence of events, complete down to M_c. We seek to make forecasts of events above magnitude M_d. This corresponds to the sequence:

$$\{(t_j, m_j) : m_j \geq M_d\}_{j=1}^k.$$

To ease in the distinction between 'all events' and the target events, we subscript the former with i and the latter with j.

Let $\lambda_0(t, m|H_t)$ denote the joint intensity function of all events above M_c. We seek to learn the intensity of events above M_d, denoted $\lambda_d(t, m|H_t)$, where the history H_t contains all events $\{(t_i, m_i)\}_{i:t_i<t}$. The ground intensity above the target threshold is found by marginalising the joint intensity over the target magnitude region,

$$\lambda_d(t|H_t) = \int_{M_d}^{\infty} \lambda_0(t, m|H_t) dm.$$

The log likelihood of target events is then given by:

$$\log L(\{t_j, m_j\}) = \sum_{j:m_j \geq M_d} \left(\log \lambda_d(t_j, m_j|H_{t_j}) - \int_{t_{j-1}}^{t_j} \lambda_d(t|H_t) dt \right).$$

For ETAS, the rate above magnitude M_d is a fraction of the rate above M_c, due to the independent distribution for magnitudes,

$$\lambda_d(t|H_t) = \int_{M_d}^{\infty} \lambda_0(t, m) dm = \left(\int_{M_d}^{\infty} f_{GR}(m) dm \right) \lambda_0(t|H_t) = p_d \cdot \lambda_0(t|H_t),$$

where $f_{GR}(m)$ is the Gutenburg-Richter law and p_d is simply the probability that $m \geq M_d$. Therefore the expression for the likelihood is relatively simple,

$$\log L(\{t_j, m_j\}) = \sum_{j:m_j \geq M_d} \left[\log \left(p_d \cdot \lambda_0(t_j, m_j|H_{t_j}) \right) - \int_{t_{j-1}}^{t_j} p_d \cdot \lambda_0(t|H_t) dt \right].$$

For the neural model, we make use of the fact that the integral of the intensity function between target events, $\{(t_j, m_j) : m_j \geq M_d\}_{j=1}^k$, can be expressed as a sum of disjoint integrals

between all events $\{(t_i, m_i)\}_{i=1}^n$,

(2.6)

$$\log L(\{t_j, m_j\}) = \sum_{j:m_j \geq M_d} \left[\log \lambda_d(t_j, m_j | H_{t_j}) - \int_{t_{j-1}}^{t_j} \lambda_d(t|H_t) dt \right]$$

(2.7)
$$= \sum_{j:m_j \geq M_d} \left[\log \lambda_d(t_j|H_{t_j}) + \log f_d(m_j|t_j, H_{t_j}) \right] - \int_{t_0}^{t_k} \lambda_d(t|H_t) dt$$

(2.8)
$$= \sum_{\substack{i:m_i \geq M_c \\ t_i \leq t_k}} \left[(\log \lambda_d(t_i|H_{t_i}) + \log f_d(m_i|t_i, H_{t_i})) \mathbf{I}\{m_i \geq M_d\} - \int_{t_{i-1}}^{t_i} \lambda_d(t|H_t) dt \right]$$

(2.9)
$$= \sum_{\substack{i:m_i \geq M_c \\ t_i \leq t_k}} \left[\left(\log \frac{\partial}{\partial \tau} \Phi(\tau_i|\mathbf{h}_i) + \log \frac{\partial}{\partial m} \Psi(m_i|\tau_i, \mathbf{h}_i) \right) \mathbf{I}\{m_i \geq M_d\} - \Phi(\tau_i|\mathbf{h}_i) \right],$$

where between (2.6) and (2.7) we have factorised the joint intensity of events above M_d, $\lambda_d(t, m|H_t)$, into the ground intensity above the target threshold, $\lambda_d(t|H_t)$, and the distribution of the next magnitude above the target threshold given the time and the history, $f_d(m|t, H_t)$.

Between (2.7) and (2.8) we changed the summation from being over target events to being over all events by adding the indicator function $\mathbf{I}\{m_i \geq M_d\}$. The integral in (2.7) becomes the sum of integrals in (2.8) through a decomposition into disjoint integrals between all events $\{(t_i, m_i)\}_{i=1}^n$. Thus the neural model may target events by adding the indicator function $\mathbf{I}\{m_i \geq M_d\}$ to the expression for the log-likelihood. Now the hazard function models the rate above M_d as a function of the time from the last event (of any magnitude),

$$\lambda_d(t|H_t) = \phi(\tau = t - t_i|\mathbf{h}_i).$$

2.3.3 Experimental Design

For both the synthetic data and real data we apply the same training and testing procedure illustrated in Figure 2.3. At a fixed point in time along the sequence we set a marker and train on data up to that point. Following that, the remainder of the sequence will be used as the test set. For the synthetic catalogs this is done at one single point in time, whereas for the Central Apennines sequence we make three partitions - each just before the Visso, Norcia and Campotosto earthquakes. By making these partitions we can see how the performance of each model is affected by the number of training datapoints as well as the number of major earthquakes.

We seek to understand how different magnitude thresholding affects the performance of each of the models. For each of the partitions, we look at the performance of both models as the magnitude threshold of the input catalog is lowered, a parameter we refer to as Mcut. For the

AVN catalog and incomplete synthetic catalog, this includes lowering Mcut into periods where Mcut$< M_c(t)$. We keep fixed the magnitude of events we wish to target at $\mathrm{M}_d = 3$ Mw.

Both models are trained by maximum likelihood estimation (MLE) on the training dataset. For ETAS we use the intensity function defined by Ogata [150] and maximise the likelihood through Nelder-Mead optimisation, chosen for its robustness. For the neural model, we maximise the likelihood defined in equation (2.9) through ADAM optimisation [98] written in Tensorflow [1]. The neural model uses validation data during training for early stopping [54]. This validation set comprises 20% of randomly sampled points from the training data.

We compare the two models' performance through the log-likelihood of the events in the testing set. We separate the temporal and magnitude terms in the likelihood equation (2.1), to analyse their predictive skills on the two target variables separately. To compare the performance across different magnitude thresholds, we also fit a homogeneous Poisson model. For the temporal log-likelihood we can therefore present the log-likelihood gain from a benchmark Poisson model, whereas for the magnitude forecasts, simply the log-likelihood is reported. We shall refer to both performance metrics as log-likelihood scores and to make general comparisons across the models, we compare the mean log-likelihood score per earthquake as well as construct a 95% confidence interval to assess the variability. The confidence interval is constructed with 1000 bootstrap samples of the log-likelihood scores.

2.4 Results

2.4.1 Synthetic Data

Despite the synthetic catalog being complete down to $M_c = 1$ Mw, we only lower Mcut down to 1.7 due to the computational time it takes to find the MLE parameters of ETAS for such a large dataset. Figure 2.4 shows the computation time (CPU hours) to train each of the models as a function of the size of the training set using an 2.4 GHz Intel E5-2680 v4 (Broadwell) CPU. The neural model is significantly faster to train than ETAS due to the likelihood function not being dependent on the full history of the sequence, giving it complexity $O(n)$ [192]. ETAS in contrast has complexity $O(n^2)$ due to the double sum in the likelihood.

Figure 2.5 shows 95% confidence intervals for the log-likelihood scores on the synthetic catalogs for the varying magnitude thresholds. By varying the value of Mcut the size of the training dataset changes, depicted by the green barplots. In Figure 2.5a) the temporal log-likelihood scores for both models on the complete synthetic catalog are displayed. Although there are fluctuations, there is no significant difference between the two models' log-likelihood gain from the same Poisson baseline for all values of Mcut, suggesting the neural model has learnt to

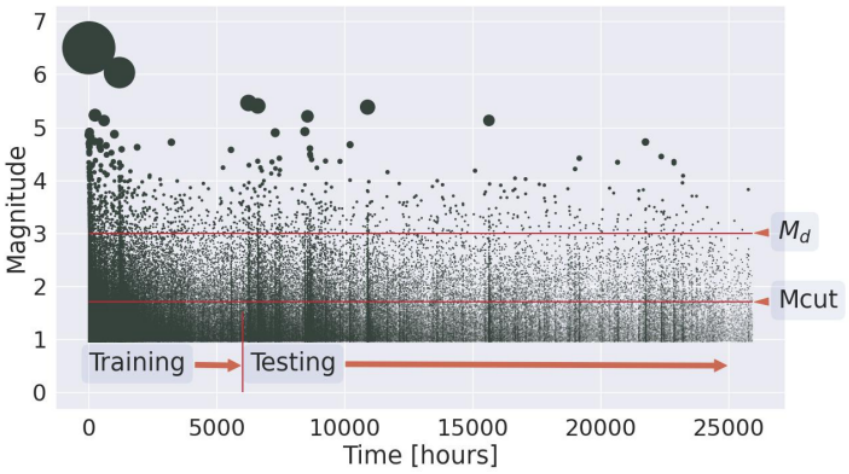


Figure 2.3: The synthetic catalog with an outline of the training and testing procedure. We train up to a fixed point in time in the catalog, following which the remainder of the catalog is used for testing. We vary the value of the threshold for the input catalog (Mcut) and keep fixed the value of the target threshold (M_d).

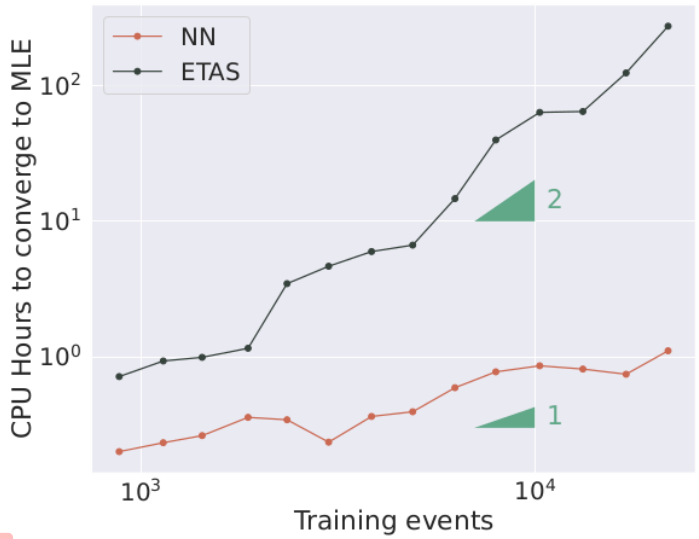


Figure 2.4: Time on a single CPU required to train each of the models as the training size increases. Each model is trained by maximising the likelihood of the training data.

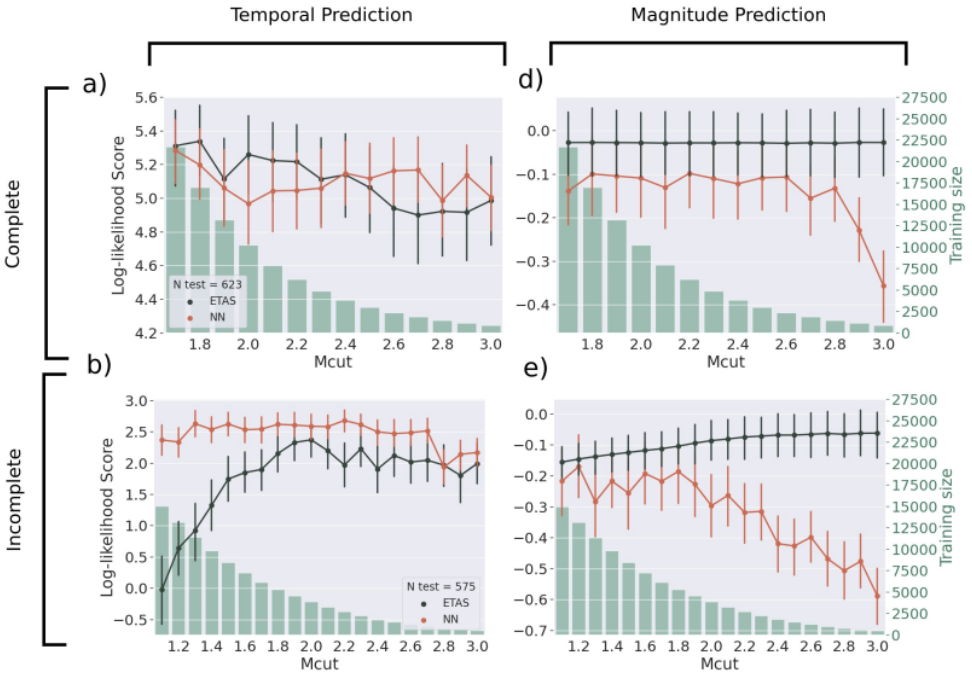


Figure 2.5: Results from the synthetic tests. 95 % confidence intervals for the log-likelihood scores for each model as a function of Mcut (the magnitude threshold of the input catalog). The size of the training set is displayed in the green barplot; the size of the testing set in the legend. a) temporal log-likelihood gain from Poisson for the complete synthetic catalog. b) temporal log-likelihood gain for the incomplete catalog. c) magnitude log-likelihood for the complete catalog. d) magnitude log-likelihood for the incomplete catalog.

capture the ETAS data sufficiently well. Although the mean of ETAS increases as we lower Mcut, whereas the mean of the neural model fluctuates, these changes are non-significant. We speculate that we do not see significant improvement as we lower Mcut due to the fact that we are fitting temporal models to spatio-temporal data.

Figure 2.5b) shows the temporal log-likelihood scores for the incomplete synthetic catalog. Just as in Figure 2.5a), ETAS remains constant down to Mcut 2.0. But now, on this incomplete dataset, the performance of ETAS significantly reduces as Mcut is lowered below this threshold. In contrast, the neural model remains constant in performance and significantly outperforms ETAS for all but the highest two thresholds, demonstrating a robustness to the missing data in this catalog. By synthetically recreating incompleteness we remove many datapoints, therefore we can fit ETAS to a lower Mcut as we do not experience the longer training times of the complete catalog.

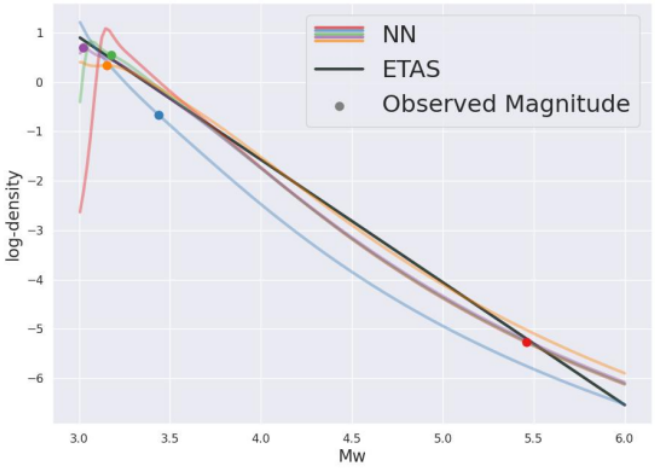


Figure 2.6: Five examples of the forecasted magnitude distributions from the complete synthetic catalog tests at Mcut = 1.7 compared with the ETAS Gutenberg-Richter law. The magnitudes of the observed events are plotted as points along the log-density for the neural model.

Figure 2.5c) shows the magnitude log-likelihood scores for the complete synthetic catalog. For the highest two thresholds, the neural model performs significantly worse than ETAS but then remains marginally worse for all other values of Mcut. For the magnitude scores for the incomplete data in Figure 2.5d), ETAS significantly outperforms the neural model at the higher thresholds. As the threshold is lowered the two perform more similarly, owing to an increase in performance from the neural model and a slight decrease from ETAS.

We can understand the marginal difference in performance between the two models at lower thresholds by looking at their respective distributions. Figure 2.6 shows five instances of the magnitude distribution learnt by both models at Mcut = 1.7 for the complete catalog. For ETAS we simply learn the b value of the Gutenberg-Richter (GR) law whereas for the neural model, a history and time-dependent distribution for the next magnitude is learnt. In these five instances, although the neural model can closely approximate the GR law, allowing it to be time-history dependent means that its predictions vary across different occurrences in the sequence and therefore in this synthetic example performs worse than the stationary data-generating distribution. Since the neural model contains orders of magnitude more parameters, this result indicates overfitting [103].

2.4.2 AVN Catalog

Figure 2.7 shows the log-likelihood scores for both models on each of the testing-training partitions on the AVN catalog, where Figure 2.7a) is for training both models up to the Visso 5.9 Mw event, Figure 2.7b) up to the Norcia 6.5 Mw event and Figure 2.7c) up to the first of the major Campotosto earthquakes. Across all training-testing partitions Figure 2.7a)-2.7c), as Mcut is lowered below 3 Mw, the performance of ETAS decreases consistently. The neural model, however, either remains constant in performance or improves as Mcut is lowered. In addition, as the neural model is trained on a longer period of time, its performance improves. For higher values of Mcut the neural model performs significantly worse than ETAS when trained up to Visso, but with the additional training data leading up to Norcia and Campotosto, it is similar to ETAS. For low values of Mcut, the performance of the neural model is significantly better than ETAS. When comparing across all values of Mcut neither model is significantly better than the other. Generally, the neural model is more robust to different values of Mcut than ETAS. Figure S1 of the Supporting Information shows how the fitted ETAS parameters change with Mcut.

The magnitude log-likelihood scores in Figure 2.7d)-2.7f) show that the time-history dependent magnitude distribution generally cannot match the predictive power of the stationary GR law. The performance of ETAS remains constant for all values of Mcut and testing-training partitions. In Figure 2.7d) and Figure 2.7e) the neural model improves in performance as Mcut is lowered, where it only performs closely to ETAS at the very lowest threshold. This and the fact that it performs much closer to ETAS when training up to Campotosto (Figure 2.7f)) suggests that it is learning and improving when shown more data.

To compare the models' performance as a function of time, Figure 2.8 displays the cumulative information gain (CIG) of the neural model over ETAS, for both models trained up to the Norcia earthquake. This information gain is simply the difference in the log-likelihood scores, where we subtract the score of ETAS from the neural model for both the magnitude and event-time term of the likelihood. The CIG is plotted per earthquake, but the evolution with time since the Norcia earthquake is also displayed. Figure 2.8a) shows the CIG for event time forecasts. Beyond the trend that the neural model improves over ETAS as we lower Mcut, the improvement varies over the testing catalog. For the thresholds that give the largest gain, (Mcut = 1.2, 1.4, 1.6, 1.8), the period of time with the greatest amount of gain, indicated by the steepest gradient of the curve, is found within the first 2 hours of the Norcia earthquake. This is followed by a reduced improvement up to 24 hours, beyond which it levels out and remains relatively linear.

Figure 2.8b) shows the CIG for the magnitude forecasts, confirming the loss in average

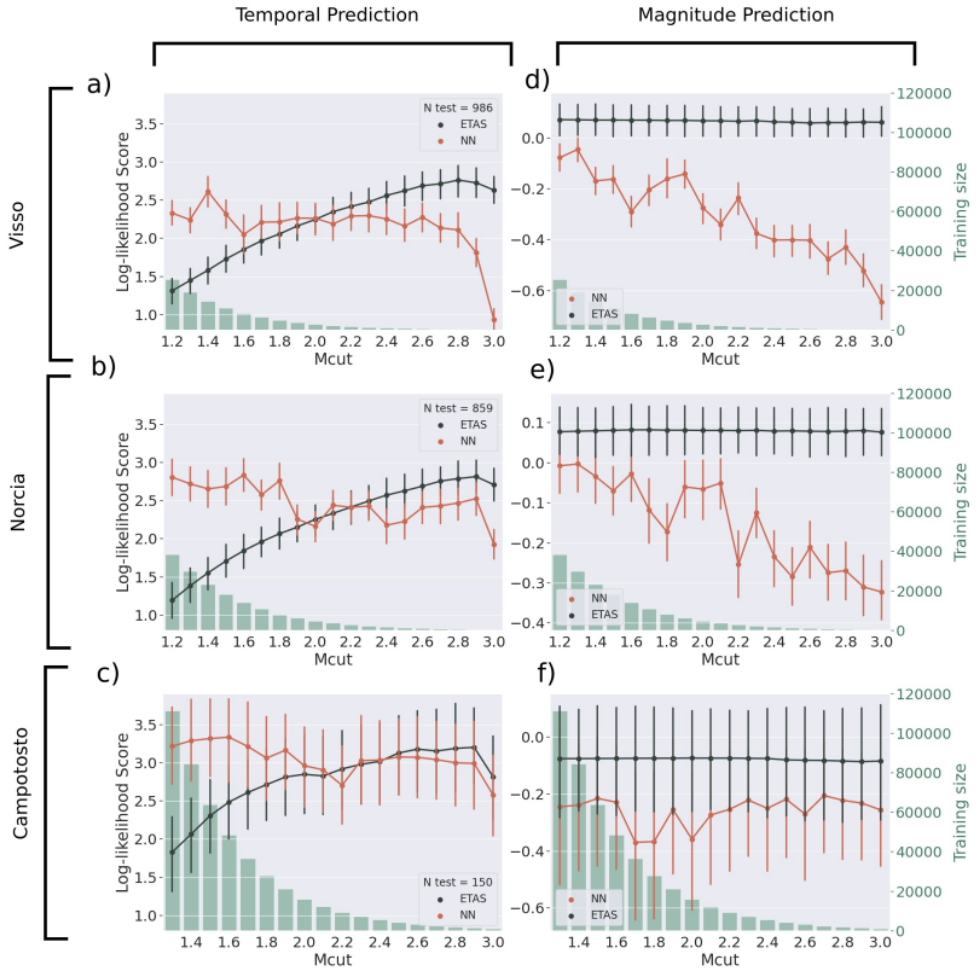


Figure 2.7: Results from tests on AVN catalog. 95 % confidence intervals for the log-likelihood score of each model for varying values of Mcut. The size of the training set is displayed in the green barplot as well as the size of the testing set in the legend. a)-c) depicts the temporal log-likelihood gain from Poisson. In a), both models are trained up to the Visso earthquake, in b) both models are trained up to the Norcia earthquake and in c) both are trained up to the Campotosto earthquakes. d) - f) depict the magnitude term of the log-likelihood for the same training-testing partitions.

performance of the neural model over ETAS. All thresholds decrease fairly steadily for nearly all of the testing period, apart from immediately following the Norcia earthquake. For the lower thresholds the period of time following Norcia sees an improvement over ETAS very briefly before declining.

Figure 2.8c) shows the IG of the neural model over ETAS but now as a function of the estimate of the completeness of the testing period. Both models are trained up to Norcia for Mcut = 1.2, 2.0, 2.8. A locally weighted scatterplot smoothing (lowess) regression [24] with 95 % confidence intervals estimates this relationship. For Mcut = 1.2, the difference between the two models is smallest for intermediate values of the incompleteness (around $M_c(t) = 2.0$), but for the most complete ($M_c(t) = 1.0$) and most incomplete ($M_c(t) = 3.0$) parts of the testing catalog, the neural model performs greatest compared to ETAS. At $M_c(t) = 2.0$ where the relative performance of ETAS is best, the confidence interval for the log-likelihood difference between the two models lies above zero, centered around 0.75. So, there is no value of completeness in the testing catalog where ETAS performs as well as the neural model. For Mcut = 2.0, the neural model outperforms ETAS during more incomplete periods of the testing period and for Mcut = 2.8, ETAS is consistently better across all values of completeness.

The forecasted magnitude distribution of each model shown in Figure 2.9 provides some insight into the cause of this immediate improvement over ETAS's GR forecast right after Norcia. Figure 2.9a) shows the magnitude distribution of each model at the time of the Norcia earthquake. The two distributions resemble each other relatively closely. At the time of the next Mw3+ event, Figure 2.9b), the neural model has shifted its density towards higher magnitude values anticipating a lack of observed smaller magnitude earthquakes due to catalog incompleteness.

2.5 Discussion

2.5.1 Approximating ETAS

The ability to approximate ETAS data using a neural point process is a benchmark goal. Demonstrating a baseline level of expressiveness is essential before any work on real data is done. Given other neural point process models regularly use univariate or discretely marked Hawkes data as a baseline [39, 157], this result provides an example of fit to continuous bivariate Hawkes data.

Specifically, the merit of this fit to ETAS data is that it uses a truncated history of events. Truncating the ETAS model to sum over only the last d events would dramatically change the evaluation of the intensity function between a significantly large earthquake d events ago or

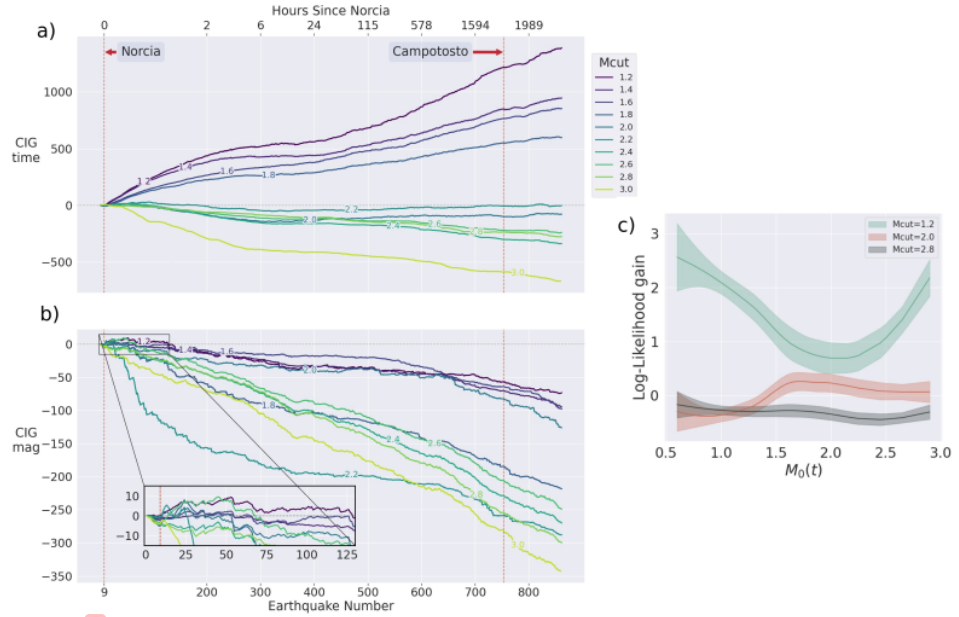


Figure 2.8: a) - b) The Cumulative Information Gain (CIG) of the neural model over ETAS for a range of values of Mcut. The models are trained up to the Norcia earthquake and the plot depicts the evolution of the CIG from the Norcia earthquake to the end of the catalog. The curve is plotted per event, however, the actual time since the Norcia earthquake is displayed on the top axis. a) displays the CIG for event-time forecasts, b) displays the CIG for magnitude forecasts. c) displays the information gain of the neural model over ETAS as a function of the completeness of the testing catalog - both models are trained up to Norcia for Mcut = 1.2, 2.0, 2.8.

$d+1$ events ago, thus instead the way in which to formulate the relationship between the intensity and these truncated events is learnt. The past d events are exhibiting behaviour based on the events prior and thus we do not directly specify the contribution from events further back in time, instead dependence on such events is learnt indirectly.

The reduction in the amount of history each forecast is dependent upon drastically improves the computational requirements for both learning and evaluation of the likelihood. To create forecasting models alongside the growing size of earthquake catalogs generated through machine learning based phase picking, we require models that scale well with the data. Shown here is that we can achieve the same predictive accuracy as ETAS (in a synthetic catalog) but with a smaller computational budget.

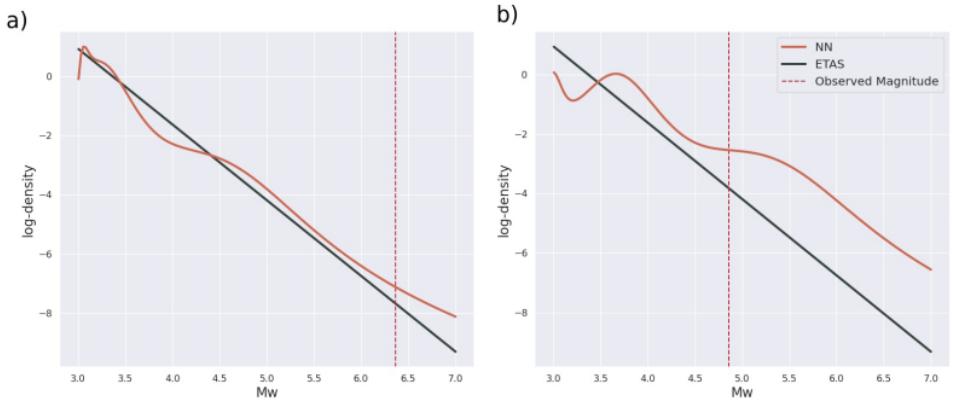


Figure 2.9: The forecasted magnitude distribution of each model, a) at the occurrence of the Norcia earthquake, and b) at the occurrence of the next Mw3+ earthquake following Norcia.

2.5.2 Embracing and Ignoring Data Incompleteness

At the previously explored thresholds of this sequence, [40, 118, 120, 125], the neural point process performs similarly to ETAS. The biggest deviations between the two models are found as the magnitude threshold is lowered into new unexplored regions revealed by this enhanced catalog. The deviations come from the fact that the neural model increases gradually in performance as the threshold is lowered whereas ETAS drastically decreases in performance. Below, we offer an interpretation for these results.

We argue that the largest gains made by the neural model are due to its ability to handle the incomplete data immediately following large earthquakes. There are two justifications for this: Given that the magnitude threshold of the input catalog is lowered into regions where there are periods when $M_c(t) >$ Mcut, we expect there to be biases in fitting ETAS [66, 183, 245]. The consequences of these biases on ETAS are reflected in the log-likelihood scores on the synthetic incomplete data figure 2.5b). In this synthetic catalog with short-term aftershock incompleteness, ETAS drastically reduces in performance in contrast to the neural model. Since a basic ETAS is formulated as completely observing all potentially triggering earthquakes it poorly captures incomplete sequences. The same shape of plot is found in the real data in Figure 2.7a)-c), where the performance of ETAS decreases as Mcut is lowered. In contrast, other studies have found decreasing the minimum triggering magnitude improves the performance of ETAS (e.g., [78, 224]). A consequence of the bias in the ETAS parameters is that the forecasting performance is only competitive with the neural model during intermediate values of incompleteness, Figure 2.8c). Even during complete periods in the testing catalog, since ETAS has been fit on incomplete data, it fails to forecast well.

The second justification comes through considering the process by which the observed data are generated, e.g. as described by [154]. The relationship between the underlying process $\lambda(t|H_t)$ and the observed process $v(t|H_t)$ that forms the catalog itself can be written as

$$v(t|H_t) = \lambda(t|H_t)\ r(t|M_c),$$

where $r(t|M_c)$ is the probability of detection at time t. For ETAS variants that deal with time-varying completeness [65, 66, 133, 154], this function has to be estimated alongside the parameters of ETAS. When fitting a temporal ETAS model to data with temporal incompleteness, bias in the fitted parameters comes from the modeling assumption that $v^*(t) = \lambda^*(t)$. Through the use of a flexible model such as this neural point process, rather than trying to learn both the underlying process and the detection rate, we instead directly learn the observed process. So in fact, in the construction of the model rather than equation (2.2), the observation process is approximated,

$$\phi(t - t_i|\mathbf{h}_i) = v^*(t),$$

where t_i is the time of the last event. As we lower Mcut into regions of temporal incompleteness, unlike ETAS, the performance of the neural model does not decrease, Figure 2.5b). This demonstrates the neural models' ability to fit to observed data as it is not biased by an increasing amount of missingness.

Learning to model the observation process requires the assumption that the process of the incompleteness is stationary for future forecasts, thus if there is new methodology in data collection, the model would have to be re-trained on this new data. This is similar to detection rate based methods, [65, 66, 133, 154], that would also need to update their detection function with new observational methodology.

To further test the effectiveness of the neural model, a comparison with other ETAS models that specifically deal with incompleteness is needed. But given that methods that deal with incompleteness only increase the computational requirements upon fitting a basic ETAS model, neural point processes could offer a more efficient way to deal with missing data. This is especially important when moving towards using enhanced catalogs such as the one used here. Temporal variations in completeness must be considered when using these catalogs and to be able to use these catalogs we must take more care with the computational efficiency of models.

Although data incompleteness is the most reasonable argument for the significant gains of the neural point process at low magnitude thresholds, we shouldn't limit ourselves to this description. We have reached this conclusion by extrapolating the forecasting results on synthetic data and through arguments about which statistical process is being approximated

by the neural model. However, we shouldn't rule out the possibility that the new low magnitude data in this enhanced catalog has contributed additional signal that is not explainable by ETAS. Even for the intermediate value of completeness that results in the best performance of ETAS compared with the neural model, there is still an average log-likelihood gain of around 0.75, Figure 2.8c). We can compare this with Figure A.3c in the appendix, where on the incomplete synthetic data we see a truncated curve of the same shape. In this synthetic experiment, there are periods of completeness where the performance between the two models is comparable, however on the real data, since there is no value of completeness that results in comparable performance, this would suggest that something additional is contributing to the gains. The question of whether there is additional signal in low magnitude events found in enhanced catalogs such as this one needs further attention beyond this study. We believe that further development in neural point processes will aid modellers in analysing this wealth of new data as neural models provide more flexible modelling alternatives and can cope with the scale of new enhanced catalogs.

2.5.3 Limitations

This study presents a flexible model that does not suffer from the same misspecification as ETAS due to short-term aftershock incompleteness. Although the size of the gains for these magnitude thresholds is large, we found, however, no significant overall improvement in forecasting ability over ETAS across magnitude thresholds. Comparing the value of Mcut that gives the greatest performance for each model finds that although the mean of the neural model is highest, the gain over ETAS is not significant. In this two dimensional time-magnitude domain, given the flexibility of this neural network and the data volume provided, there is insufficient signal in the data to learn anything significantly better than ETAS. This would suggest that the time and magnitude data from low magnitude events alone does not give us additional information in forecasting M3+ events. This motivates considering whether additional features can aid in the forecasting ability of neural point processes. Given that operationally we also require spatial forecasts, this is an obvious future extension to the model. It is natural to expect that including spatial covariates would improve forecasting performance [151, 211], however, it is not clear that considering them as an additional dimension to the input of an RNN would learn any spatial structure from the data. Neural point processes for spatio-temporal data do not utilise RNNs which are primarily sequence encoders and instead consider models based on Ordinary Differential Equations (ODEs) [11, 19]. We believe such models should outperform RNN-based models on spatio-temporal data.

By modelling the magnitudes by a completely unconstrained density that is also time-history dependent, we create lots of potential for over-fitting to the data [234]. This is exactly what is observed during the tests on synthetic data where we learn a 'noisy' Gutenberg-Richter law. It

is the likely source for the performance which is on average worse than ETAS on the AVN catalog. However, by letting this function be unconstrained, the model was able to make improvements over ETAS immediately following Norcia. This isn't too surprising since deviations from a stationary GR law have been observed, either through fluctuations of the b-value in space or time [61, 180] or short-term aftershock incompleteness [79, 95, 230].

A final limitation of the model presented here is its (in)ability to simulate events into the future. Where simulation of ETAS can be done due its equivalent branching process formulation, simulation from the neural model can only be leveraged through inverse transform sampling (Algorithm 1), where $\Lambda^{-1}(\cdot)$ is found through a numerical root finding method, or by thinning (Algorithm 2).

2.6 Conclusion

We present an initial investigation into the viability of neural point processes for the forecasting of short term seismicity. The neural point process is formulated in a similar way to the ETAS model, only with a much more flexible way of representing the intensity function. Now with much larger earthquake catalogs, data-driven point process models present us with an opportunity to investigate whether these new data may offer some deviation to the parameterization of ETAS as well as providing more computationally efficient models that are robust to missing data.

We extend the existing point process model of Omi et al. [157] so that it also models the magnitudes associated with the events contained in earthquake sequences. We also show how this model can be used to forecast earthquakes above some target threshold magnitude through decomposing the cumulative hazard function between target events. A notable feature of the presented model is that a forecast is only dependent on a fixed length vector representing the history of events, making the evaluation of the likelihood scale linearly with the sample size.

With an experiment on data simulated from the ETAS model we demonstrate this computational advantage by showing a stark improvement on the time to train the neural point process against the ETAS model, whilst still obtaining a similar likelihood score on test data. We find that defining a more flexible time-history dependent magnitude distribution leads to overfitting and consequentially the magnitude likelihood scores are worse than when using a stationary Gutenberg-Richter law.

Through artificially removing events from the synthetic catalog we create a dataset that mimics short-term aftershock incompleteness. We find that on this altered catalog, the

performance of ETAS now decreases as the magnitude threshold is lowered. In contrast, the neural model remains constant in performance, suggesting that it is more robust to the missing data found in typical earthquake catalogs.

On real data from the Amatrice-Visso-Norcia sequence the performance of both models vary with respect to the magnitude threshold of the input catalog. Both models perform similarly at previously explored thresholds (Mw3+), but when lowered into magnitude regions revealed by the new catalog, ETAS decreases in performance unlike the neural point process. We argue that these gains are due to the neural model's ability to handle the incomplete data found in this enhanced catalog. This experiment both motivates the need for considering temporal completeness when using enhanced catalogs and motivates further work into what the spatial covariates of this dataset might offer when combined with flexible point process models such as this one.

2.7 Open Research

The Amatrice-Visso-Norcia catalog produced by Tan et al. [206] is accessible at the Zenodo repository https://doi.org/10.5281/zenodo.4736089 [205]. The ETAS simulator used to generate the synthetic data was written for Mizrahi et al. [132] and Mizrahi et al. [133], and is available at `https://github.com/lmizrahi/etas`, [134]. Both synthetic and real datasets are found in the reproducibility package along with the models and the experimental design used in this study [200].

Chapter 3

SB-ETAS: using simulation based inference for scalable, likelihood-free inference for the ETAS model of earthquake occurrences

Declaration

The methodology, experimentation and writing of this chapter was undertaken by me, Samuel Stockman, with guidance from my two supervisors: Maximilian Werner and Daniel Lawson. This work benefited from external reviews by Robert Shcherbakov and an anonymous reviewer, with all changes made by me.

The following chapter was published in Statistics and Computing on August 29th 2024 [202]:
Stockman, Samuel, Daniel J. Lawson, and Maximilian J. Werner. SB-ETAS: using simulation based inference for scalable, likelihood-free inference for the ETAS model of earthquake occurrences. Statistics and Computing, 34(5), 174.
https://doi.org/10.1007/s11222-024-10486-6

This work appears in this thesis in near-identical format to the original publication. An additional section (3.3) describes how I extend the model of the previous chapter for simulation based inference and reports experimental results on a 3 parameter Hawkes process. The Supplementary Material has been included as Appendix B.

Abstract

The rapid growth of earthquake catalogs, driven by machine learning-based phase picking and denser seismic networks, calls for the application of a broader range of models to determine whether the new data enhances earthquake forecasting capabilities. Additionally, this growth demands that existing forecasting models efficiently scale to handle the increased data volume. Approximate inference methods such as `inlabru`, which is based on the Integrated nested Laplace approximation (INLA), offer improved computational efficiencies and the ability to perform inference on more complex point-process models compared to traditional MCMC approaches. We present SB-ETAS: a simulation based inference procedure for the Epidemic-Type Aftershock Sequence (ETAS) model. This approximate Bayesian method uses Sequential Neural Posterior Estimation (SNPE) to learn posterior distributions from simulations, rather than typical MCMC sampling using the likelihood. On synthetic earthquake catalogs, SB-ETAS provides better coverage of ETAS posterior distributions compared with `inlabru`. Furthermore, we demonstrate that using a simulation based procedure for inference improves the scalability from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$. This makes it feasible to fit to very large earthquake catalogs, such as one for Southern California dating back to 1981. SB-ETAS can find Bayesian estimates of ETAS parameters for this catalog in less than 10 hours on a standard laptop, a task that would have taken over 2 weeks using MCMC. Beyond the standard ETAS model, this simulation based framework allows earthquake modellers to define and infer parameters for much more complex models by removing the need to define a likelihood function.

3.1 Introduction

In recent years, an accelerated growth in the number of seismic sensors and machine learning algorithms for detecting the arrival times of earthquake phases [e.g. 242], has meant that the size earthquake catalogs have grown by several orders of magnitude [99]. In California, a deployment of a dense network of seismic sensors over the last century combined with an active tectonic regime has resulted in a comprehensive dataset of earthquakes in the region [88]. Furthermore, in more specific areas of California, through machine learning based seismic phase picking and template matching, enhanced earthquake catalogs have been created which contain many small previously undetected earthquakes [175, 225]. It is fair to assume that these datasets will only continue to grow in the future as past continuous data is reprocessed and future earthquakes are recorded. Determining whether increased data size leads to improved earthquake forecasts is a crucial question for the seismological community. The growing volume of data requires an expansion of modeling capabilities within the field. This not only necessitates that existing models can scale with the increasing datasets but also calls for a broader range of models to be fit to the data.

In this work we propose using Simulation Based Inference (SBI) to address this modeling expansion. We present SB-ETAS: a simulation based estimation procedure for the Epidemic Type Aftershock Sequence (ETAS) model, the most widely used earthquake model among seismologists. SBI is a family of approximate procedures which infer posterior distributions for parameters using simulations in place of the likelihood [9, 26]. By specifying a model through simulation rather than the likelihood, SBI broadens the scope of available models to encompass greater complexity. This study also demonstrates that for the ETAS model, SBI improves the scalability from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$.

While there is extensive literature on SBI, its application to Hawkes process models [76], of which ETAS is a member, is limited. This work builds upon earlier studies by Ertekin et al. [42] and Deutsch and Ross [33], which applied SBI to 1-dimensional Hawkes processes with exponential kernels. We expand upon their choice of summary statistics to fit the more complex ETAS model, which includes a magnitude (mark) domain and power law kernels. We add that since both simulation and summary statistic computation can be performed with time complexity $\mathcal{O}(n \log n)$, then SBI offers the additional benefit of scalability. Additionally, we enhance inference performance by using Sequential Neural Posterior Estimation (SNPE). SNPE trains a neural density estimator to approximate the posterior distribution from pairs of simulations and parameters. Section 3.2 provides an overview of SNPE and other SBI methods.

In SBI, models are defined through a simulator, eliminating the need to specify a likelihood function for inference. This approach has been adopted in other scientific fields where the likelihood is intractable, such as when it involves integrating over numerous unobserved latent variables. Seismology already encounters such intractable likelihoods. For instance, models that account for triggering from undetected earthquakes [196] and those that incorporate geological features [46] present estimation challenges. By linking earthquake modeling with SBI, this study introduces a framework for fitting these complex models while also providing an immediate scalability benefit for simpler models.

The remainder of this chapter is structured as follows: In section 3.2 we give an overview of SBI, following which we show how the model introduced in the last chapter (Chapter 2) can be extended to perform SBI in section 3.3. We then describe the details of SB-ETAS in section 3.4. We present empirical results based on synthetic earthquake catalogs in section 3.5 and observational earthquake data from Southern California in section 3.6, before finishing with a discussion in section 3.7.

3.2 Simulation Based Inference

A family of Bayesian inference methods have evolved from application settings in science, economics or engineering where stochastic models are used to describe complex phenomena. In this setting, the model may simulate data from a given set of input parameters, however, the likelihood of observing data given parameters is intractable. The task in this setting is to approximate the posterior $p(\theta|\mathbf{Y}_{\text{obs}}) \propto p(\mathbf{Y}_{\text{obs}}|\theta)p(\theta)$, with the restriction that we cannot evaluate $p(\mathbf{Y}|\theta)$ but we have access to the likelihood implicitly through samples $\mathbf{Y}_r \sim p(\mathbf{Y}|\theta_r)$ from a simulator, for $r = 1, \ldots, R$ and where $\theta_r \sim p(\theta)$. This approach is commonly referred to as Simulation Based Inference (SBI) or likelihood-free inference.

Until recently, the predominant approach for SBI was Approximate Bayesian Computation [9]. In its simplest form, parameters are chosen from the prior $\theta_r \sim p(\theta),\ r = 1, \ldots, R$, the simulator then generates samples $\mathbf{Y}_r \sim p(\mathbf{Y}|\theta_r),\ r = 1, \ldots, R$, and each sample is kept if it is within some tolerance ϵ of the observed data, i.e. $d(\mathbf{Y}_r, \mathbf{Y}_{\text{obs}}) < \epsilon$ for a given distance function $d(\cdot, \cdot)$.

This approach, although exact when $\epsilon \to 0$, is inefficient with the use of simulations. Sufficiently small ϵ requires simulating an impractical number of times, and this issue scales poorly with the dimension of $\mathbf{Y}$. In light of this an MCMC approach to ABC makes proposals for new simulator parameters $\theta_r \sim q(\cdot|\theta_{r-1})$ using a Metropolis-Hastings kernel [9, 121]. This leads to a far higher acceptance of proposed simulator parameters but still scales poorly with the dimension of $\mathbf{Y}$.

In order to cope with high dimensional simulator outputs $\mathbf{Y} \in \mathbb{R}^n$, summary statistics $S(\mathbf{Y}) \in \mathbb{R}^d$ are chosen to reduce the dimension of the sample whilst still retaining as much information as possible. These are often chosen from domain knowledge or can be learnt as part of the inference procedure [166]. Summary statistics $S(\mathbf{Y})$ are then used in place of $\mathbf{Y}$ in any of the described methods for SBI.

3.2.1 Neural Density Estimation

Recently, SBI has seen more rapid development as a result of neural network based density estimators [116, 160, 161], which seek to approximate the density $p(y)$ given samples of points $y \sim p(y)$. A popular method for neural density estimation is normalising flows [170], in which a neural network parameterizes an invertible transformation $y = g_\phi(u)$, of a variable u from a simple base distribution $p(u)$ into the target distribution of interest. In practice, the transformation is typically composed of a stack of invertible transformations, which allows it to learn the complex target density. The parameters of the transformation are trained through

maximising the likelihood of observing $p_g(y)$, which is given by the change of variables formula. Since y is expressed as a transformation of a simple distribution $u \sim p(u)$, samples from the learnt distribution $p_g(y)$ can be generated by sampling from $p(u)$ and passing the samples through the transformation. Neural density estimators may also be generalised to learn conditional densities $p(\mathbf{Y}|\mathbf{z})$ by conditioning the transformation g_ϕ on the variable z [161].

In the task of SBI, a neural density estimator can be trained on pairs of samples $\theta_r \sim p(\theta), \mathbf{Y}_r \sim p(\mathbf{Y}|\theta_r)$ to approximate either the likelihood $p(\mathbf{Y}_{\text{obs}}|\theta)$ or the posterior density $p(\theta|\mathbf{Y}_{\text{obs}})$, from which posterior samples can be obtained. If the posterior density is estimated, in a procedure known as Neural Posterior Estimation (NPE) [116], then samples can be drawn from the normalising flow. If the likelihood is estimated, known as Neural Likelihood Estimation (NLE) [162], then the approximate likelihood can be used in place of the true likelihood in a MCMC sampling algorithm to obtain posterior samples. Neural density estimation techniques consistently outperform ABC-based methods in benchmarking experiments, since they can efficiently interpolate between different simulations [Figure 3.1, 117]. Other neural network methods exist for SBI such as ratio estimation [90] or score matching [51, 186], however, we direct the reader to Cranmer et al. [26] for a more comprehensive review of modern SBI.

3.3 SBI using Neural Point Processes

In the previous two chapters (1.4.1, 2.3.1), we saw how through a sequence encoding of the most recent history $\mathbf{h}_i = \sigma(t_i, \dots, t_{i-d})$, and by modelling the cumulative hazard function,

$$\Phi(\tau|\mathbf{h}_i) = \int_{t_i}^{\tau} \lambda(s|\mathbf{h}_i)ds, \tag{3.1}$$

one could construct the density of the next event,

$$\log p(t_{i+1}|t_i, \dots, t_{i-d}) = \log \frac{\partial}{\partial \tau}\Phi(\tau_i = t_{i+1} - t_i|\mathbf{h}_i) - \Phi(\tau_i|\mathbf{h}_i), \tag{3.2}$$

where the log-likelihood of all events can be expressed as the sum of these conditional densities,

$$\log p(\{t_i\}_{i=1}^n) = \sum_i \left[\log \frac{\partial}{\partial \tau}\Phi(\tau_i|\mathbf{h}_i) - \Phi(\tau_i|\mathbf{h}_i)\right]. \tag{3.3}$$

This density estimation task is related to normalising flows through the Random Time Change Theorem (1.1), where the "simple" distribution $u \sim p(u)$ is a an exponential distribution (the inter-event distribution of a Poisson process) [191].

By extending the modeling of next-event densities $p(t_i|\mathcal{H}_{t_i})$, to be conditional on ETAS model parameters $p(t_i|\mathcal{H}_{t_i}, \theta)$ extends this approach to be usable for SBI for the ETAS model. The

SBI approach then becomes Neural Likelihood Estimation (NLE),

$$\log p(\mathbf{Y} = \{t_i\}_{i=1}^n | \theta) = \sum_{i=1}^{n} \log p(t_i | \mathcal{H}_{t_i}, \theta) \tag{3.4}$$

$$= \sum_{i=1}^{n} \left[\log \frac{\partial}{\partial \tau} \Phi(\tau_i | \mathbf{h}_i, \theta) - \Phi(\tau_i | \mathbf{h}_i, \theta) \right], \tag{3.5}$$

where the cumulative hazard function $\Phi(\tau_i|\mathbf{h}_i, \theta)$ now depends on θ. Following [157, 201], Φ is modeled by a feed-forward neural network with enforced positive weights to ensure positivity and monotonicity (in τ).

The neural likelihood estimator is trained by maximising the total log-likelihood of R parameter-simulation pairs,

$$\sum_{r=1}^{R} \log p(\mathbf{Y}_r | \theta_r). \tag{3.6}$$

samples from the approximate posterior can then be obtained through MCMC slice sampling [145, 162].

3.3.1 Results

Before extending to the ETAS model, initial tests perform Bayesian inference for the much simpler 3-parameter univariate Hawkes process,

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{t_i < t} k e^{\upsilon(t - t_i)} \tag{3.7}$$

For this model, since we have access to the likelihood function, we can compare NLE to posterior samples obtained through MCMC. Further comparison is also done to ABC-Hawkes introduced by Deutsch and Ross [33], an ABC-MCMC method which uses summary statistics $S(\mathbf{Y})$ which we will introduce later. Finally, we develop upon ABC-Hawkes using their same summary statistics $S(\mathbf{Y})$, but instead with Sequential Neural Posterior Estimation (SNPE) [116, 160].

Figure 3.1 shows posterior samples for the simple Hawkes process from all four approaches. Unfortunately the posterior samples using the likelihood approximation do not appear to fit the MCMC posterior as well as the other two methods that use summary statistics. NLE appears biased, particularly for the parameter υ. ABC-MCMC gives much wider posterior estimates than SNPE, despite using many more simulations and the same summary statistics $S(\mathbf{Y})$.

3.3.2 Discussion

What differentiates Hawkes process models from other simulator models used in Neural-SBI is that the output of the simulator $\mathbf{Y} = t_1, \ldots, t_n$ itself has random dimension. For a specified

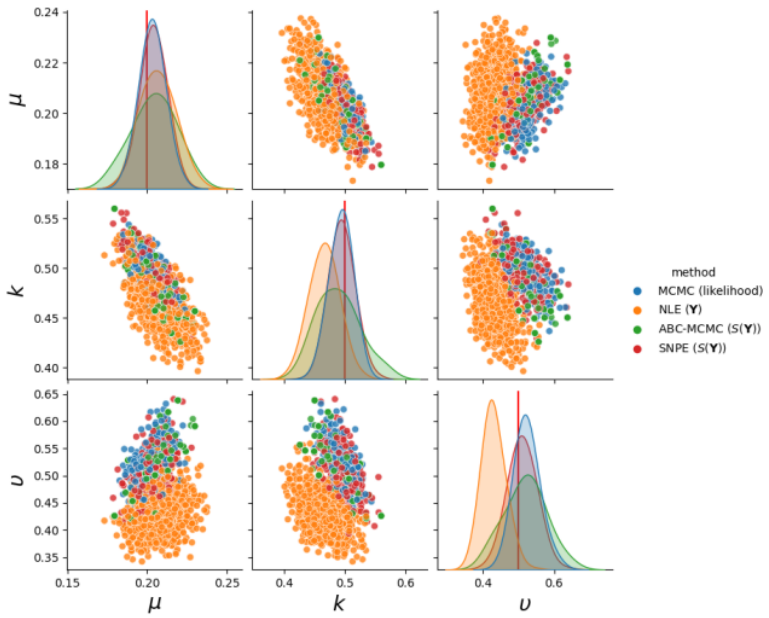


Figure 3.1: Posterior densities for a univariate Hawkes process with exponential kernel. The 'observed' data contains 4806 events and was simulated from parameters indicated in red on the diagonal plots. In blue are posterior samples found using MCMC sampling with likelihood function. In orange are posterior samples from the neural likelihood approximation (NLE) using 100,000 simulations. In green are posterior samples found using ABC-MCMC using 300,000 simulations. In blue are posterior samples from SNPE using the same summary statistics as ABC-MCMC but only 10,000 simulations. A Uniform($[0.05, 0, 0], [0.85, 0.9, 3]$) prior was used for all three methods.

time interval over which to simulate earthquakes $[0, T]$, one particular parameter θ_1 will generate different numbers of events if simulated repeatedly. This is problematic for off-the-shelf neural density estimators since even though they are successful over high dimensional data, they require a fixed dimensional input. For neural likelihood estimation, this was achieved through a truncated history encoding, $d = 40$. Testing performance over many values of d revealed no obvious improvement as it was increased (Figure B.1), likely due to vanishing gradients [85]. In the last chapter, Figure 2.5 demonstrates that next-event density estimation was possible using this particular architecture ($d = 20$ in that case), however we speculate that a longer event history is needed in order to successfully condition on model parameters θ. In light of this we proceed by pursuing a summary statistic approach to SBI, rather than a full likelihood approximation.

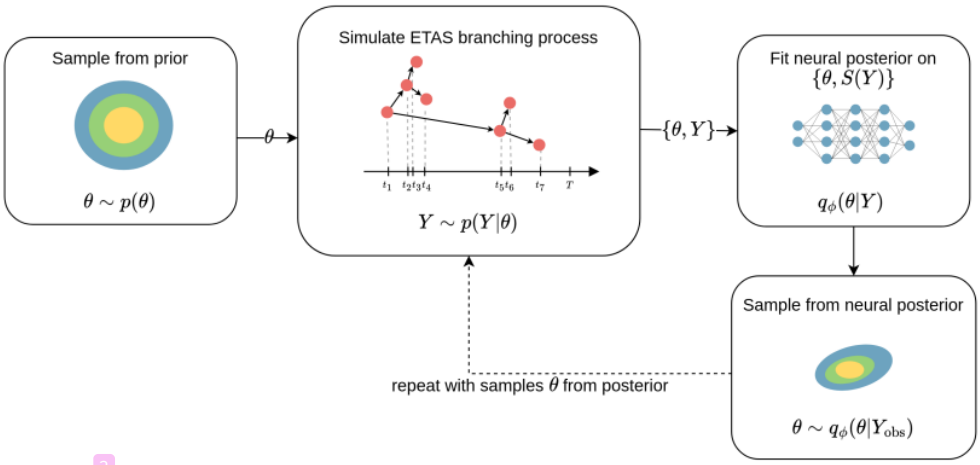


Figure 3.2: An outline of the SB-ETAS inference procedure. Samples from the prior distribution are used to simulate many ETAS sequences. A neural density estimator is then trained on the parameters and simulator outputs to approximate the posterior distribution. Samples from the posterior given the observed earthquake sequence can then be used to improve the estimate over rounds or are returned as the final posterior samples.

3.4 SB-ETAS

We now present SB-ETAS, our simulation based inference method for the ETAS model. The method avoids computing the likelihood function and instead leverages fast simulation from the ETAS branching process. The inference method uses Sequential Neural Posterior Estimation (SNPE) [116, 160], a modified version of NPE which performs inference over rounds. Each round, an estimate of the posterior proposes new samples for the simulator, a neural density estimator is trained on those samples and the estimated posterior is updated (Figure 3.2, Algorithm 4). SNPE was chosen over other methods of Neural-SBI, as it avoids the need to perform MCMC sampling, a slow procedure. Instead, sampling from the posterior is fast since the approximate posterior is a normalising flow.

3.4.1 Summary Statistics

Works to perform ABC on the univariate Hawkes process with exponential decay kernel (Equation 3.7) have found summary statistics that perform well in that setting. Ertekin et al. [42] use the histogram of inter-event times as summary statistics as well as the number of events. Deutsch and Ross [33] extend these summary statistics by adding Ripley's K statistic [173], which is a popular choice of summary statistic in spatial point processes [137]. Figure 3.1 shows the performance of the ABC-MCMC method developed by Deutsch and Ross [33], using

Algorithm 4 SB-ETAS

Input: observed data $\mathbf{Y}_{\text{obs}}$, summary statistic $S(\mathbf{Y})$, estimator $q_\phi(\theta|\mathbf{Y})$, prior $p(\theta)$, number of rounds K, simulations per round L.

$q_\phi^0(\theta|\mathbf{Y}_{\text{obs}}) = p(\theta)$ and $\mathcal{D} = \{\}$
for $k = 1 : K$ **do**
  for $l = 1 : L$ **do**
    sample $\theta_l \sim q_\phi^{k-1}(\theta|\mathbf{Y}_{\text{obs}})$
    simulate $\mathbf{Y}_l \sim p(\mathbf{Y}|\theta_l)$ using Algorithm 3
    add $(\theta_l, S(\mathbf{Y}_l))$ to $\mathcal{D}$
  end for
  train $q_\phi^k(\theta|\mathbf{Y}_{\text{obs}})$ on $\mathcal{D}$
end for

Output: approximate posterior $\hat{p}(\theta|\mathbf{Y}_{\text{obs}}) = \hat{q}_\phi^K(\theta|\mathbf{Y}_{\text{obs}})$

the aforementioned summary statistics. Using SNPE on the same summary statistics yields a more confident estimation of the "true" posterior which is found through MCMC sampling using the likelihood and requires far fewer simulations (10,000 versus 300,000). The ETAS model is more complex than a univariate Hawkes process since it is both marked (i.e. it contains earthquake magnitudes) and contains a power law decay kernel which decays much more slowly than exponential, making it harder to estimate [8]. For SB-ETAS we borrow similar summary statistics to [42], namely $S_1(\mathbf{Y}) = \log(\#\text{ events})$, $S_2, \ldots, S_4(\mathbf{Y}) =$20th, 50th and 90th quantiles of the inter-event time histogram. Similar to Deutsch and Ross [33], we use another statistic $S_5(\mathbf{Y})$, which is the ratio of the mean and median of the inter-event time histogram.

3.4.1.1 Ripley's K Statistic

For the remaining summary statistics, we develop upon the introduction of Ripley's K statistic by Deutsch and Ross [33]. For a univariate point process $\mathbf{Y} = (t_1, \ldots, t_n)$, Ripley's K statistic is [37],

$$K(\mathbf{Y}, w) = \frac{1}{\lambda}\mathbb{E}(\#\text{ of events within } w \text{ of a random event}). \tag{3.8}$$

Here, λ is the unconditional rate of events in the time window $[0, T]$. An estimator for the K-statistic is derived by Diggle [36],

$$\hat{K}(\mathbf{Y}, w) = \frac{T}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{I}(0 < t_j - t_i \leq w). \tag{3.9}$$

Despite containing a double-sum, computation of this estimator has complexity $\mathcal{O}(n)$ since $\{t_i\}_{i=1}^n$ is an ordered sequence, i.e. $(t_3 - t_1 < w) \Rightarrow (t_3 - t_2 < w)$. Calculation of Ripley

K-statistic therefore satisfies the complexity requirement of our procedure if the number of windows w for which we evaluate $\hat{K}(\mathbf{Y}, w)$ is less than $\log n$. In fact, our results suggest that less than 20 are required.

The use of Ripley's K-statistic for non-marked Hawkes data is motivated by Bacry and Muzy [7], who show that second-order properties fully characterise a Hawkes process and can be used to estimate a non-parametric triggering kernel. Bacry et al. [8] go on to give a recommendation for a binning strategy to estimate slow decay kernels such as a power law, using a combination of linear and log-scaling. It therefore seems reasonable to define $S_6(\mathbf{Y}) \dots S_{23}(\mathbf{Y}) = \hat{K}(\mathbf{Y}, w)$, where w scales logarithmically between $[0, 1]$ and linearly above 1.

We modify Ripley's K-statistic to account for the particular interaction between marks and points in the ETAS model. Namely, the magnitude of an earthquake directly affects the clustering that occurs following it, expressed in the productivity relationship (1.15). In light of this, we define a magnitude thresholded Ripley K-statistic,

$$K_T(\mathbf{Y}, w, M_T) = \frac{1}{\lambda_T}\mathbb{E}(\#\text{ events within } w \text{ of an event } m_i \geq M_T), \tag{3.10}$$

where λ_T is the unconditional rate of events above M_T. One can see that $K_T(\mathbf{Y}, w, M_c) = K(\mathbf{Y}, w)$. We estimate K_T with

$$\hat{K}_T(\mathbf{Y}, w, M_T) = \frac{T}{\nu^2} \sum_{i:m_i \geq M_T} \sum_{j \neq i} \mathbb{I}(0 < t_i - t_j \leq w), \tag{3.11}$$

where ν is the number of events above magnitude threshold M_T. For general M_T, we lose the $\mathcal{O}(n)$ complexity that the previous statistic has, instead it is $\mathcal{O}(\nu n)$. However, if the threshold is chosen to be large enough, evaluation of this estimator is fast. In our experiments, M_T is chosen to be $(4.5, 5, 5.5, 6)$, with $w = (0.2, 0.5, 1, 3)$. This defines the remaining statistics $S_{24}(\mathbf{Y}), \dots, S_{39}(\mathbf{Y})$.

3.5 Experiments and Results

To evaluate the performance of SB-ETAS, we conduct inference experiments on a series of synthetic ETAS catalogs. On each simulated catalog we seek to obtain 5000 samples from the posterior distribution of ETAS parameters. The latent variable MCMC inference procedure, `bayesianETAS`, will be used as a reference model in our experiments since it uses the ETAS likelihood without making any approximations. We compare samples from this exact method with samples from approximate methods, `inlabru` and SB-ETAS.

We begin with an experiment to test the scalability of all 3 methods. Following that we evaluate the performance of SB-ETAS on parameter sets estimated from real earthquake catalogs.

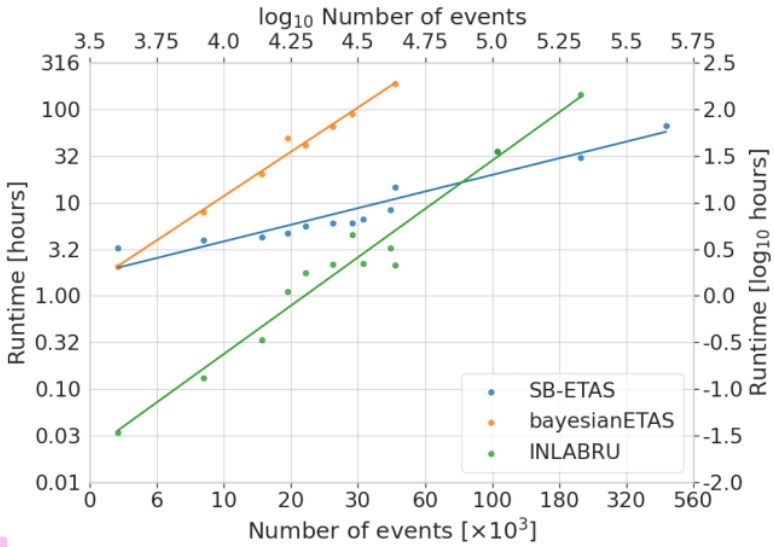


Figure 3.3: The runtime for parameter inference versus the catalog size for SB-ETAS, `inlabru` and `bayesianETAS`. Separate ETAS catalogs were generated with the same intensity function parameters but for varying size time-windows. The runtime in hours and the number of events are plotted in log-log space.

3.5.1 Scalability

Multiple catalogs are simulated from a fixed set of ETAS parameters, $(\mu, k, \alpha, c, p) = (0.2, 0.2, 1.5, 0.5, 2)$ with magnitude of completeness $M_c = 3$ and Gutenberg-Richter distribution parameter $\beta = 2.4$. Each new catalog is simulated in a time window $[0, T]$, where $T \in (10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 250, 500, 1000) \times 10^3$.

Figure 3.3 shows the runtime of each inference method as a function of the number of events in each catalog. Each method was run on a high-performance computing node with eight 2.4 GHz Intel E5-2680 v4 (Broadwell) CPUs, which is equivalent to what is commonly available on a standard laptop. On the catalogs with up to 100,000 events, `inlabru` is the fastest inference method, around ten times quicker on a catalog of 20,000 events. However, the superior scaling of SB-ETAS allows it to be run on the catalog of $\sim 500,000$ events, which was unfeasible for `inlabru` given the same computational resources i.e. it exceeded a two week time limit. The gradient of 2 for both `bayesianETAS` and `inlabru` in log-log space confirm the $\mathcal{O}(n^2)$ time complexity of both inference methods. SB-ETAS, on the other hand, has gradient $\frac{2}{3}$ which suggests that the theoretical $\mathcal{O}(n \log n)$ time complexity is a conservative upper-bound.

The prior distributions for each implementation are not identical since each has its own

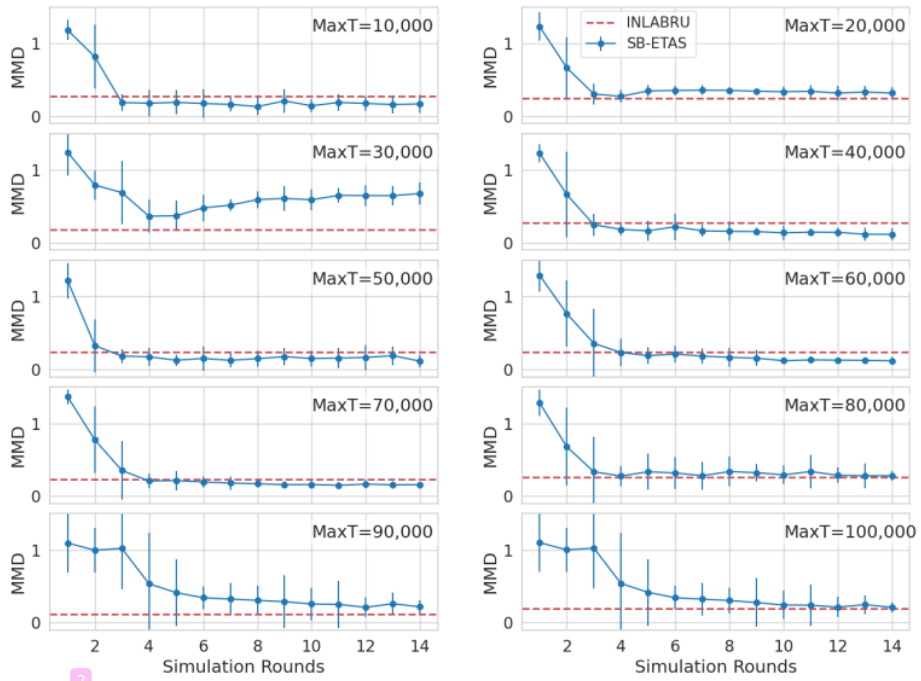


Figure 3.4: Maximum Mean Discrepancy for samples from each round of simulations in SB-ETAS. Each plot corresponds to a different simulated ETAS catalog simulated with identical model parameters but over a different length time-window (MaxT). In red is the performance metric evaluated for samples from `inlabru`. 95% confidence intervals are plotted for SB-ETAS across 10 different initial seeds.

requirements. Priors are chosen to replicate the fixed implementation in the `bayesianETAS` package,

$$\mu \sim \text{Gamma}(0.1, 0.1) \tag{3.12}$$

$$K, \alpha, c \sim \text{Unif}(0, 10) \tag{3.13}$$

$$p \sim \text{Unif}(1, 10). \tag{3.14}$$

The implementation of `inlabru` uses a transformation $K_b = \frac{K(p-1)}{c}$, with prior $K_b \sim \text{Log-Normal}(-1, 2.03)$ chosen by matching 1% and 99% quantiles with the `bayesianETAS` prior for K. SB-ETAS uses a $\mu \sim \text{Unif}(0.05, 0.3)$ prior in place of the gamma prior as well as enforcing a sub-critical parameter region $K\beta < \beta - \alpha$ [243]. Both the uniform prior and the restriction on K and α stop unnecessarily long or infinite simulations.

Once samples are obtained from SB-ETAS and `inlabru`, we measure their (dis)similarity with samples from the exact method `bayesianETAS` using the Maximum Mean Discrepancy (MMD)

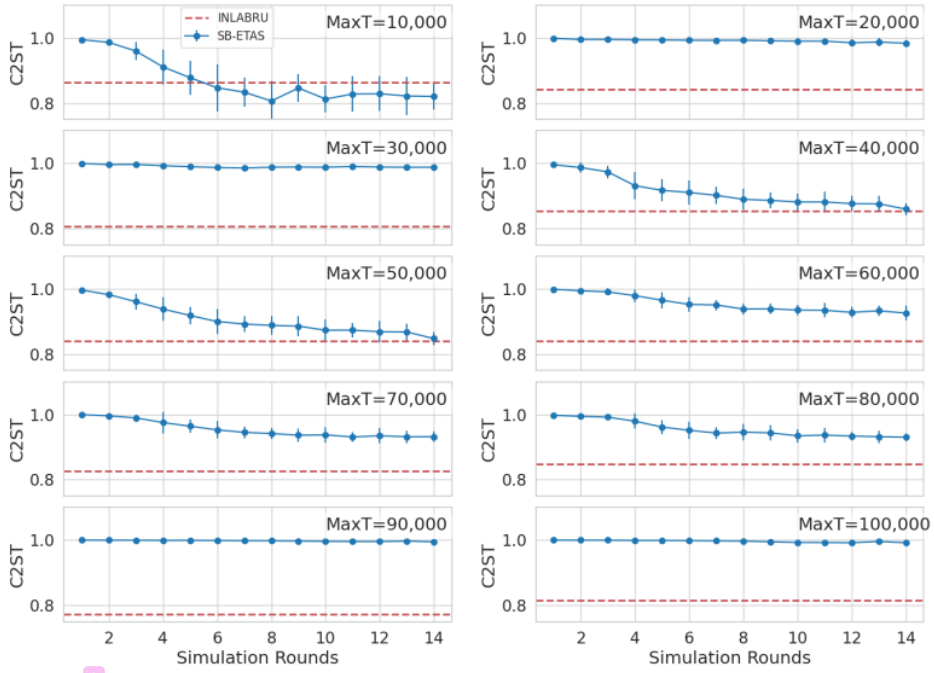


Figure 3.5: Classifier Two-Sample Test scores for samples from each round of simulations in SB-ETAS. Each plot corresponds to a different simulated ETAS catalog simulated with identical model parameters but over a different length time-window (MaxT). In red is the performance metric evaluated for samples from `inlabru`. 95% confidence intervals are plotted for SB-ETAS across 10 different initial seeds.

[59] and the Classifier Two-Sample Test (C2ST) [104, 115]. Figures 3.4 and 3.5 show the values of these performance metrics for samples from each of 15 rounds of simulations in SB-ETAS compared with the performance of `inlabru`. Since SB-ETAS involves random sampling in the procedure, we repeat it across 10 different seeds and plot a 95% confidence intervals. In general across the 10 synthetic catalogs, SB-ETAS and `inlabru` are comparable in terms of MMD (Figure 3.4) and `inlabru` performs best in terms of C2ST (Figure 3.5). Figure 3.6 shows samples from the $T = 60,000$. Samples from `inlabru` are overconfident with respect to the `bayesianETAS` samples, whereas SB-ETAS samples are more conservative. This phenomenon is shared across the samples from all the simulated catalogs and we speculate that it accounts for the difference between the two metrics.

A common measure for the appropriateness of a prediction's uncertainty is the coverage property [165, 233]. The coverage of an approximate posterior assesses the quality of its

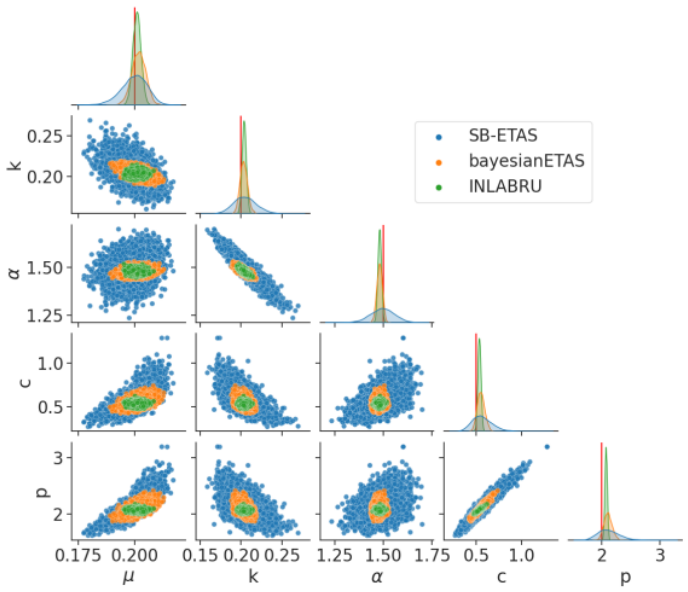


Figure 3.6: Samples from the posterior distribution of ETAS parameters for the simulated catalog with $T = 60,000$, for `bayesianETAS`, `inlabru` and SB-ETAS. The data generating parameters are marked in red in the diagonal plots.

credible regions $\hat{C}_{\mathbf{Y}_{obs}}$ which satisfy,

$$\gamma = \mathbb{E}_{\hat{p}(\theta|\mathbf{Y}_{obs})}\left(\mathbb{I}\{\hat{\theta} \in \hat{C}_{\mathbf{Y}_{obs}}\}\right) \tag{3.15}$$

An approximate posterior has perfect coverage if its operational coverage,

$$b(\mathbf{Y}_{obs}) = \mathbb{E}_{p(\theta|\mathbf{Y}_{obs})}\left(\mathbb{I}\{\hat{\theta} \in \hat{C}_{\mathbf{Y}_{obs}}\}\right) \tag{3.16}$$

is equal to the credibility level γ. The approximation is conservative if it has operational coverage $b(\mathbf{Y}_{obs}) > \gamma$ and is overconfident if $b(\mathbf{Y}_{obs}) < \gamma$ [82]. Expectations in equations (3.15)-(3.16) cannot be computed exactly and so are replaced with Monte Carlo averages, resulting in empirical coverage $c(\mathbf{Y}_{obs})$. Figure 3.7 shows the empirical coverage for both SB-ETAS, averaged across the 10 initial seeds, along with `inlabru` on the 10 synthetic catalogs. `inlabru` consistently gives overconfident approximations, as the empirical coverage lies well below the credibility level. SB-ETAS has empirical coverage that indicates conservative estimates, but that is generally closer to the credibility level.

3.5.2 Synthetic Catalogs

We now perform further tests to evaluate the performance of SB-ETAS on parameter sets estimated from real earthquake catalogs (Table 3.1). We consider MLE estimates of ETAS for

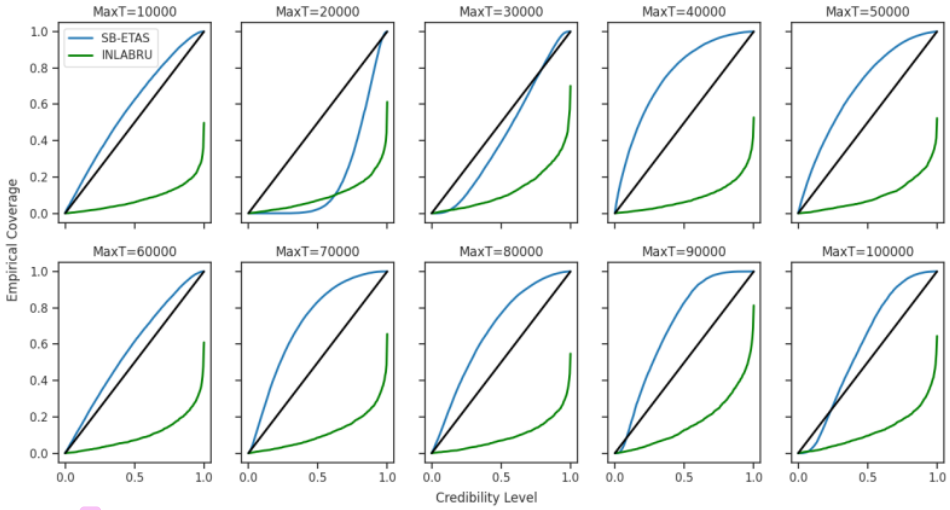


Figure 3.7: Empirical estimates of the coverage of both SB-ETAS and `inlabru`. Coverage below the black line $y = x$ indicates an overconfident approximation, whereas coverage below $y = x$ indicates a conservative approximation.

the Amatrice earthquake sequence, taken from Stockman et al. [201], for both Landers and Ridgecrest earthquakes, taken from Hainzl [67] and finally for the Kumamoto earthquake, taken from Zhuang et al. [245]. From each of these parameter sets, we simulate an earthquake catalog of around $6,000$ events and compare posterior samples using `bayesianETAS` with both SB-ETAS and `inlabru`.

Sequence	μ	K	α	c	p	β	M_0
Amatrice	0.084	0.422	1.34	0.00211	1.108	2.50	3.0
Landers	0.20	0.0674	0.36	0.06458	1.31	2.14	2.0
Ridgecrest	0.10	0.2891	0.66	0.02638	1.24	1.82	2.0
Kumamoto	0.073	0.0932	2.021	0.0091	1.157	2.30	3.5

Table 3.1: Parameter values used to generate the synthetic earthquake catalogs. Amatrice parameters were taken from [201], Landers and Ridgecrest parameters were taken from [67] and Kumamoto parameters were taken from [245]. The parameter K has been transformed for Landers, Ridgecrest and Kumamoto to account for the unnormalised Omori-Utsu law.

Figure 3.8 displays the MMD and C2ST scores for samples from each of 15 rounds of simulations in SB-ETAS compared with the performance of `inlabru`. SB-ETAS outperforms `inlabru` on the synthetic Amatrice, Landers and Ridgecrest catalogs across both metrics. This superior performance is attributed to the posterior distributions from SB-ETAS exhibiting less

bias and providing better coverage of the "ground truth" MCMC posteriors (Figures B.2-B.7). While `inlabru` provides the closest approximation for the synthetic Kumamoto catalog (Figure B.5), its posteriors are generally overconfident, leading to a lack of coverage for the MCMC posteriors whenever there is bias. Furthermore, the posterior distribution for the synthetic Landers catalog exhibits weak identifiability between parameters (c, p). While SB-ETAS expresses this in the posterior (Figure B.6), `inlabru` is unable to (Figure B.7).

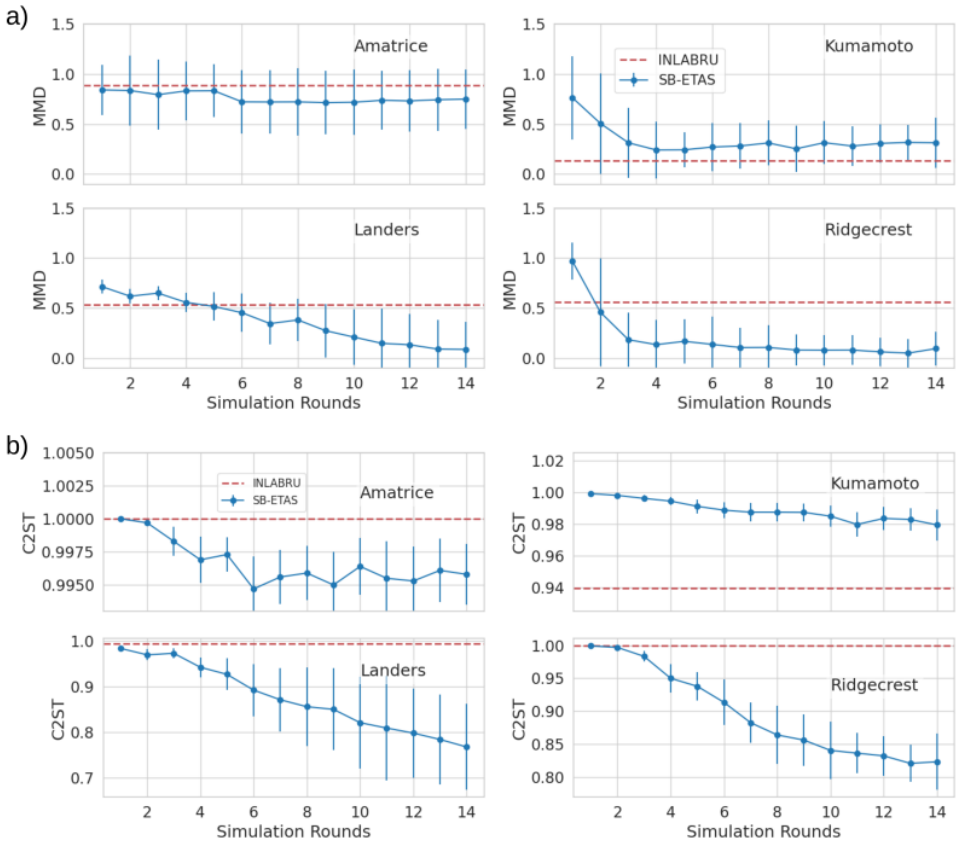


Figure 3.8: **a)** Maximum Mean Discrepancy (MMD) and **b)** Classifier Two-Sample Test (C2ST) scores for samples from each round of simulations in SB-ETAS. Each plot corresponds to a different synthetic ETAS catalog simulated using MLE parameters taken from the Amatrice, Kumamoto, Landers and Ridgecrest earthquake sequences. In red is the performance metric evaluated for samples from `inlabru`. 95% confidence intervals are plotted for SB-ETAS across 10 different initial seeds.

3.6 SCEDC Catalog

We now evaluate SB-ETAS on some observational data from Southern California. The Southern California Seismic Network has produced an earthquake catalog for Southern California going back to 1932 [88]. This catalog contains many infamous large earthquakes such as the 1992 M_W 7.3 Landers, 1999 M_W 7.1 Hector Mine and M_W7.1 Ridgecrest sequences. We use $N = 43,537$ events from 01/01/1981 - 31/12/2021 with earthquake magnitudes $\geq M_W$ 2.5 since this assures the most data completeness [88]. The catalog can be downloaded from the Southern California Earthquake Data Center `https://service.scedc.caltech.edu/ftp/catalogs/SCSN/`.

This size of catalog contains too many events to find ETAS posteriors using `bayesianETAS` (i.e. it would take longer than 2 weeks). Therefore we run only SB-ETAS and `inlabru` on the entire catalog and validate their performance by comparing the compensator, $\Lambda^*(t;\theta) = \int_0^t \lambda^*(s;\theta)ds$, with the observed cumulative number of events in the catalog $N(t)$. $\Lambda^*(t;\theta)$ gives the expected number of events at time t, and therefore a model and its parameters are consistent with the observed data if $\Lambda^*(t;\theta) = N(t)$.

We generate 5,000 samples using SB-ETAS and `inlabru` and use each sample to generate a compensator curve $\Lambda^*(t;\theta)$. We display 95% confidence intervals of these curves in Figure 3.9, along with a curve for the Maximum Likelihood Estimate (MLE). Consistent with the synthetic experiments, we find that SB-ETAS gives a conservative estimate of the cumulative number of events across the catalog, whereas `inlabru` is overconfident and does not contain the observed number of events within its very narrow confidence interval. Both `inlabru` and the MLE match the total observed number of events in the catalog, since this value, $\Lambda^*(T)$, is a dominant term in each of their loss functions (the likelihood) during estimation.

For both the MLE and SB-ETAS, we fix the α parameter equal to the β parameter of the Gutenberg-Richter law $f_{GR(m)}$, a result that is consistent with other temporal only studies of Southern California [43, 81], and which reproduces Båth's law for aftershocks [44]. We were unable to successfully fix α for `inlabru` and therefore use the 5 parameter implementation of ETAS.

Posterior distributions are displayed in Figures 3.10 and 3.11 including $\alpha = \beta$ and free α implementations of the MLE. Although the modes of the marginal distributions do not match the MLE, the SB-ETAS posteriors contain the MLE parameters within their wider confidence ranges. Since `inlabru` has much narrower confidence, although the distributions are relatively close to the MLE, the confidence ranges do not contain MLE parameters.

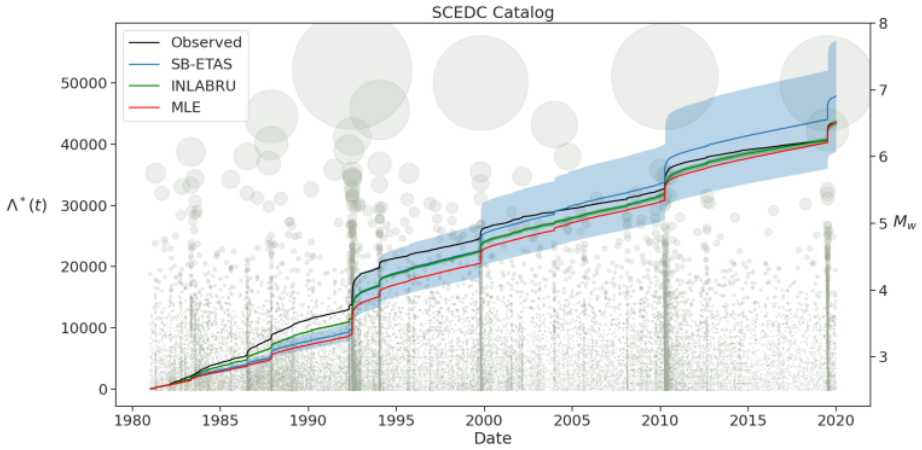


Figure 3.9: The compensator $\Lambda^*(t)$ found from estimating the ETAS posterior distribution on the SCEDC catalog (events displayed in background). 5,000 Samples from the posterior using both SB-ETAS and `inlabru` were used to generate a mean and 95% confidence interval. The compensator is compared against the observed cumulative number of events in the catalog along with the MLE.

3.7 Discussion and Conclusion

The growing size of earthquake catalogs generated through machine learning based phase picking and an increased density of seismic networks, calls for the application of a broader range of models to assess whether the new data enhances forecasting capabilities. Furthermore, this growth demands that our existing models scale effectively to handle the new volume of data. We propose using a simulation-based approach, where models are defined by a simulator without the need for a likelihood function, thereby alleviating some modeling constraints. Simulation based inference (SBI) performs Bayesian inference for such models using outputs of the simulator in place of the likelihood. SB-ETAS: our simulation based estimation procedure for the Epidemic Type Aftershock Sequence (ETAS) model, establishes an initial connection between earthquake modeling and simulation based inference, demonstrating improved scalability over previous methods.

In our study, using SB-ETAS we generate samples of the ETAS posterior distribution for a series of synthetic catalogs as well as a real earthquake catalog from southern California. Additionally, we generate samples using another approximate inference method: `inlabru`. Our general finding is that `inlabru` produces overconfident and sometimes biased posterior estimates, while SB-ETAS provides more conservative and less biased estimates. Although it

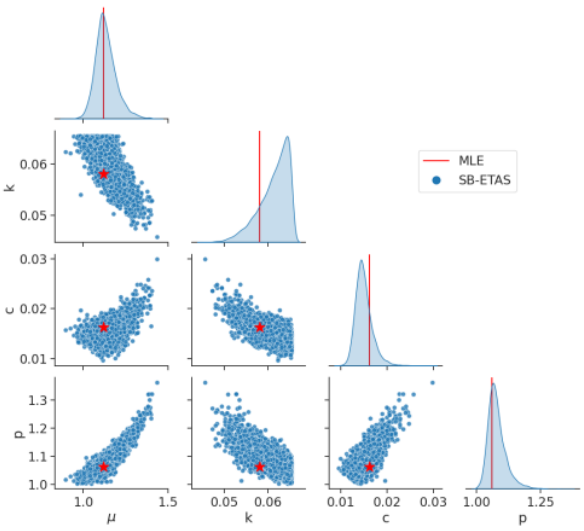


Figure 3.10: The posterior distribution of ETAS parameters found on the SCEDC catalog using SB-ETAS. This implementation of ETAS fixes $\alpha = \beta$. MLE parameters are plotted for comparison.

might seem reasonable to judge an approximate posterior by its closeness to the exact posterior, for practical use, overconfident estimates should be penalised more than under-confident ones. Bayesian inference seeks to identify a range of parameter values which are then used to give confidence over a range of earthquake forecasts. However, failure to identify regions of the parameter space that give likely parameters, would result in omission of a range of likely forecasts.

Although improvements have been made to reduce the computational time of performing Bayesian inference for the ETAS model, first with `bayesianETAS` followed by `inlabru`, neither of these approaches improve upon the scalability of inference. Therefore as catalogs continue to grow in size, these methods become less feasible to use. On experiments where we give SB-ETAS, `bayesianETAS` and `inlabru` access to the same 8 CPUs, only SB-ETAS could be used to fit a catalog of 500,000 events and was the fastest method for catalogs above 100,000 events. Both `inlabru` and SB-ETAS are parallelized methods and would therefore see a reduction in runtime if given access to more CPUs. This is unlike `bayesianETAS` which is not parrallelized in its current implementation. It is also worth noting that although SB-ETAS and `inlabru` were given the same CPUs, `inlabru` required over 4 times the amount of memory than SB-ETAS with catalogs over 100,000 events, (Figure B.8). This additional memory demand far exceeds the capacity typically available on standard laptops.

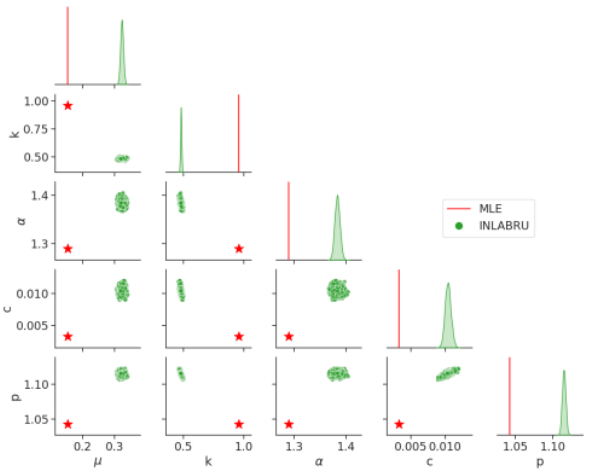


Figure 3.11: The posterior distribution of ETAS parameters found on the SCEDC catalog using `inlabru`. This implementation of ETAS has a free α parameter. MLE parameters are plotted for comparison.

A clear limitation of this inference procedure is that the posterior distribution must lie in the sub-critical region of the parameter space. Super-critical parameters, which lie outside this region, result in simulations that explode with non-zero probability. That is, infinitely many earthquakes would be simulated within the finite time window. In our experiments, to avoid this we enforce a sub-critical parameter region using the prior. There is however, the possibility that the "true" posterior lies outside of the prior. While this may an immediate problem for SB-ETAS, MCMC or `inlabru` do not circumvent the problem when forecasts are made. Generating forecasts requires simulating multiple earthquake catalogs, and therefore super-critical parameters will result in explosive forecasts. The practical solution is to discard such forecasts, however this ignores the fact that the model is unable to successfully recreate real earthquake sequences over extended time periods: we do not observe infinitely many earthquakes occurring in nature.

An inability to replicate nature indicates a poorly fit or misspecified model. Restricted by our need for non-critical simulations we wish to advocate for models which are sub-critical. Developing models in a simulation based way could ensure that fitted models better resemble nature. Using a truncated magnitude distribution [195], which expands the size of the sub-critical region, or by fixing the alpha parameter, provide small model alterations which reduce criticality. For the SCEDC catalog the branching ratio of the 5 parameter MLE was $\eta = 2.033$, compared with $\eta = 0.699$ for the 4 parameter implementation. More significant alterations such as considering a spatially varying background rate [143, 144] have led to

sub-critical models, compared with super-critical one that use a uniform background rate. Furthermore, time-varying parameters may account for the "intermittent" criticality of the system [12, 74].

Equally, improperly considering boundary effects, in space, time and magnitude can lead to poor estimation of a model's criticality [183, 196, 222]. Models that consider events outside of the observed space-time-magnitude region, may better replicate nature. This could include simulating additional observed events [e.g. 187, 188], or unobserved events [e.g. 33] that both have triggering capabilities.

SB-ETAS is particularly well aligned for modeling such contributions from unobserved events. For example, consider the same ETAS branching process used in this study, but instead events are deleted with a time varying probability $h(t)$. The induced likelihood of this process,

$$p(\mathbf{x}|\theta) \propto \int p(\mathbf{x}, \mathbf{x}_u|\theta) \prod_{t_i \in \mathbf{x}} h(t_i) \prod_{t_j \in \mathbf{x}_u} (1 - h(t_j)) d\mathbf{x}_u, \tag{3.17}$$

is intractable since it involves integrating over the set of unobserved events $\mathbf{x}_u$ [33]. Current methods to deal missing data estimate the true earthquake rate from the apparent earthquake rate assuming no contribution from undetected events [65]. A likelihood-free method of inference such as SB-ETAS could avoid the biases from ignoring such triggering [196].

There is a natural extension to SB-ETAS for the spatio-temporal form of the ETAS model. The spatio-temporal ETAS extends the temporal model used in the study by modeling earthquake spatial interactions with an isotropic Gaussian spatial triggering kernel [151]. It is also defined as a branching process and so retains the $\mathcal{O}(n \log n)$ complexity of simulation. This study has illustrated that the Ripley K-statistic is an informative summary statistic for the triggering parameters of the temporal ETAS model. It seems fair to assume that the spatio-temporal Ripley K-statistic,

$$\hat{K}(\mathbf{x}, w_t, w_s) = \frac{AT}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(0 < t_j - t_i \leq w_t) \mathbb{I}(||s_j - s_i||_2 \leq w_s).$$

where A is the area of the study region, would be a reasonable choice for the spatio-temporal form of SB-ETAS. This statistic loses the $\mathcal{O}(n)$ efficiency that the purely temporal one benefits from. Instead Wang et al. [223] have developed a distributed procedure for calculating this statistic with $\mathcal{O}(n \log n)$ complexity that would retain the overall time complexity that SB-ETAS has.

Ideally, the value of the Ripley K-statistic $\hat{K}(\mathbf{x}, w)$ for all $w \in \mathbb{R}_+$ would be used as the summary statistic for the observed data $\mathbf{x}$. However, since the neural density estimator requires

a fixed length vector as input, we have to sample this function at pre-specified intervals. Increasing the number of samples would increase the dimension of this fixed length vector, making the density estimation task more challenging. On the other hand, using fewer samples w, would make the density estimation task easier but would reduce the information contained in the summary statistic. Future work, should address how to balance the number of samples of the Ripley K-statistic as well as moving beyond the hand chosen values used in this study. We speculate that the loss of information from under-sampling the K-statistic, weakens the generalisation of the method in its current form, e.g. the MMD for the MaxT=30000 experiment does not decrease over the simulation rounds (Figure 3.4).

Further model expansion using this simulation based framework could help estimate earthquake branching models that include complex physical dependencies. One possible example would be to calibrate the Third Uniform California Earthquake Rupture Forecast ETAS Model (UCERF3-ETAS), a unified model for fault rupture and ETAS earthquake clustering [46]. This model extends the standard ETAS model by explicitly modeling fault ruptures in California and includes a variable magnitude distribution which significantly affects the triggering probabilities of large earthquakes. This model is only defined as a simulator and uses ETAS parameters found independently to the joint ETAS and fault model. In fact, Page and van der Elst [158] validate the models performance through a comparison of summary statistics from the outputs of the model. This validation could be extended to comprise part of the inference procedure for model parameters using the same simulation based framework as SB-ETAS.

Chapter 4

EarthquakeNPP: Benchmark Datasets for Earthquake Forecasting with Neural Point Processes

Declaration

The methodology, experimentation and writing of this chapter was undertaken by me, Samuel Stockman, with guidance from my two supervisors: Maximilian Werner and Daniel Lawson.

The chapter closely follows a paper currently under review: Stockman, Samuel, Daniel J. Lawson, and Maximilian J. Werner. EarthquakeNPP: Benchmark Datasets for Earthquake Forecasting with Neural Point Processes. The Supplementary Material has been included as Appendix C.

Abstract

Classical point process models, such as the epidemic-type aftershock sequence (ETAS) model, have been widely used for forecasting the event times and locations of earthquakes for decades. Recent advances have led to Neural Point Processes (NPPs), which promise greater flexibility and improvements over classical models. However, the currently-used benchmark dataset for NPPs does not represent an up-to-date challenge in the seismological community since it lacks a key earthquake sequence from the region and improperly splits training and testing data. Furthermore, initial earthquake forecast benchmarking lacks a comparison to state-of-the-art earthquake forecasting models typically used by the seismological community. To address these gaps, we introduce EarthquakeNPP: a collection of benchmark datasets to facilitate testing of NPPs on earthquake data, accompanied by a credible implementation of the ETAS model. The datasets cover a range of small to large target regions within California, dating from 1971 to 2021, and include different methodologies for dataset generation. In a benchmarking experiment, we compare three spatio-temporal NPPs against ETAS and find that none outperform ETAS in either spatial or temporal log-likelihood. These results indicate that current NPP implementations are not yet suitable for practical earthquake forecasting. However, EarthquakeNPP will serve as a platform for collaboration between the seismology and machine learning communities with the goal of improving earthquake predictability.

4.1 Introduction

Operational earthquake forecasting by global governmental organisations such as the US Geological Survey (USGS) necessitates the development of models which can forecast the times and locations of damaging earthquakes. While model development is ongoing in the seismology community, recent improvements have relied upon refinement of a spatio-temporal point process model known as the Epidemic-Type Aftershock Sequence (ETAS) model [150, 151], despite significant growth in available data [139, 140, 175, 194, 204, 206, 225].

In contrast, the machine learning community has offered promising advancements over classical point process models like ETAS with Neural Point Process (NPP) models, showcasing greater flexibility [17, 39, 91, 157, 190, 239, 240]. While some initial benchmarking of these models has been conducted on an earthquake dataset in Japan, these experiments lack relevance for stakeholders in the seismology community. The benchmark lacks a key earthquake sequence from the region, fails to recreate an operational setting with proper train-test splits, and doesn't compare against state-of-the-art models like ETAS.

Here, we introduce EarthquakeNPP: a curated collection of datasets designed for benchmarking NPP models in earthquake forecasting, accompanied by a state-of-the-art

benchmark model. These datasets are derived from publicly available raw data, which we process and configure within our platform to facilitate meaningful forecasting experiments relevant to stakeholders in the seismology community. Covering various regions of California, these datasets represent typical forecasting zones and encompass data commonly utilized by forecast issuers. Moreover, employing modern techniques, some datasets include smaller magnitude earthquakes, exploring the potential of numerous small events to enhance forecasting performance through flexible NPPs. To unify efforts, we present an operational-level implementation of the ETAS model alongside the datasets, serving as a benchmark for NPPs.

Although initial benchmarking finds that none of the 3 tested NPP implementations outperform ETAS, EarthquakeNPP aims to serve as a platform for future NPP development. The platform facilitates the generative evaluation procedure used for rigorous benchmarking in the seismology community, directing the impact of future NPPs to stakeholders in seismology. Access to the dataset collection, along with comprehensive documentation and notebooks, can be found at `https://github.com/ss15859/EarthquakeNPP`.

4.1.1 Related Work

Existing Benchmark Dataset. Chen et al. [17] introduced an earthquake dataset for benchmarking the Neural Spatio-temporal Point Process (NSTPP) model using a global dataset from the U.S. Geological Survey, focusing on Japan from 1990 to 2020. They considered earthquakes with magnitudes above 2.5, splitting the data into month-long segments with a 7-day offset. They exclude earthquakes from November 2010 to December 2011, deeming these sequences "too long" and "outliers." However, this period includes the 2011 Tohoku earthquake [138], the largest earthquake recorded in Japan and the fourth largest in the world, at magnitude 9.0. This exclusion renders the benchmarking experiment irrelevant for seismologists, as it is precisely these large earthquakes and their aftershocks that are crucial to forecast due to their damaging impact. Additionally, these events are of significant scientific interest because they provide valuable insights into the earthquake rupture process.

The dataset segments are divided for training, testing, and validation. Instead of a chronological partitioning that mirrors operational forecasting, the segments are assigned in an alternating pattern. This approach misrepresents a realistic forecasting scenario and inflates performance measures due to earthquake triggering [50]. Since the model is tested on windows immediately preceding training windows, it exploits causal dependencies backwards it time.

Although earthquakes with magnitudes above 2.5 are considered by Chen et al. [17], following a change in USGS policy on global data collection, from 2009 onwards, only events above magnitude 4.0 are recorded in the dataset. For earthquake forecasting in Japan, seismologists

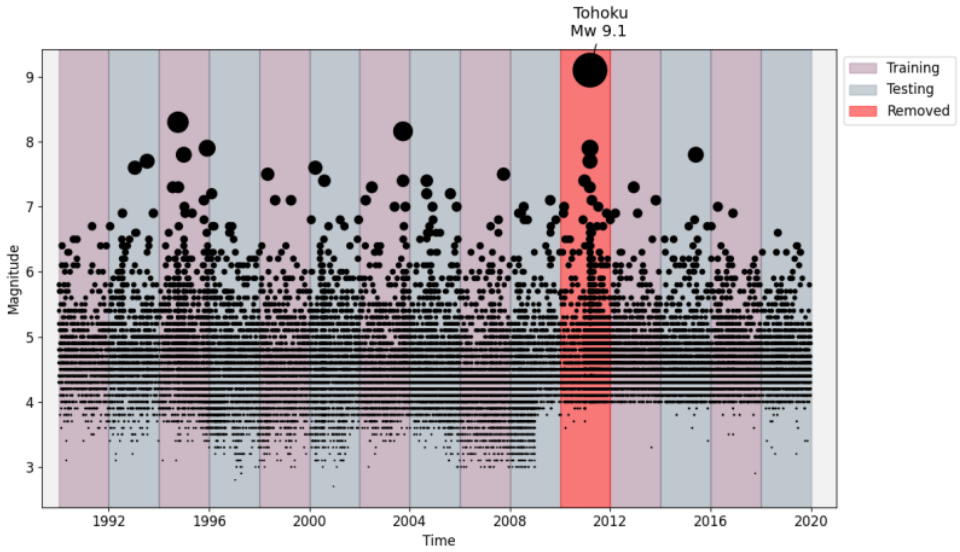


Figure 4.1: ANSS Comprehensive Earthquake Catalog, focusing on Japan from 1990 to 2020, constructed by Chen et al. [17] to benchmark NPPs. Earthquakes above $M_w 2.5$ are considered and the data is partitioned for training and testing in an alternating pattern. For the pattern, the authors use month long segments with a 7 day overlap, however to aid illustration we plot 2 year segments. The authors also exclude the Tōhoku earthquake sequence, under the pretext of removing outliers.

use datasets from Japanese data centers since they are more comprehensive and complete than global datasets. Section C.1.2 describes the biases incurred from such data missingness.

Chen et al. [17] benchmark their model against another spatio-temporal model, Neural Jump SDEs [91], and a temporal-only Hawkes process, even though a spatio-temporal Hawkes process would provide a more rigorous benchmark. Subsequent papers adopting this benchmark [235, 239, 240] similarly lack comparisons to a spatio-temporal Hawkes process, benchmarking instead against temporal-only or spatial-only baselines or other spatio-temporal NPPs.

Temporal-NPP Benchmarking on Earthquake Data. Two existing works benchmark NPPs for earthquake forecasting within the seismology community. The first by Dascher-Cousineau et al. [31] extends a temporal-only NPP from Shchur et al. [190] to include earthquake magnitudes. The second by Stockman et al. [201] extends another temporal-only model by Omi et al. [157] to target larger magnitude events. Both models are benchmarked against a temporal ETAS model, showing moderate improvements over the baseline. Extending these models to include spatial data is necessary for further testing and potential operational

use in the seismological community.

Benchmarking within the Seismology Community. Model comparison has been crucial in the development of earthquake forecasting models since their inception [96, 150]. The Collaboratory for the Study of Earthquake Predictability (CSEP) [89, 128, 177, 181] [`https://cseptesting.org/`] aims to unify the framework for earthquake model testing and evaluation, hosting retrospective and fully prospective forecasting experiments globally. CSEP benchmarks short-term models using performance metrics that require forecasts to be generated by simulating many repeat sequences over a specified time horizon (typically one day). These simulated forecasts are compared by discretizing time and space intervals, with test statistics calculated for event counts, magnitudes, locations, and times. The simulation-based approach allows the inclusion of generative models that don't output explicit earthquake probabilities (i.e., a likelihood), and enables evaluation of the full distribution of entire sampled sequences.

4.1.2 Scope of this work

Since generating repeated sequences over forecast horizons is computationally costly, the seismology community uses the mean log-likelihood on held-out data for a more streamlined metric during model development [75, 150]. Our platform uses this metric in the NPP benchmarking experiment and provides detailed guidance on CSEP's simulation-based procedure, enabling future NPP implementations and evaluations within CSEP experiments.

The goal of this work is to allow Machine Learning researchers to have seismological impact by defining a baseline target for which NPP models can be compared to state-of-the-art domain-based models. NPPs that can generate log-likelihoods are in scope, whilst those that do not [e.g. 107, 235] are out of scope because a valid score does not currently exist. The popular next-event point prediction metrics (e.g. Root Mean Square Error (RMSE) and related scores) are considered to be flawed and misleading for seismological prediction [86], because the predictive distribution is strongly skewed and therefore far from Gaussian. To have seismological relevance, authors of NPP models are challenged to implement long-term predictions using CSEP's evaluation procedure, benchmarking against the reported performance for the ETAS model.

4.2 Background

4.2.1 Spatio-Temporal Point Processes

A spatio-temporal point process is a continuous-time stochastic process that models the random number of events $N(S \times (t_a, t_b])$ which occur in a space-time interval

$\mathcal{S} \times (t_a, t_b]$, $\mathcal{S} \in \mathbb{R}^2$, $(t_a, t_b] \in \mathbb{R}^+$. This process is typically defined by a non-negative *conditional intensity function*

$$\lambda(t, \mathbf{x}|\mathcal{H}_t) := \lim_{\Delta t, \Delta \mathbf{x} \to 0} \frac{\mathbb{E}\left[N([t, t+\Delta t) \times B(\mathbf{x}, \Delta \mathbf{x})|\mathcal{H}_t\right]}{|B(\mathbf{x}, \Delta \mathbf{x})|}, \tag{4.1}$$

where $\mathcal{H}_t = \{(t_i, \mathbf{x}_i)|t_i < t\}$ denotes the history of events preceding time t and $|B(\mathbf{x}, \Delta \mathbf{x})|$ is the Lebesgue measure of the ball $B(\mathbf{x}, \Delta \mathbf{x})$ with radius $\Delta \mathbf{x}$. Given we observe a history of events up to t_i, the probability density function (pdf) of observing an event at time t and location $\mathbf{x}$ is given by

$$p(t, \mathbf{x}|\mathcal{H}_{t_i}) = \lambda(t, \mathbf{x}|\mathcal{H}_{t_i}) \cdot \exp\left(-\int_{t_i}^{t} \int_{\mathcal{S}} \lambda(s, \mathbf{z}|\mathcal{H}_s) d\mathbf{z} ds\right). \tag{4.2}$$

Most models specify the conditional intensity function, though some [e.g. 17, 190, 235], directly model this pdf. Model parameters are typically estimated by maximizing the log-likelihood of observed events within a training time interval $[T_0, T_1]$ and spatial region $\mathcal{S}$,

$$\log p(\mathcal{H}_T) = \underbrace{\sum_{i=0}^{n} \log \lambda(t_i|\mathcal{H}_{t_i}) - \int_{T_0}^{T_1} \int_{\mathcal{S}} \lambda(s, \mathbf{z}|\mathcal{H}_s) d\mathbf{z} ds}_{\text{Temporal log-likelihood}} + \underbrace{\sum_{i=0}^{n} \log f(\mathbf{x}_i|t_i, \mathcal{H}_{t_i})}_{\text{Spatial log-likelihood}}, \tag{4.3}$$

where the decomposition of the spatio-temporal conditional intensity function, $\lambda(t_i, \mathbf{x}_i|\mathcal{H}_{t_i}) = \lambda(t_i|\mathcal{H}_{t_i}) \cdot f(\mathbf{x}_i|t_i, \mathcal{H}_{t_i})$, allows the log-likelihood to be written as contributions from the temporal and spatial components. In practice, this exact function is often not maximized directly during training: for models specified through the conditional intensity function, an analytical solution to the integral term is generally not possible and is approximated numerically.

For model evaluation and comparison, the log-likelihood of observing events in the test set can be used as a performance metric. This is consistent with a wealth of literature in the seismology community [see 236, and references therein] as well as the wider general point process literature [28], which now more recently includes neural point processes [192]. The metric evaluates models that output probability distributions over their predictions and consequently penalises models that are overconfident. Although evaluating on events in the test set, the test log-likelihood, $\log p\left((t_i, \mathbf{x}_i)|t_i \in [T_2, T_3], \mathcal{H}_{T_2}\right)$, may still contain dependence upon events prior to the test window $[T_2, T_3]$, typically contained in the history $\mathcal{H}_{T_2}$ of the intensity function. Comparing the mean log-likelihood per event provides the *information gain* from one model to another [28].

4.2.2 ETAS

The Epidemic Type Aftershock Sequence (ETAS) model [151] is a spatio-temporal Hawkes process which models how earthquakes cluster in time and space. It has been adopted for

operational earthquake forecasting by government agencies in California [131], New-Zealand [23], Italy [197], Japan [156] and Switzerland [135], and performs consistently well in CSEP's retrospective and fully prospective forecasting experiments [e.g. 16, 118–120, 171, 207, 228]. The general formulation of the model is

$$\lambda(t, \mathbf{x} | \mathcal{H}_t; \theta) = \mu + \sum_{i:t_i<t} g(t - t_i, ||\mathbf{x} - \mathbf{x}_i||_2^2, m_i), \tag{4.4}$$

where μ is a constant background rate of events, $g(\cdot, \cdot, \cdot)$ is a non-negative excitation kernel which describes how past events contribute to the likelihood of future events and m_i are the associated magnitudes of each event. The equivalent formulation as a Hawkes branching process accompanies a causal branching structure $\mathbf{B}$. This concept broadly aligns with the understanding of the physics of earthquake triggering and interaction, e.g. via dynamic wave triggering [14] and static stress triggering [58, 119].

Although ETAS can be fit by maximizing the log-likelihood function directly, parameter estimation is typically performed by simultaneously estimating the branching structure $\mathbf{B}$. Veen and Schoenberg [221] developed an Expectation Maximisation (EM) procedure, which maximises the marginal likelihood over the unobserved branching structure, $\log \int p(\mathcal{H}_{T_1} | \mathbf{B}, \theta) p(\mathbf{B}|\theta) d\mathbf{B}$ through the iteration

$$\theta^{(k+1)} = \arg\max_{\theta} \mathbb{E}_{\mathbf{B} \sim p(\cdot | \mathcal{H}_{T_1}, \theta^{(k)})} \left[\log p(\mathcal{H}_{T_1}, \mathbf{B} | \theta) \right]. \tag{4.5}$$

This avoids the need to numerically approximate the integral term in the likelihood, provides more stability during estimation and simultaneously estimates the causal structure.

The formulation of the ETAS model we present with the EarthquakeNPP datasets is implemented in the `etas` python package by Mizrahi et al. [134]. It defines the triggering kernel as

$$g(t, r^2, m) = \frac{e^{-t/\tau} \cdot k \cdot e^{a(m - M_{cut})}}{(t + c)^{1+\omega} \cdot \left(r^2 + d \cdot e^{\gamma(m - M_{cut})}\right)^{1+\rho}}, \tag{4.6}$$

where r^2 is the squared distance between events and $k, a, c, \omega, \tau, d, \gamma, \rho$ are the learnable parameters along with the constant background rate μ.

4.3 EarthquakeNPP Datasets

The EarthquakeNPP datasets encompass earthquake records, including timestamps, geographical coordinates, and magnitudes, documented within California from 1971 to 2021. California, with its dense network and high seismic hazard, has been extensively studied, demonstrating the utility of forecasting algorithms [45, 47, 52]. It encompasses the San

Andreas fault plate boundary system [246] and includes modern high-resolution catalogs with numerous small magnitude earthquakes, offering potential for new, more expressive models.

Notebooks to access and preprocess these public datasets along with the associated benchmarking experiment are publicly accessible at `https://github.com/ss15859/EarthquakeNPP`, accompanied by more detailed documentation for each dataset. A summary of how earthquake datasets are generated, along with the associated challenges of using earthquake catalog data can be found in Appendix C.1. The following subsections provide a short overview of each EarthquakeNPP dataset.

4.3.1 ANSS Comprehensive Earthquake Catalog (ComCat)

The U.S. Geological Survey (USGS) National Earthquake Information Center (NEIC) monitors global earthquakes (Mw 4.5 or larger) and provides complete seismic monitoring of the United States for all significant earthquakes (> Mw 3.0 or felt). Its contributing seismic networks have produced the Advanced National Seismic System (ANSS) Comprehensive Catalog of Earthquake Events and Products. We focus on the California region defined by Schorlemmer and Gerstenberger [178], with a test period consistent with CSEP experiments [237].

4.3.2 Southern California Earthquake Data Center (SCEDC) Catalog

The Southern California Seismic Network (SCSN) has developed and maintained the standard earthquake catalog for Southern California [88] since the Caltech Seismological Laboratory began routine operations in 1932. Significant network improvements since the 1970s and 1980s reduced the catalog completeness from Mw 3.25 to Mw 1.8. We use three magnitude thresholds (Mw 2.0, 2.5, 3.0) to explore the effect of truncation on forecasting model performance. Training includes the Mw 7.3 Landers and 1999 Mw 7.1 Hector Mine earthquakes, while testing involves the 2019 Mw 7.1 Ridgecrest sequence.The USGS utilizes both `ComCat` and `SCEDC` datasets in the aftershock forecasts they release to the public. The inclusion of these datasets determines whether NPPs can exploit the datasets currently being used for operational forecasting.

4.3.3 Detailed Earthquake Catalog for the San Jacinto Fault-Zone Region

White et al. [225] created an enhanced catalog focusing on the San Jacinto fault region, using an dense seismic network in Southern California. This denser network, combined with automated phase picking (STA/LTA), ensures a 99% detection rate for earthquakes greater than Mw 0.6 in a specific subregion [225]. The training window includes the 2010 Mw 5.4 Borrego Springs and 2013 ML 4.7 Anza Borrego earthquakes. This catalog is named `White` after the authors.

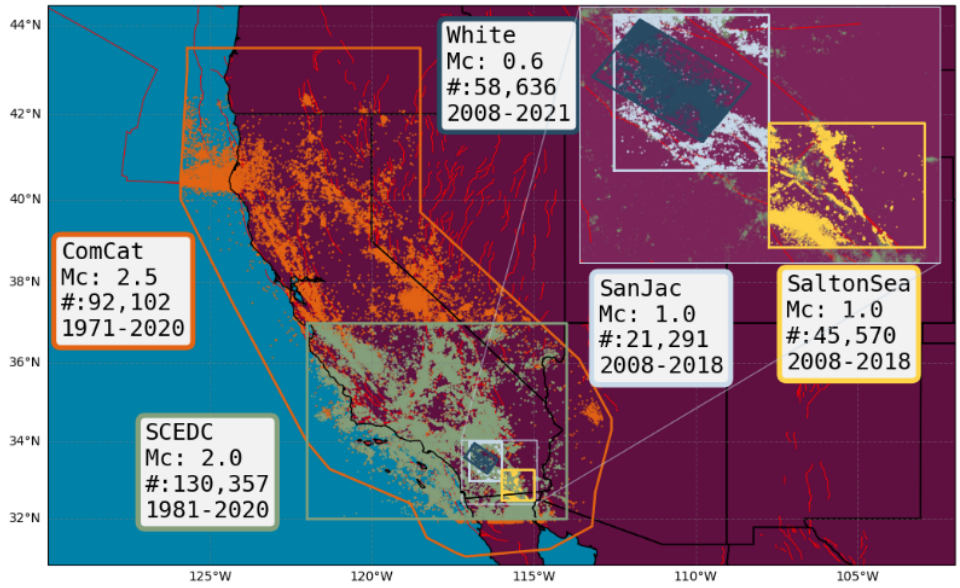


Figure 4.2: Earthquakes contained in the observational datasets found in EarthquakeNPP. Colours indicate the respective datasets, including the target region, magnitude of completeness M_c, number of events and the time period that the dataset spans. In red is a fault map from the GEM Global Active Faults Database [203].

4.3.4 Quake Template Matching (QTM) Catalog

Using data collected by the SCSN, Ross et al. [175] generated a denser catalog by reanalyzing the same waveform data with a template matching procedure that looks for cross-correlations with the wavetrains of previously detected events. The catalog contains 1.81 million earthquakes complete down to Mw 0.3. Following Dascher-Cousineau et al. [31], we use a more conservative completeness estimate of Mw 1.0 and split the catalog into two focus regions: the San Jacinto fault region and the Salton Sea. The inclusion of `White`, `QTM_SanJac` and `QTM_SaltonSea` determines whether very low magnitude earthquakes exhibit different behaviour to those of the larger earthquakes, an assumption which ETAS makes.

4.3.5 Additional Datasets

Beyond the official EarthquakeNPP datasets, we include 3 further datasets that either provide additional scientific insight or continuity from previous benchmarking works.

Synthetic ETAS Catalogs. We simulate a synthetic catalog using the ETAS model with parameters estimated from ComCat, at M_c 2.5, within the same California region. A second catalog emulates the time-varying data-missingness present in observational catalogs by

removing events using the time-dependent formula from Page et al. [159],

$$M_c(M,t) = M/2 - 0.25 - \log_{10}(t), \tag{4.7}$$

where M is the mainshock magnitude. Events below this threshold are removed using mainshocks of Mw 5.2 and above. The inclusion of these datasets allows us to test whether NPPs are inhibited by data missingness to the same extent that ETAS is.

Deprecated Catalog of Japan. To provide continuity from the previous benchmarking for NPPs on earthquakes, we also provide results on the Japanese dataset from Chen et al. [17], however with a chronological train-test split and without removing any supposed outlier events. To reflect our recommendation not to use this dataset in any future benchmarking following the dataset completeness issues mentioned above, we name this dataset `Japan_Deprecated`.

4.4 Benchmarking Experiment

We now use EarthquakeNPP to benchmark three spatio-temporal NPPs with prior positive claims on earthquake forecasting.

Neural Spatio-Temporal Point Process (NSTPP) [17]**:** a pdf based NPP that parameterizes the spatial pdf with continuous-time normalizing flows (CNFs). We use their Attentive CNF model for its computational efficiency and overall performance versus their other model Jump CNF [17].

Deep Spatio-Temporal Point Process (Deep-STPP) [240]**:** a conditional intensity function based NPP that constructs a non parametric space-time intensity function governed by a deep latent process. The intensity function enjoys a closed form integration, avoiding the need for numerical approximation.

Automatic Integration for Spatiotemporal Neural Point Processes (AutoSTPP) [239]**:** a conditional intensity function based NPP which jointly models the 3D space-time integral of the intensity along with its derivative (the intensity function) using a dual network approach.

The benchmark is against the **ETAS** model defined in section 4.2.2, as well as a homogeneous **Poisson** process. The Poisson model is fit to events in the auxiliary, training and validation windows to provide a baseline score against which to compare all four other models.

Validation is typically not part of the estimation procedure for ETAS, so it is fit using the combined training and validation windows. NPPs follow the standard

training/validation/testing procedure of machine learning. When possible, a model's likelihood for training, validation, and testing can depend on events occurring before the splits through memory in its history. The exception is NSTPP, lacking a direct dependency on prior events. Nonetheless, its likelihood is evaluated on the same events as the other models. The definition of the ETAS model (equation 4.4) specifies how the magnitudes of earthquakes in the history contribute towards the intensity function. This earthquake magnitude dependence is not implemented in any of the NPPs we benchmark, since it requires modeling choices beyond the scope of this work.

Figures 4.3 and 4.4 report the temporal and spatial log-likelihood scores of all models on the EarthquakeNPP datasets. The ETAS model achieves the highest temporal and spatial log-likelihood across all datasets, with some NPP models achieving comparable temporal performance on `ComCat`, `QTM_SaltonSea`, `QTM_SanJac`, and `White` catalogs. Amongst the NPP models, Deep-STPP generally performs best in terms of temporal log-likelihood, whereas AutoSTPP performs best in terms of spatial log-likelihood. The improved relative temporal performance of all NPPs compared to ETAS as the magnitude threshold is lowered from 3.0 to 2.0 in the `SCEDC` dataset, as well as the comparable performance to ETAS using the low magnitude catalogs `QTM_SaltonSea`, `QTM_SanJac`, and `White`, indicates that low magnitude earthquakes provide valuable predictive information for NPPs.

4.4.1 Additional Benchmark Results

Figures 4.5 and 4.6 report the temporal and spatial log-likelihood scores of all the benchmarked models on additional datasets. On synthetic data generated by the ETAS model the performance of NPPs mirrors the results on the observational data (Figures 4.3 and 4.4). The performance of NPPs is more comparable to ETAS in terms of temporal log-likelihood however they cannot capture the distribution of earthquake locations. Change in temporal performance of models between the `ETAS` and `ETAS_incomplete` datasets reveal each model's robustness to the missing data typically present in earthquake catalogs. Auto-STPP and ETAS reduce in performance upon the removal earthquakes during aftershock sequences, whereas DeepSTPP and NSTPP maintain the same performance indicating a robustness to the data missingness.

On the `Japan_Deprecated` dataset, whilst ETAS remains the best performing model for spatial prediction, for temporal prediction it performs comparably to NSTPP and is even marginally outperformed by DeepSTPP. This performance can be attributed to the data completeness issues of the `Japan_Deprecated` dataset (see section 4.1.1), where the test period is missing all earthquakes bellow magnitude 4.0.

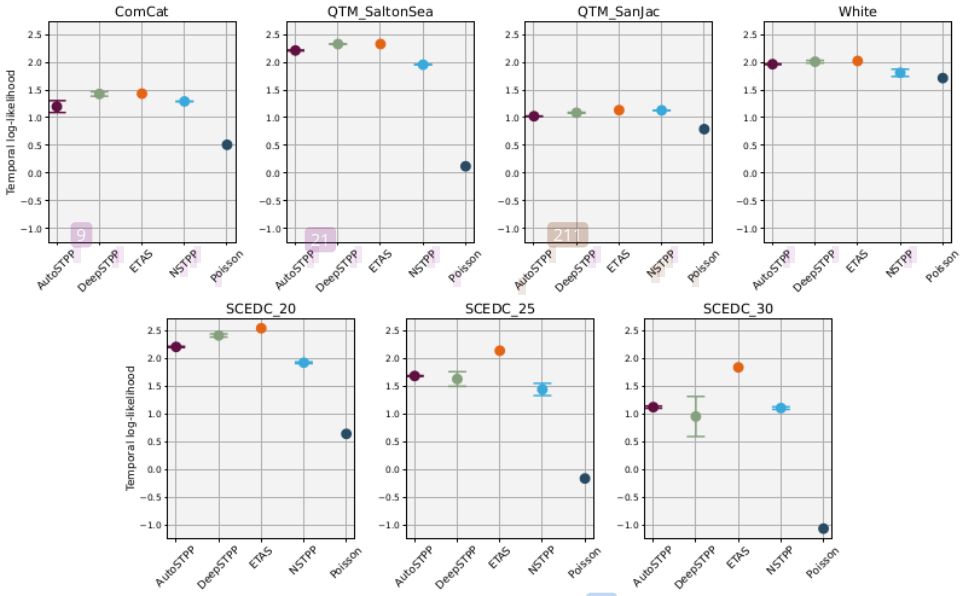


Figure 4.3: Test temporal log-likelihood scores for all the spatio-temporal point process models on each of the EarthquakeNPP datasets. Error bars of the mean and standard deviation are constructed for the NPPs using three repeat runs.

4.5 CSEP Consistency Tests

EarthquakeNPP supports the earthquake forecast evaluation protocol developed by the Collaboratory for the Study of Earthquake Predictability (CSEP). In this procedure a model generates 24-hour forecasts through 10,000 repeat simulations of earthquake sequences at the beginning of every day in the testing period. This procedure exactly mimics how earthquake forecasts are generated in an operational setting [217]. Models can then be evaluated by comparing the observed sequence with the distribution over model simulations. Three test statistics target the temporal, spatial and magnitude components of the forecasts, where a test is failed if the observed statistic falls within a pre-defined rejection region (Figure 4.7). We demonstrate this procedure for the ETAS model and report performance scores as a benchmark for future immplementations of NPPs.

4.5.1 Number Test

The number test evaluates the temporal component of the forecast by checking the consistency of the forecasted number of events, N with those observed in the forecast horizon, N_{obs}. Upper and lower quantiles are estimated using the empirical cumulative distribution from the repeat

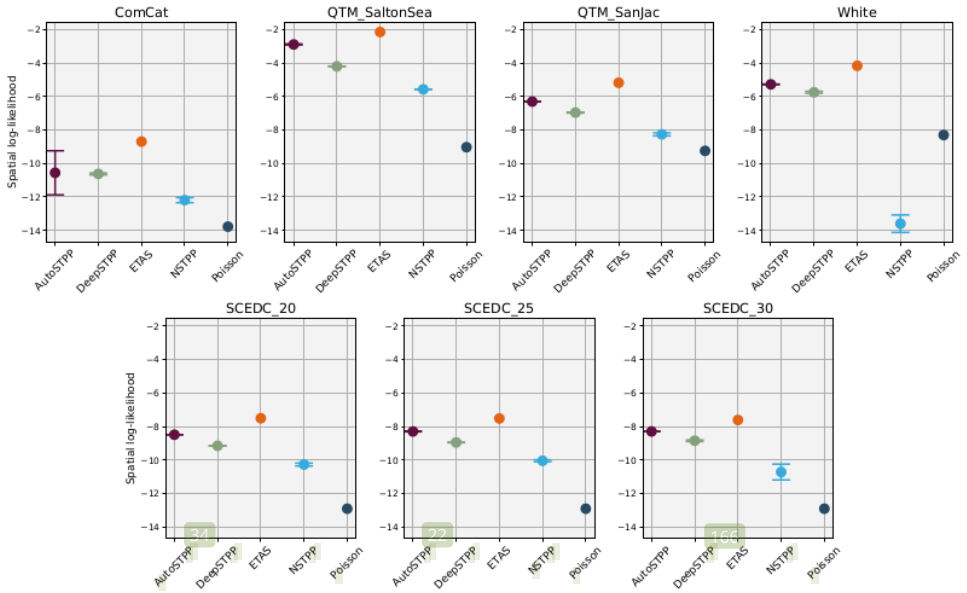


Figure 4.4: Test spatial log-likelihood scores for all the spatio-temporal point process models on each of the EarthquakeNPP datasets. Error bars of the mean and standard deviation are constructed for the NPPs using three repeat runs.

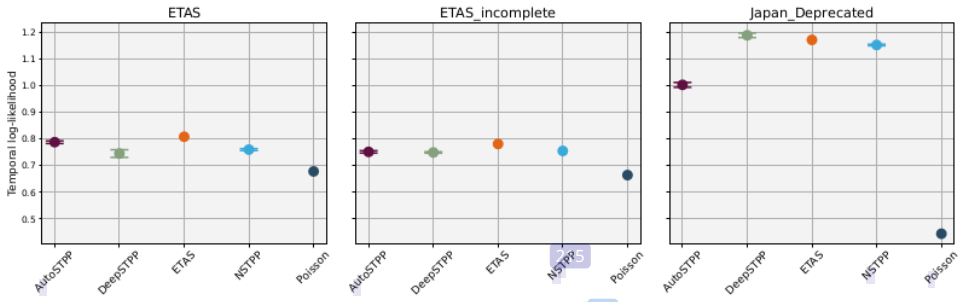


Figure 4.5: Test temporal log-likelihood scores for all the spatio-temporal point process models on each of the additional datasets. Error bars of the mean and standard deviation are constructed for the NPPs using three repeat runs.

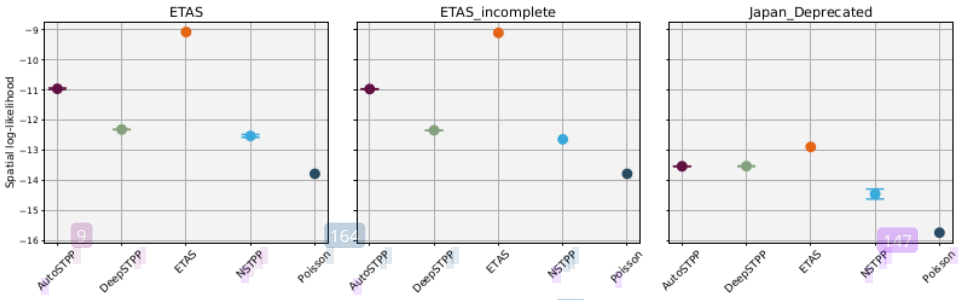


Figure 4.6: Test spatial log-likelihood scores for all the spatio-temporal point process models on each of the additional datasets. Error bars of the mean and standard deviation are constructed for the NPPs using three repeat runs.

simulations, F_N,

$$\delta_1 = \mathbb{P}(N \geq N_{\text{obs}}) = 1 - F_N(N_{\text{obs}} - 1) \tag{4.8}$$

$$\delta_2 = \mathbb{P}(N \leq N_{\text{obs}}) = F_N(N_{\text{obs}}). \tag{4.9}$$

4.5.2 Spatial Test

To evaluate the spatial component of the forecast, a test statistic aggregates the forecasted rates of earthquakes over a regular grid,

$$S = \left[\sum_{i=1}^{N} \log \hat{\lambda}(k_i) \right] N^{-1}, \tag{4.10}$$

where $\hat{\lambda}(k_i)$ is the approximate rate in the cell k where the i^{th} event is located. Upper and lower quantiles are estimated by comparing the observed statistic

$$S_{\text{obs}} = \left[\sum_{i=1}^{N_{\text{obs}}} \log \hat{\lambda}(k_i) \right] N_{\text{obs}}^{-1}, \tag{4.11}$$

with the empirical cumulative distribution of S using the repeat simulations, F_S

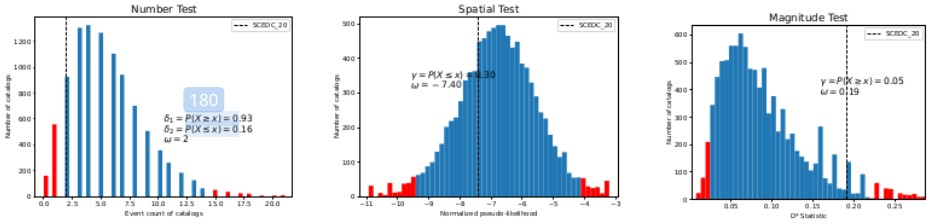
$$\gamma_s = \mathbb{P}(S \leq S_{\text{obs}}) = F_S(S_{\text{obs}}). \tag{4.12}$$

The grid is constructed from $\{0.1^\circ, 0.05^\circ, 0.01^\circ\}$ squares for `ComCat`, `SCEDC` and {`QTM_Salton_Sea`, `QTM_SanJac`, `White`} respectively.

4.5.3 Magnitude Test

To evaluate the earthquake magnitude component of the forecast, a test statistic compares the histogram of a forecast's magnitudes $\Lambda^{(m)}$, against the mean histogram over all forecasts $\bar{\Lambda}^{(m)}$,

$$D = \sum_k \left(\log \left[\bar{\Lambda}^{(m)}(k) + 1 \right] - \log \left[\Lambda^{(m)}(k) + 1 \right] \right)^2, \tag{4.13}$$



(a) (b) (c)

Figure 4.7: CSEP consistency tests on the ETAS model for the first day (01/01/2014) of the testing period in the SCEDC catalog. A total of 10,000 simulations are generated to compute empirical distributions of the test statistics for each of the three consistency tests: (a) Number test, (b) Spatial test, and (c) Magnitude test. The test fails if the observed statistic falls within the rejection region (red), defined by the 0.05 and 0.95 quantiles of the distribution.

where $\Lambda^{(m)}(k)$ and $\bar{\Lambda}^{(m)}(k)$ are the counts in the k^{th} bin of the forecast and mean histograms, normalised to have the same total counts as the observed catalog. Upper and lower quantiles are estimated by comparing the observed statistic

$$D_{\text{obs}} = \sum_k \left(\log\left[\bar{\Lambda}^{(m)}(k) + 1\right] - \log\left[\Lambda^{(m)}_{\text{obs}}(k) + 1\right] \right)^2, \tag{4.14}$$

with the empirical distribution of D using the repeat simulations, F_D

$$\gamma_m = \mathbb{P}(D \leq D_{\text{obs}}) = F_D(D_{\text{obs}}). \tag{4.15}$$

Histogram bins of size $\delta_m = 0.1$ are used across all datasets.

4.5.4 Evaluating Multiple Forecasting Periods

Savran et al. [176] describe how to assess a model's performance across the multiple days in the testing period. By construction, quantile scores over multiple periods should be uniformly distributed if the model is the data generator [55]. Therefore comparing quantile scores against standard uniform quantiles (y = x), highlights discrepancies between the observed data and the forecast. Additional statements can be made about over-prediction or under-prediction of each test statistic (quantile curves above/bellow y=x respectively). The Kolmogorov-Smirnov (KS) statistic then quantifies the degree of difference to the uniform distribution for each of the tests.

Further documentation of how to perform the CSEP evaluation procedure can be found on the platform, where we demonstrate the procedure for the ETAS model. Table 4.1 reports the benchmark performance scores taken from the quantile plots in Figure 4.8. The performance of ETAS is higher for the more typical higher magnitude catalogs such as ComCat and SCEDC, whereas it performs worse at the lower magnitude catalogs of QTM_SanJac, QTM_SaltonSea and

Dataset	Number Test		Spatial Test		Magnitude Test	
	Pass Rate	KS-Statistic	Pass Rate	KS-Statistic	Pass Rate	KS-Statistic
`ComCat`	62.3%	0.392	85.3%	0.128	75.3%	0.318
`SCEDC`	74.4%	0.161	87.5%	0.123	80.5%	0.153
`QTM_SanJac`	59.2%	0.461	96.7%	0.145	66.2%	0.406
`QTM_SaltonSea`	54.2%	0.441	82.1%	0.216	79.0%	0.311
`White`	17.1%	0.750	98.0%	0.373	25.0%	0.741

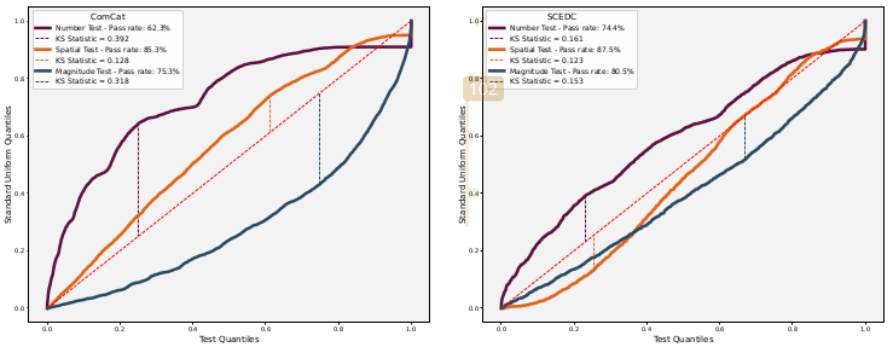
Table 4.1: CSEP consistency tests evaluate the calibration of all daily ETAS forecasts on EarthquakeNPP datasets. A test is performed at the $\alpha = 0.05$ significance level on each day in the testing period. The pass rate indicates the success of ETAS across all testing days. By construction quantile scores of the tests should be uniformly distributed if the model is the data generator. The KS-Statistic reports the difference of the quantile distribution to uniform, taken from the quantile plots in Figure 4.8.

`White`. Spatial prediction is consistently the best performing component of the ETAS forecast, whereas earthquake numbers are overpredicted by the model and earthquake magnitudes are generally not well predicted (Figure 4.8) . All results indicate significant room for improvement beyond the predictive performance of the ETAS model.

4.6 Discussion and Conclusion

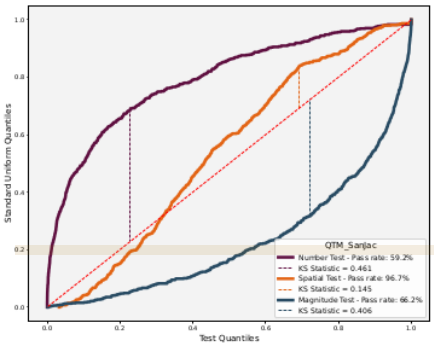
We introduce the EarthquakeNPP datasets to facilitate the benchmarking of NPPs against a community-endorsed ETAS model for earthquake forecasting. These datasets cover various regions of California, representing typical forecasting zones and the data commonly available to forecast issuers. Several datasets use modern methods of detection, which enables the inclusion of much smaller magnitude earthquakes.

In a benchmarking experiment, we compared three NPP models against ETAS and a baseline Poisson process. None of the NPP models outperformed ETAS, indicating that current NPP implementations are not yet suitable for operational earthquake forecasting. ETAS explicitly defines how larger earthquake magnitudes increase the likelihood of future earthquakes in both time and space, with an empirical relationship derived from observational studies [212, 214]. Since the NPPs lack any direct dependence on magnitudes, this is the likely cause for their performance relative to ETAS. Future implementations should exploit this additional feature for improved temporal and spatial performance. Encouragingly, the comparable temporal performance to ETAS without this additional feature suggests that incorporating magnitude dependence would enhance NPP performance beyond that of ETAS.

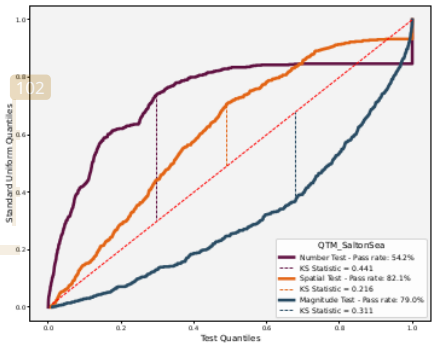


(a)

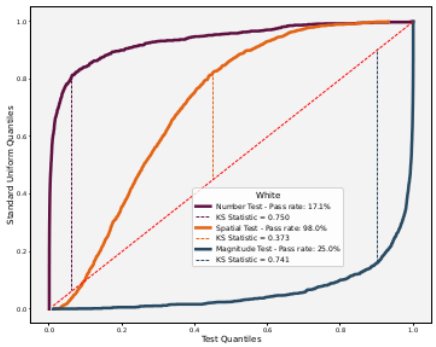
(b)



(c)



(d)



(e)

Figure 4.8: Quantile-quantile plots showing the calibration of all daily ETAS forecasts on a) `ComCat`,b) `SCEDC`, c) `QTM_San_Jac`, d) `QTM_Salton_Sea`, e) `White`. By construction quantile scores over multiple periods should be uniformly distributed if the model is the data generator. Comparing quantile scores against standard uniform quantiles ($y = x$), highlights discrepancies between the observed data and the forecast. Pass rates of each test are indicated in the legend. The Kolmogorov-Smirnov statistic, quantifies the degree of difference to the uniform distribution.

EarthquakeNPP supports the earthquake forecast evaluation procedure developed by the Collaboratory for the Study of Earthquake Predictability (CSEP). The procedure replicates how earthquakes forecasts are generated in an operational setting, requiring models to simulate many repeat event sequences over a day-long forecast horizon. Benchmark performance for the ETAS model enables future comparison of NPPs that are implemented for this procedure and enables their promotion to the fully prospective CSEP experiments. Notably, this procedure allows the evaluation of generative NPP models without explicit likelihoods [106, 235], by assessing their performance over the full trajectory of future events. This approach offers stakeholders a more comprehensive understanding of earthquake hazard than metrics focused on predicting the next event (e.g., Root Mean Square Error (RMSE)). The procedure also follows the recommendation by Shchur et al. [192] to move away from next-event point prediction for NPPs.

The EarthquakeNPP datasets, available at `https://github.com/ss15859/EarthquakeNPP`, provide a platform for future NPP developments to be benchmarked against these initial results. The platform is under ongoing development and in the future will see the direct comparison of emerging and other existing models models developed within the seismology community, as well as an expansion of datasets included to other seismically active global regions. Successful NPP models on these datasets, for both log-likelihood and CSEP metrics, will be directly impactful to stakeholders in seismology, ultimately enabling their integration into operational earthquake forecasting by government agencies.

Chapter 5

Conclusion

Short-term earthquake forecasting enables probabilistic hazard and risk assessments during ongoing aftershock sequences. Such forecasts are vital for informed decision-making to protect lives from subsequent destructive earthquakes following a mainshock. To date, the most successful forecasting models rely on statistical models known as point processes which represent earthquakes as points in time and space. The epidemic-type aftershock sequence (ETAS) model is the most successful of such models, describing the successive triggering of earthquakes with empirical laws derived from observational studies. Although useful, there are many known modeling inaccuracies and challenges for the ETAS model that motivate seismologists to continue to explore other options. Significant improvements to data collection methodologies in recent years have led to new earthquake catalogs containing a ten-fold increase in the number of recorded earthquakes, through including much lower magnitude earthquakes. This volume of new data offers potential improvement and challenges for current forecasting models as well as offering a chance to explore new modeling approaches.

In this thesis I apply methodologies from statistical machine learning to point process models of earthquakes. Machine learning is a highly data driven approach to statistical modeling which excels at learning complex non-linear relationships from large datasets. The recent availability of high-resolution earthquake catalogs is a testament to the successful application of machine learning techniques to continuous waveform data recorded by seismometers. However, the application of machine learning to model the point-like nature of earthquake occurrences in these catalogs is still in its infancy. Neural point processes (NPPs) are a recent development from the machine learning community which expand beyond the parametric approaches of classical point process modeling to encompass greater flexibility. NPPs span a broad spectrum of machine learning methodologies, including various sequence encoding techniques (Recurrent Neural Networks (RNN)[39], Long Short-Term Memory (LSTM) [201], Gated Recurrent Units (GRU) [31], Transformers [247]), density estimation methods

(normalising flows [190], continuous-time normalising flows [17], cumulative density networks [201]), function approximation approaches (neural spectral decomposition [38, 241]) and generative modeling techniques (diffusion models [235], score matching [106], variational autoencoder [240]). This thesis represents an initial exploration into the potential of NPPs for modeling seismicity, providing a foundation for further research in this rapidly evolving field. Given the diversity and ongoing advancements in NPP methodologies, this work offers valuable insights and direction for the future development of machine learning-based approaches to earthquake forecasting.

Forecasting the 2016–2017 Central Apennines Earthquake Sequence With a Neural Point Process

In Chapter 2, I extended an existing temporal NPP [157] to the magnitude domain and showed how this model can forecast earthquakes above a target magnitude threshold whilst being dependent on smaller magnitude earthquakes. This separation between target and dependent events is a crucial objective for all NPPs, as larger earthquakes pose greater hazards, making it critical to prioritize forecasting these significant events. This is a deviation from ETAS forecasting, which assumes self-similarity between small and large magnitude events through the independent GR law (1.2), where low magnitude threshold forecasts can be thinned to give higher magnitude forecasts (Section 2.3.2). Separating dependent and target events allowed me to forecast $M_w > 3$ events using new low magnitude data below the magnitude of completeness using an enhanced catalog from the Central Apennines earthquake sequence in Italy [206]. At such low magnitude thresholds ($\approx M_w 1$), the information gain over the neural model over the ETAS model is significant since ETAS is poorly specified for incomplete data. This provides a forecasting model which circumvents modeling choices often made by seismologists to deal with such completeness [67, 133]. However, the information gain from ETAS at each model's best performing magnitude threshold is marginal (Figure 2.7), indicating that further improvements are required for NPPs to significantly supersede the forecasting power of models such as ETAS.

The NPP model from chapter 2 used an LSTM encoding of the history of the past 20 events for the construction of the probability density for the next event. While it may seem intuitive that a longer event history would improve prediction, LSTM and other recurrent neural networks (RNNs, GRUs) suffer from vanishing gradients, limiting their effectiveness in capturing dependencies from distant past events. We found that there was no obvious improvement to extending the history beyond 20 (Figure A.4), in fact more instability was introduced. Transformers, on the other hand, have been shown to provide superior event encodings compared to recurrent architectures, owing to their longer memory capabilities [240, 247]. However, this advantage comes at the cost of significantly higher computational complexity ($\mathcal{O}(n^3)$) [220], restricting the effective sequence length to around 500 events [34].

For earthquake catalogs with very low completeness magnitudes, a truncated history limits the capability of a NPP model since important events in the history are masked by more recent small events. Any dependence on an event prior to the truncated event history is only captured implicitly through the most recent events. This enforces dependence constraints on the model, which limit expressivity. In contrast, the ETAS model has an infinite memory, defined through the summation over all past events. There are two potential solutions to this challenge. The first involves extending the sequence length of transformers by modifying attention mechanisms to reduce computational complexity. Techniques such as sparse attention [21], which enforces a sparse dependency structure on the past, or hierarchical attention [3], which encodes subsequences of history in blocks before feeding them into a higher-level sequence encoder, could allow for longer input sequences. Alternatively, some NPPs retain the Hawkes process formulation [38, 239–241], allowing a flexible triggering kernel to be applied across all historical events.

SB-ETAS: using simulation based inference for scalable, likelihood-free inference for the ETAS model of earthquake occurrences

Chapter 3 addresses improving Bayesian inference for existing earthquake forecasting models. Simulation based inference (SBI), though not directly a machine learning procedure, has seen significant improvements in recent years through the inclusion of machine learning approaches. By specifying a model through simulation rather than the likelihood, SBI broadens the scope of available models to encompass greater complexity. Since the NPP model of Chapter 2 is trained through density estimation of the next earthquake, I attempt to extend this model to approximate the likelihood of a 3 parameter Hawkes process by adding a dependence on simulator parameters. I find that this approach is unsuccessful, likely due to the truncated history of the model. I speculate that a longer history is necessary to successfully learn how the density of the next event depends on simulator parameters, whereas it is not necessary purely for next event prediction. Since only one NPP model was tested, I cannot rule out the success of other models, particularly those proposed by Zhu et al. [241], Zhou et al. [240], Dong et al. [38], and Zhou and Yu [239], which can incorporate a full event history, might perform better in this context.

I opt instead to use summary statistics which I extend from previous works Ertekin et al. [42] and Deutsch and Ross [33], to perform Bayesian inference for a temporal ETAS model. This approach provided better coverage of ETAS posterior distributions than a competitor approach `inlabru` [184], and the efficient computation of these summary statistics allowed the procedure to scale at $\mathcal{O}(n \log n)$ with the number of events in the earthquake catalog, improving upon the $\mathcal{O}(n^2)$ scaling of prior methods. Bayesian inference is widely used in operational earthquake forecasting, especially to accommodate the substantial variability in ETAS parameters

observed between different aftershock sequences [159, 217]. The methodology presented in Chapter 3 would allow practitioners to continue employing a Bayesian approach for forecasting with much larger earthquake catalogs and would allow for parameter inference for models without a tractable likelihood. Chapter 2 highlighted the increased importance of data completeness for low magnitude catalogs, emphasizing that any model using such data must account for this artifact. While some ETAS approaches consider data missingness [67, 133], none fully address the potential triggering from unobserved events. The intractable likelihood from such a model lends itself to a simulation based approach and the scalability further lends itself to inference on such large low magnitude catalogs. Further improvements could be made by refining the hand-tuned summary statistics used in this study, particularly regarding which intervals of the Ripley K function to sample from and the number of samples to use. "Learning" optimal summary statistics is a common practice in simulation-based inference [26, 166] and could be applied to optimize such hyperparameters during the inference process.

A far more complex forecasting model combines the third Uniform California Earthquake Rupture Forecast (UCERF3), a fault based long-term model, with the short-term earthquake clustering of ETAS. UCERF3-ETAS [46], is run prospectively in California following large magnitude earthquakes with forecasts generated by simulating many synthetic catalogs. The model is defined purely as a simulator model and possesses no tractable likelihood function. ETAS clustering parameters for the model are taken from independent parameter estimates [72], without any formal estimation using the combined model. Whilst the use of these parameters has proven generally effective in validation studies [131, 158, 176], combining ETAS with the fault model breaks the dependence structure of parameters from the independent estimates since UCERF3-ETAS uses both a variable magnitude distribution (incorporating elastic rebound) as well as a different background rate that includes the fault model. Simulation based inference could present a framework for both calibrating model parameters as well providing a way to update parameters during ongoing aftershock sequences. In fact Page and van der Elst [158] construct a set of "Turing tests" which compare the consistency of URCERF3-ETAS simulations with observed data. These tests, although used here for model validation, are in essence summary statistics and could be repurposed for model calibration along with those I consider in Chapter 3. Following this validation study, UCERF3-ETAS was later updated [47] to account for the inter-sequence variability of aftershock productivity using a hand tuned distribution. Simulation based inference could offer a more formal construction on the distribution of the productivity parameter in a Bayesian manner similar to Page et al. [159]. Despite the additional computational demands due to the model's complexity, various SBI methods designed to handle expensive simulators [26] would make this approach more feasible.

EarthquakeNPP: Benchmark Datasets for Earthquake Forecasting with Neural Point Processes

Finally, Chapter 4 directs the future application of NPPs to earthquake forecasting. Motivated by the poor construction of an earthquake benchmarking experiment for NPPs circulating machine learning literature [17, 235, 239, 240], I preprocess and present several earthquake datasets from California alongside a benchmark implementation of the ETAS model in a platform named EarthquakeNPP. The relevant datasets, benchmark model and evaluation procedure ensure that successful NPP models will have direct impact to earthquake forecasting. In a benchmarking experiment I find that none of 3 NPPs - with implementations taken directly from their original studies [17, 239, 240] - outperform the ETAS model. This is not entirely surprising, as these implementations do not use earthquake magnitudes as a feature. However, the comparable temporal log-likelihood performance against ETAS for Deep-STPP [240] across all but one of the datasets indicates the strong potential of future modifications to this model. There is a greater difference between ETAS and NPPs for spatial log-likelihood scores, likely due to the absence of magnitude information in the NPPs. For the ETAS model, earthquake magnitudes affect future rates (temporally) through the productivity scaling relation 1.15, with larger magnitudes triggering on average more events. An NPP without any magnitude information could implicitly learn that a larger earthquake has occurred through an observed increase in rate and adapt accordingly [78]. On the other hand, earthquake magnitude affects spatial triggering by larger earthquakes triggering events at greater distances [153]. Without magnitudes, NPP models lack the knowledge of the centroid location of a spatial aftershock kernel and therefore must rely on background or secondary triggering from the large event to infer spatial locations. These benchmark results provide important motivation for those working on NPPs within the machine learning community, and the need to include magnitudes provides a clear suggestion for model improvement.

The development of EarthquakeNPP was motivated by a clear lack of understanding of seismology best practices and objectives within the machine learning community. While many earthquake catalogs are publicly accessible, they typically include events below the magnitude of completeness, a challenge not readily apparent to those without seismology experience. This led to Chen et al. [17] using incomplete data in their benchmarking experiment, which combined with the omission of a key earthquake sequence and improper training-testing splits, rendered the experiment irrelevant to stakeholders in seismology. Since it is unreasonable to expect all data centers to guide users on constructing accurate forecasting experiments, EarthquakeNPP aims to bridge this gap by offering clear instructions and fostering better communication between machine learning researchers and seismologists.

The construction of the platform also highlighted some of the challenges of evaluating classical

and neural point process models. Model evaluation in this thesis has primarily relied on computing the log-likelihood of events on held-out test data, either averaged over all test events or for each next future event. Since point processes provide distributions over future events, the log-likelihood metric is effective for comparing models due to its nature as a strictly proper scoring rule [56]. However, its utility is limited to model comparison and lacks direct interpretability. Metrics like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) offer more interpretability by measuring error in time, space, or magnitude units, but they only assess point predictions and disregard the probabilistic information that is central to these models. These metrics are also considered to be flawed and misleading for earthquake prediction [86], since the next-event distribution is strongly skewed and heavy-tailed [25] and therefore far from Gaussian.

Consistency metrics and the evaluation procedure developed by the Collaboratory for the Study of Earthquake Predictability (CSEP) offer both interpretability and probabilistic evaluation. This evaluation procedure is common within the seismological community, however has yet to be applied to any NPP. This is essential for future NPP development for several reasons: By generating forecasts through simulating many repeat event sequences over a forecast horizon, the procedure directly evaluates models in the manner in which they would be used in an operational setting. Additionally, the evaluation procedure enables the comparison of models which are defined purely generatively, a more recent direction of NPPs [106, 235] and machine learning in general. CSEPs consistency tests also move beyond next event prediction, evaluating the empirical distribution over entire event sequences, a far more informative quantity for estimating hazard during destructive aftershocks sequences. Adopting this evaluation framework is crucial for making NPPs operationally viable, as strong performance in prospective CSEP experiments will build confidence in these models. EarthquakeNPP facilitates this by incorporating CSEP metrics and advocating for their use in future NPP research.

5.1 Future Work and Final Comments

While extensive future testing of NPP models (and any model) is essential to build trust before they can be deployed operationally, it is equally important to enhance model interpretability. Next-event probability distributions or intensity functions are not inherently interpretable, which makes models based on them [17, 31, 39, 91, 107, 157, 201, 235] more "black-box" in nature. However, Hawkes process-based NPPs offer more interpretable features, such as background rates, triggering kernels (separable or otherwise) [38, 239, 241], or "aftershock" parameters [240], while still benefiting from the flexibility of deep learning models. Additionally, these models allow for parts of the process to be constrained — for example,

enabling flexible spatial triggering kernels while preserving the temporal Omori decay structure. However, moving beyond the purely excitatory nature typical of Hawkes process models requires additional training steps (see Section 4.1 of Dong et al. [38]) to incorporate inhibition.

While the most effective short-term earthquake forecasting models rely primarily on earthquake catalog data, various geophysical features—such as tectonic settings, source event characteristics, and regional crustal properties—have shown significant correlations with observed seismicity in aftershock sequences [30, 73, 209]. Traditional forecasting models require a clearly defined statistical relationship between these physical features and earthquake rates. However, given the assumed non-linear nature of these relationships, machine learning and non-parametric models have been explored to predict aftershock productivity and location in preliminary studies [30, 73, 209]. Because the ETAS model does not have a natural mechanism to integrate such features, NPPs offer a promising alternative, providing the flexibility to incorporate both geophysical features and earthquake catalog data. Future NPP development should leverage insights from these studies to integrate relevant geophysical features, not only to enhance earthquake predictability but also to advance our understanding of the underlying physical processes by identifying the most informative features.

Efforts to move beyond the point-like representation of earthquakes have included incorporating the rupture geometry of large events into point process models like ETAS [6, 62, 63]. By modeling large earthquakes as surfaces in 2D or 3D space while retaining the point-like representation for the event time, rupture geometry ETAS models can more accurately capture the anisotropic distribution of observed aftershocks [62]. This concept could be extended to NPP models constructed as Hawkes processes [38, 239–241], applying the same same fault geometry representation as Bach and Hainzl [6], Guo et al. [62] or Guo et al. [63], but using a more flexible NPP spatial kernel instead of the traditional stationary isotropic ETAS kernels. However, incorporating a finite fault representation of earthquakes into NPPs that model intensity functions [39, 157], probability densities over next events [17, 190] or are defined generatively [106, 235] remains an open challenge for future research.

Earthquake swarms, characterized by a sudden and seemingly spontaneous increase in seismic activity, pose a significant challenge for forecasting due to their deviation from the typical mainshock-aftershock clustering pattern. These swarms are believed to be driven by transient external processes such as fluid migration or aseismic creep [68, 84, 111, 114] and are typically modeled using ETAS frameworks with non-stationary background rates [100, 110] or varying productivity parameters [100]. The unpredictability and inherent non-stationarity of earthquake swarms present a promising opportunity for applying NPPs, which can potentially capture complex, time-varying patterns beyond the capabilities of classical models. Future

research should focus on isolating earthquake swarms as a distinct phenomenon and rigorously evaluating the performance of NPPs in this context.

Point process models can exploit the predictable properties of earthquake clustering to provide critical hazard information during destructive aftershock sequences. In this thesis, I have aimed to illustrate some ways in which statistical machine learning can enhance classical point process modeling of earthquakes. This work represents an initial exploration on use of neural point processes in earthquake forecasting and is far from exhaustive. The successful advancement of these models necessitates a deep, interdisciplinary understanding of both statistical machine learning and seismology. It is my hope that this work bridges the gap between these fields, offering valuable insights to researchers in both domains.

Appendix A

Appendix to Chapter 2

Contents of this file

1. Figures A.1 to A.3

Introduction

In this appendix, we share the learnt ETAS parameters for the Amatrice-Visso-Norcia catalog for all the values of Mcut and all training testing partitions (Figure A.1). Furthermore we show the cummulative information gain (CIG) of the neural model over ETAS on the complete synthetic catalog for both time and magnitude forecasting (Figure A.2), as well as for the incomplete synthetic catalog (Figure A.3). Finally, we show validation over the choice of history truncation parameter 'time_step' (Figure A.4)

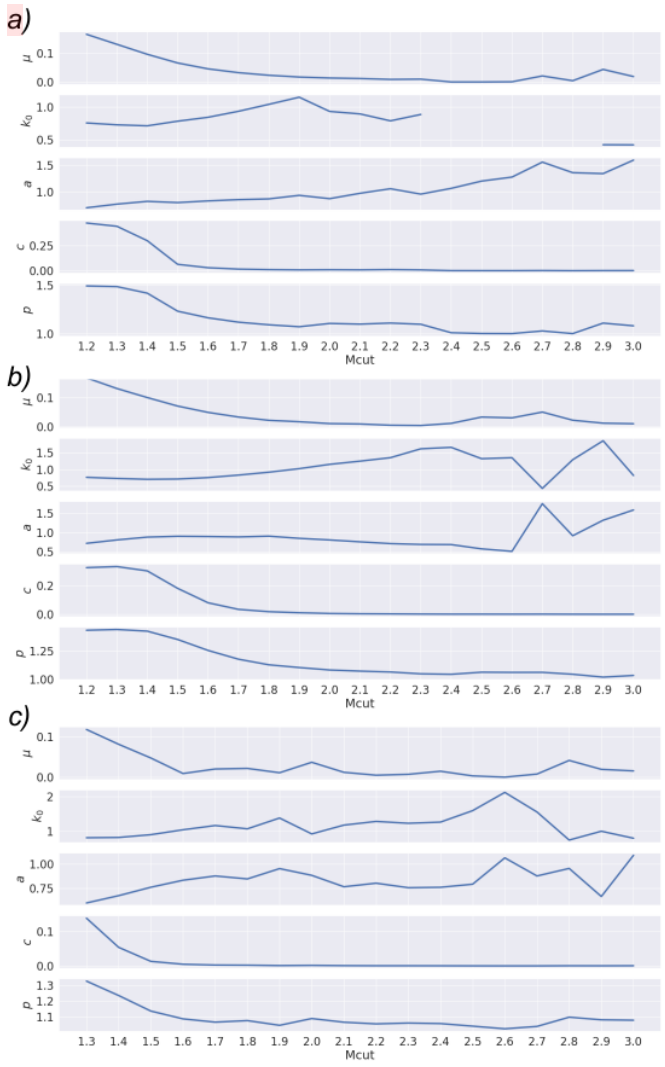


Figure A.1: Fitted ETAS parameters as a function of Mcut. **a)** training up to the Visso earthquake. k_0 parameters for Mcut 2.4-2.8 have been removed from the plot to aid in visualisation. These parameter values are orders of magnitude larger. **b)** training up to the Norcia earthquake. **c)** training up to the Campotosto earthquakes. The unit of time is hours.

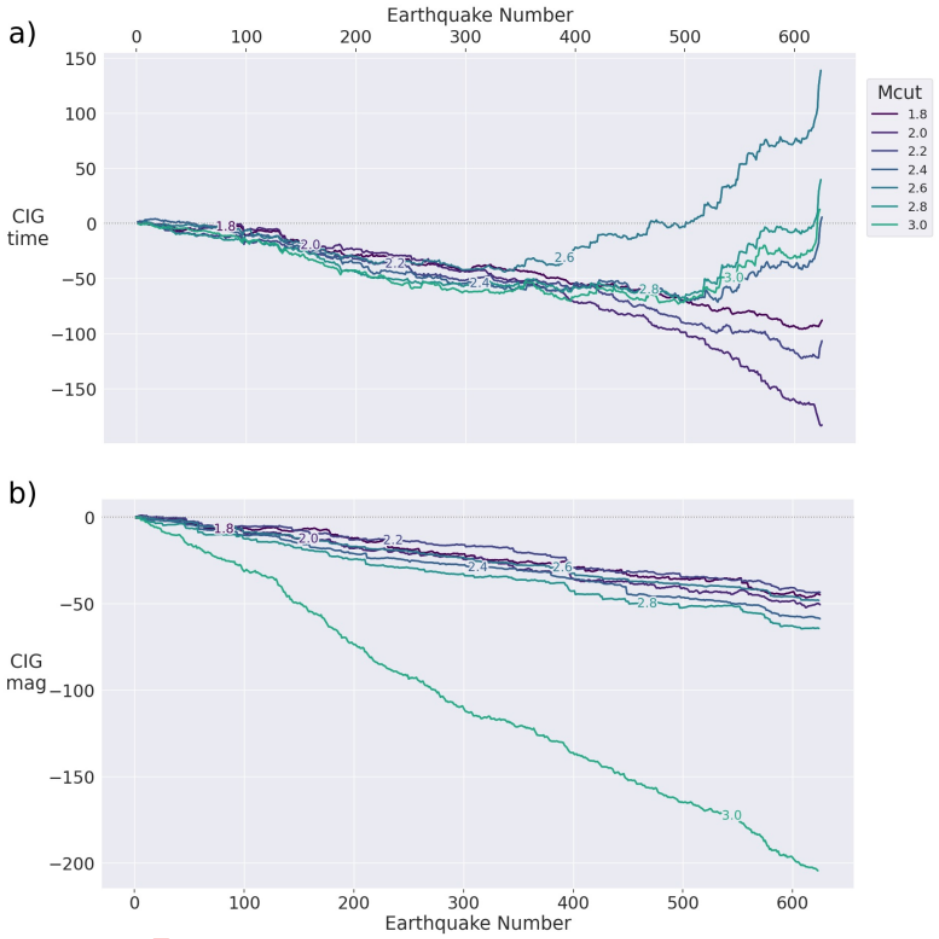


Figure A.2: a) - b) The Cumulative Information Gain (CIG) of the neural model over ETAS for a range of values of Mcut. The models are trained and forecasted on the complete synthetic catalog and the plot depicts the evolution of the CIG from the beginning of the testing period to the end of the catalog. a) displays the CIG for event-time forecasts, b) displays the CIG for magnitude forecasts.

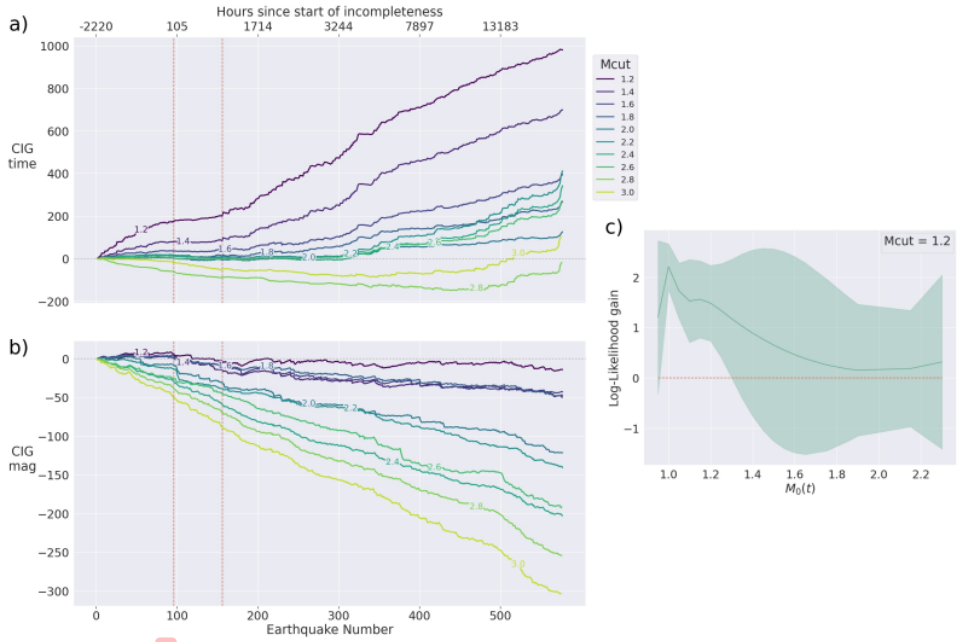


Figure A.3: a) - b) The Cumulative Information Gain (CIG) of the neural model over ETAS for a range of values of Mcut. The models are trained and forecasted on the incomplete synthetic catalog and the plot depicts the evolution of the CIG from the beginning of the testing period to the end of the catalog. The curve is plotted per event, however, the time since the start of a period of incompleteness is displayed on the top axis. a) displays the CIG for event-time forecasts, b) displays the CIG for magnitude forecasts. c) displays the information gain of the neural model over ETAS as a function of the completeness of the testing catalog - both models are trained with Mcut = 1.2.

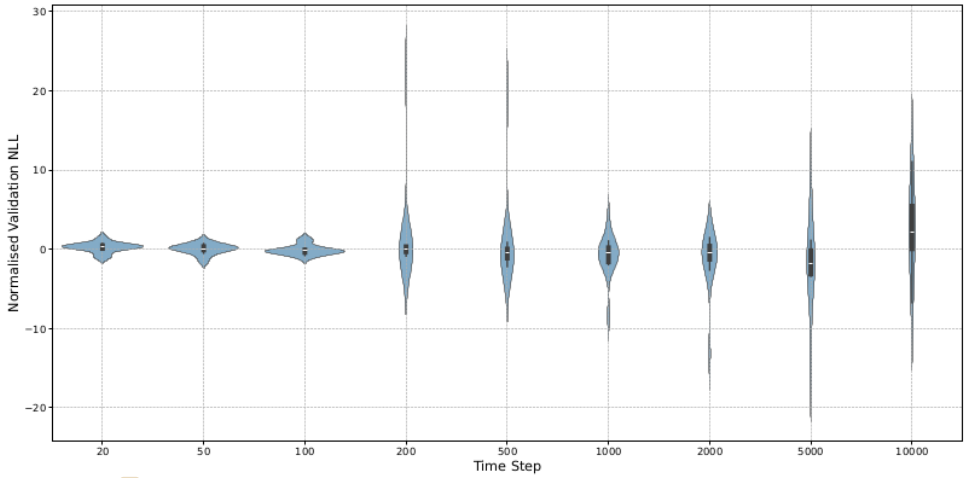
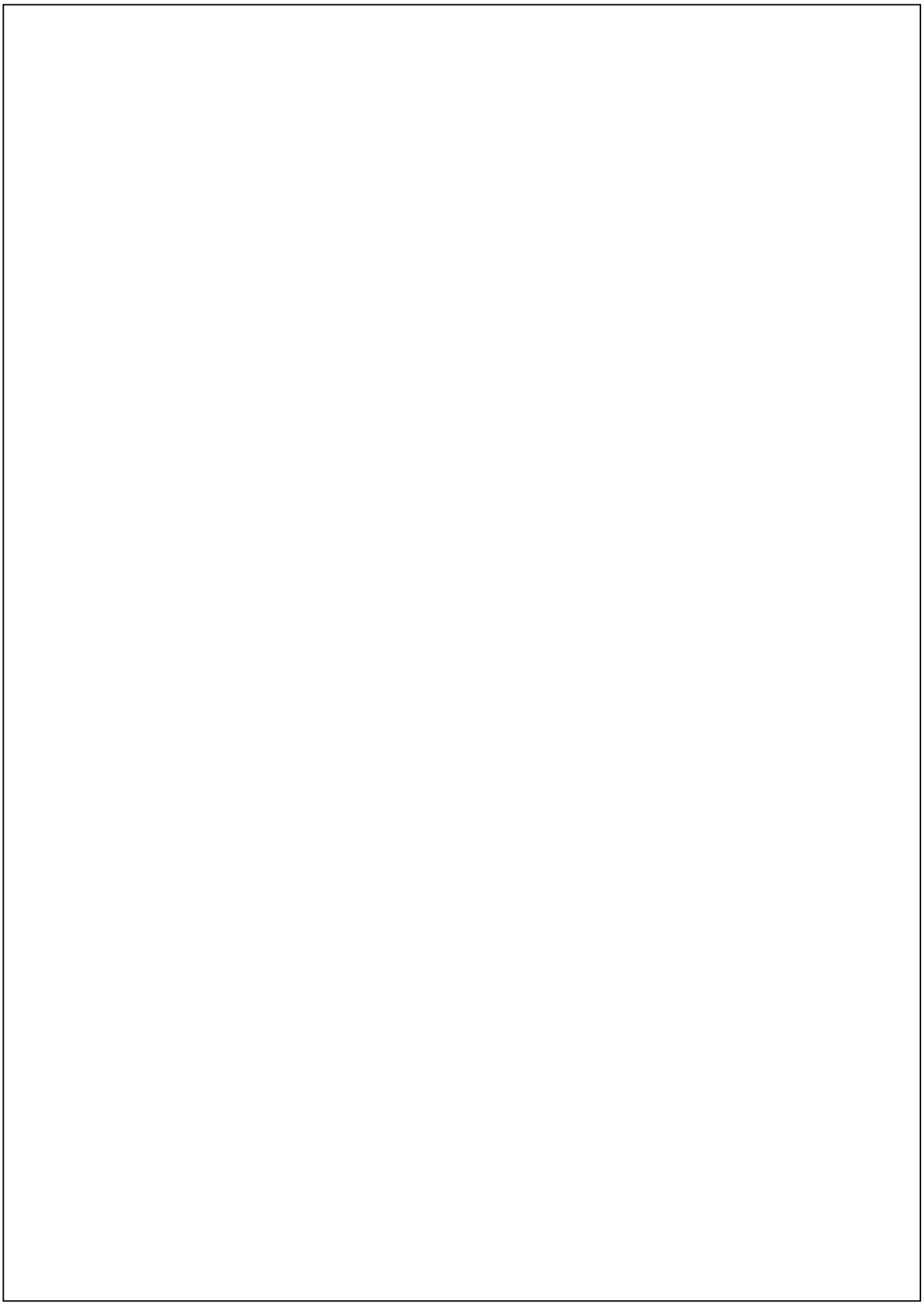


Figure A.4: Negative Log-Likelihood (NLL) scores on the validation dataset during the training of the neural model up to the Norcia earthquake. The validation set comprises 20% of randomly sampled points from the training data. The Mcut values range from 1.2 to 3.0, and the NLL for each Mcut is normalized to have a mean of zero. Normalised validation NLL is plotted as a function of 'Time Step' (also referred to as parameter d in the main text), representing the length of event history truncated for input into the neural model.



Appendix B

Appendix to Chapter 3

B.1 MMD vs. Time Step

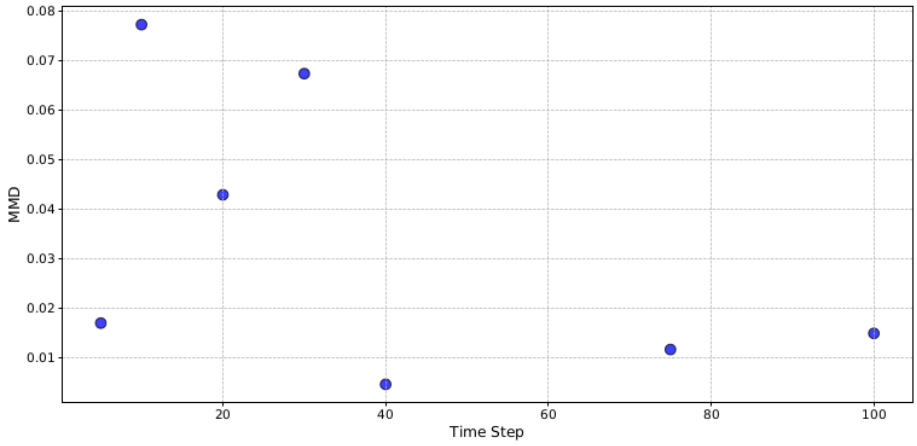


Figure B.1: The Maximum Mean Discrepancy (MMD) between samples using Neural Likelihood Estimation and MCMC using the likelihood function. Posteriors are estimated using for univariate Hawkes process with exponential kernel generated with parameters $(\mu, k, \upsilon) = (0.2, 0.5, 0.5)$, using a Uniform$([0.05, 0, 0], [0.85, 0.9, 3])$ prior. The MMD is plotted as a function of 'Time Step', which is the length of history that is truncated in the sequence encoding of the Neural Likelihood estimator.

B.2 Posterior Distributions

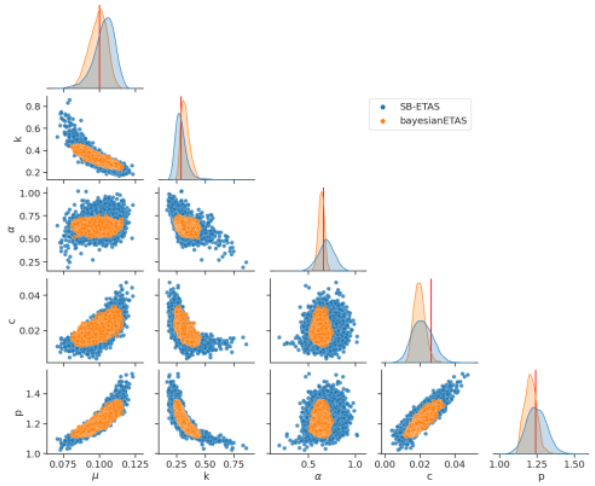


Figure B.2: Samples from the posterior distribution of ETAS parameters for the synthetic Ridgecrest catalog (5528 events, $M_0 = 2.0$), using `bayesianETAS` and SB-ETAS. The data generating parameters are marked in red in the diagonal plots.

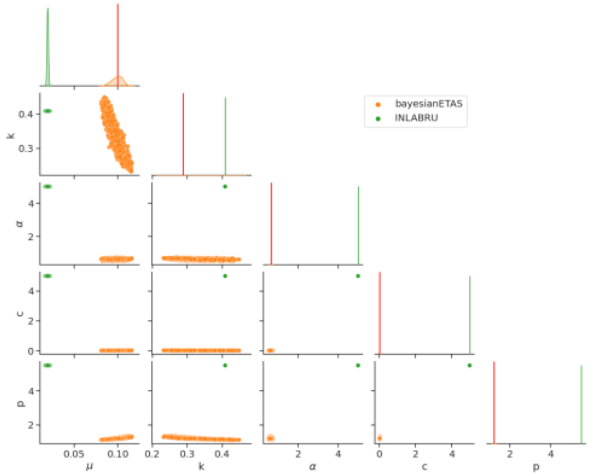


Figure B.3: Samples from the posterior distribution of ETAS parameters for the synthetic Ridgecrest catalog (5528 events, $M_0 = 2.0$), using `bayesianETAS` and `inlabru`. The data generating parameters are marked in red in the diagonal plots.

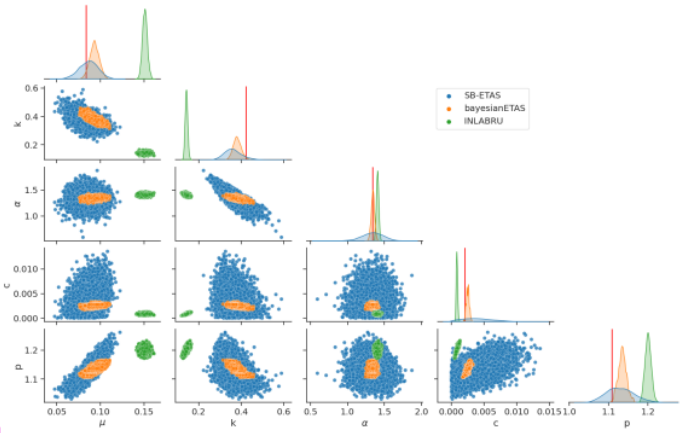


Figure B.4: Samples from the posterior distribution of ETAS parameters for the synthetic Amatrice catalog (6673 events, $M_0 = 3.0$), using `bayesianETAS`, `inlabru` and SB-ETAS. The data generating parameters are marked in red in the diagonal plots.

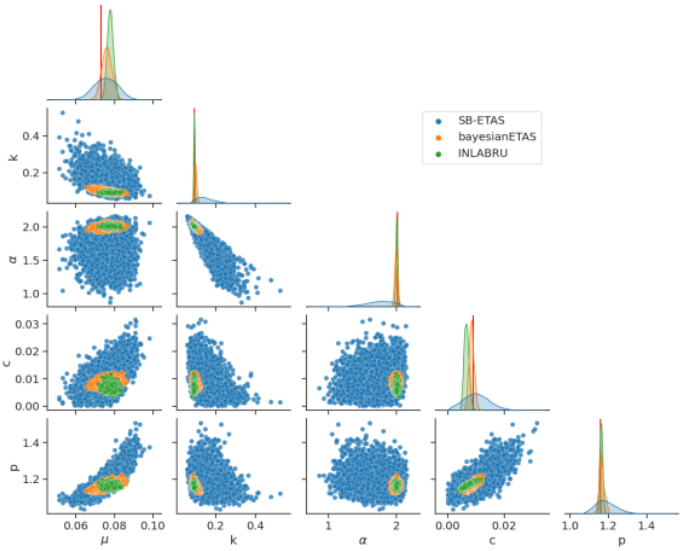


Figure B.5: Samples from the posterior distribution of ETAS parameters for the synthetic Kumamoto catalog (5340 events, $M_0 = 3.5$), using `bayesianETAS`, `inlabru` and SB-ETAS. The data generating parameters are marked in red in the diagonal plots.

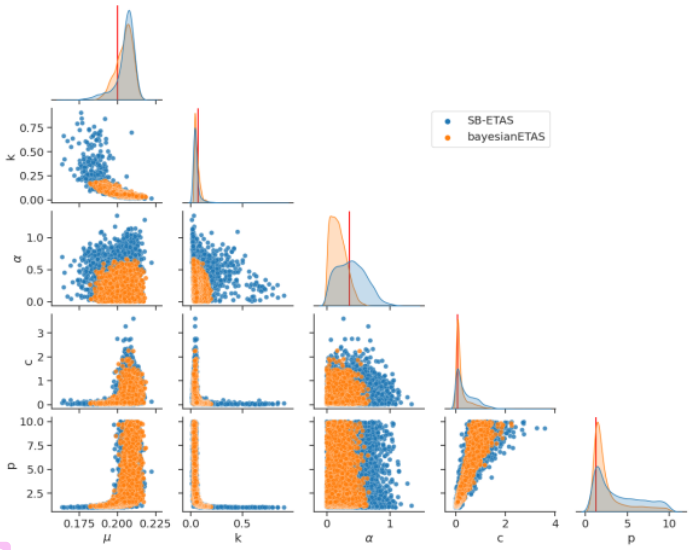


Figure B.6: Samples from the posterior distribution of ETAS parameters for the synthetic Landers catalog (6538 events, $M_0 = 2.0$), using `bayesianETAS` and SB-ETAS. The data generating parameters are marked in red in the diagonal plots.

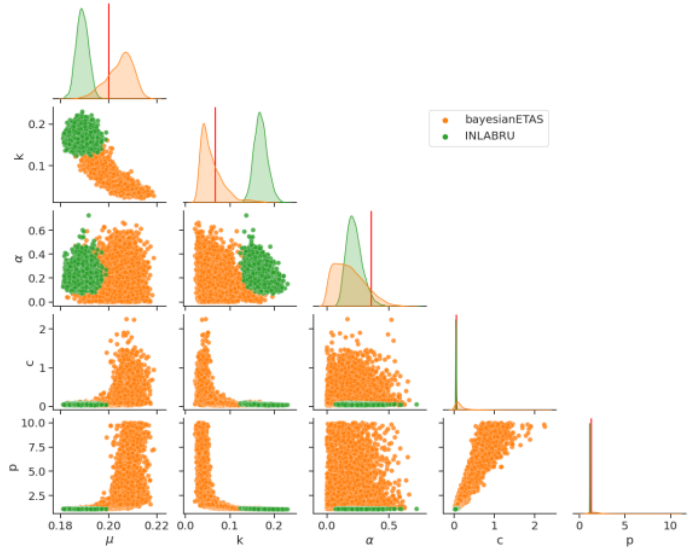


Figure B.7: Samples from the posterior distribution of ETAS parameters for the synthetic Landers catalog (6538 events, $M_0 = 2.0$), using `bayesianETAS` and `inlabru`. The data generating parameters are marked in red in the diagonal plots.

B.3 Memory

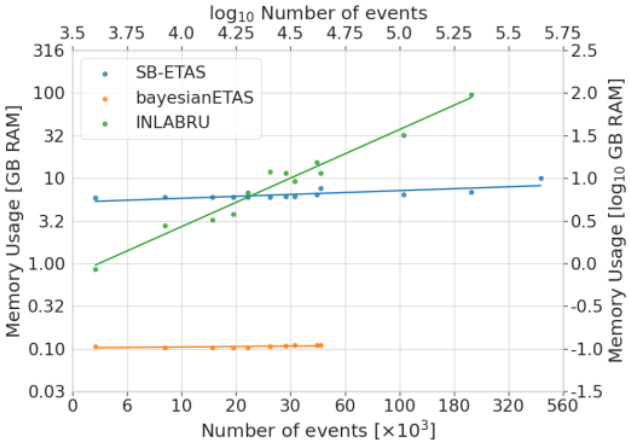


Figure B.8: The memory usage for parameter inference versus the catalog size for SB-ETAS, `inlabru` and `bayesianETAS`. Separate ETAS catalogs were generated with the same intensity function parameters but for varying size time-windows. The memory usage and the number of events are plotted in log-log space.

Appendix C

Appendix to Chapter 4

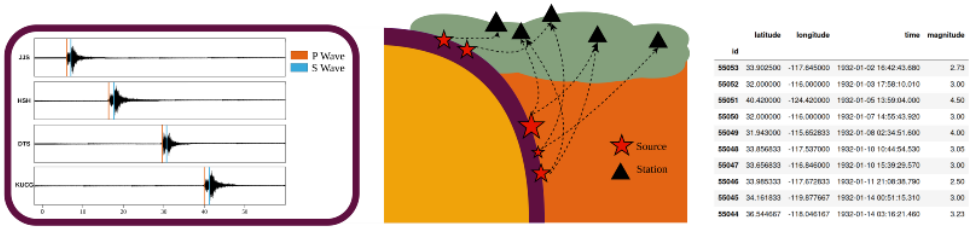
C.1 Earthquake Catalog Data

C.1.1 Earthquake Catalog Generation

Data missingness, referred to in seismology as catalog (in)completeness, is the primary challenge faced with earthquake catalogs. It is an important and unavoidable feature, and is a result of how earthquakes are detected and characterised. Below, we briefly overview the process of generating an earthquake catalog to illustrate the data quality issues. In the subsequent section, we review catalog incompleteness and its potential impact on the performance and evaluation of forecast models.

Seismometers and Seismic Networks. A seismometer is an instrument that detects and records the vibrations caused by seismic waves [193, 199]. It consists of a sensor to detect ground motion and a recording system to log three-dimensional ground motion over time, typically vertical and horizontal velocities. Seismic networks, comprising multiple seismometers, monitor seismic activity at regional, national or global scales (see, e.g., [227] and references therein). High-density networks with modern, sensitive equipment provide more detailed and accurate data, enhancing the ability to detect and analyse smaller and more distant earthquakes.

From Waveforms to Phase Picking. The process of converting raw continuous seismic waveforms into useful earthquake data begins with phase picking, which identifies the arrival times of the primary (P) and secondary (S) waves of an earthquake. Historically, this was done manually, but now automated algorithms, such as the STA/LTA algorithm, detect wave arrivals by analyzing signal amplitude changes [4]. Recent algorithms, such as machine learning classifiers [e.g. 102, 242] and template-matching [e.g. 175], can process much higher volumes of



id	latitude	longitude	time	magnitude
59053	33.902500	-117.645000	1992-01-02 16:42:43.680	2.73
59052	32.000000	-116.000000	1992-01-03 17:58:10.010	3.00
59051	40.420000	-124.420000	1992-01-05 13:59:04.000	4.50
59050	32.000000	-116.000000	1992-01-07 14:55:43.920	3.00
59049	31.943000	-115.652833	1992-01-08 02:34:51.500	4.00
59048	33.856833	-117.537000	1992-01-10 10:44:54.530	3.05
59047	33.656833	-116.846000	1992-01-10 15:39:29.570	3.00
59046	33.985333	-117.672833	1992-01-11 21:08:38.790	2.50
59045	34.161833	-119.877667	1992-01-14 00:51:15.310	3.00
59044	36.544667	-118.046167	1992-01-14 03:16:21.460	3.23

Figure C.1: Generating an earthquake catalog involves several key steps: seismic phase picking, magnitude estimation, and the association and location of seismic sources. This process transforms raw waveform data recorded at seismic stations to locations, times, and magnitudes of earthquakes.

data efficiently and are often able to detect events of much smaller magnitudes.

Earthquake Association and Location After phase picking, the next step is to associate phases from different seismometers with the same earthquake. Simple algorithms require at least four phase arrivals to be detected on different stations within a short time interval to declare an event. Once phases are associated, location estimation determines the earthquake's hypocenter and origin time by minimizing travel-time residuals using linearized or global inversion algorithms [112, 208]. Given the potential for misidentified or mis-associated phase arrivals due to low signal-to-noise of small events or the near-simultaneous occurrence during very active aftershock sequences, an automated system typically first picks arrival times and determines a preliminary location, which is subsequently reviewed by a seismologist [e.g. 227, and references therein]. Locations are typically reported as the geographical coordinates and depths where earthquakes first nucleated (hypocenters), although some catalogs report the centroid location, a central measure of the extended earthquake rupture.

Earthquake Magnitude Calculation The magnitude of an earthquake quantifies the energy released at the source and was originally defined in the seminal paper by [172]. The original definition, now referred to as the local magnitude (ML), is calculated from the logarithm of the amplitude of waves recorded by seismometers. This scale, however, "saturates" at higher magnitudes, meaning it underestimates magnitudes for various reasons. This led to introduction of the moment magnitude scale (Mw) [71], which computes the magnitude based on the estimated seismic moment M_0, which can be related to the physical rupture process via

$$M_0 = \text{rigidity} \times \text{rupture area} \times \text{slip}, \tag{C.1}$$

where rigidity is a mechanical property of the rock along the fault, rupture area is the area of the fault that slipped, and slip is the distance the fault moved. Mw is determined seismologically via a spectral fitting process to the earthquake waveforms. In practice, it can be challenging to use a single magnitude scale for a broad range of magnitudes, therefore a range

of scales may be present within a single catalog, and approximate magnitude conversion equations may be used to homogenize the scales [e.g. 83, and references therein].

C.1.2 Earthquake Catalog Completeness

All of the EarthquakeNPP datasets are made publicly available by their respective data centers in raw format. However, constructing a suitable retrospective forecasting experiment from this raw data requires appropriate pre-processing. This typically involves truncating the dataset above a magnitude threshold M_{cut} and within a target spatial region to address incomplete data, known as catalog completeness M_c [e.g., 129, 130].

There are several reasons why an earthquake may not be detected by a seismic network. Small events may be indistinguishable from noise at a single station, or insufficiently corroborated across multiple stations. Another significant cause of missing events occurs during the aftershock sequence of large earthquakes, when the seismicty rate is high [67, 96]. Human or algorithmic detection abilities are hampered when numerous events occur in quick succession, e.g. when phase arrivals of different events overlap at different stations or the amplitudes of small events are swamped by those of large events. Since catalog incompleteness increases for lower magnitude events, typically the task is to find the value M_c above which there is approximately 100% detection probability. Choosing a truncation threshold M_{cut} that is too high removes usable data. Where NPPs have demonstrated an ability to perform well with incomplete data [201], typically a threshold below the completeness biases classical models such as ETAS [183]. Seismologists often investigate the biases of different magnitude thresholds by performing repeat forecasting experiments for different thresholds [e.g. 120, 201], which we also facilitate in our datasets.

Typically M_c is determined by comparing the raw earthquake catalog to the Gutenberg-Richter law [64], which states that the distribution of earthquake magnitudes follows an exponential probability density function

$$f_{GR}(m) = \beta e^{\beta(m-M_c)} \quad : m \geq M_c. \tag{C.2}$$

where β is a rate parameter related to the b-value by $\beta = b \log 10$. Histogram-based approaches, such as the simple Maximum Curvature method [226] as well as many others [e.g. 83, and references therein], identify the magnitude at which the observed catalog deviates from this law, indicating incompleteness (See Figure C.2b).

In practice, catalog completeness varies in both time and space $M_c(t, \mathbf{x})$ [e.g. 179]. During aftershock sequences, $M_c(t)$ can be very high [e.g., 2, 66] (See Figure C.2a). Thresholding at the maximum value might remove too much data. Instead, modelers either omit particularly

incomplete periods during training and testing [69, 94], model the incompleteness itself [65–67, 79, 133, 154, 224], or accept known biases from disregarding this issue [196]. Spatially, catalogs are less complete farther from the seismic network [129], so the spatial region can be constrained to remove outer, more incomplete areas (See Figure C.2c).

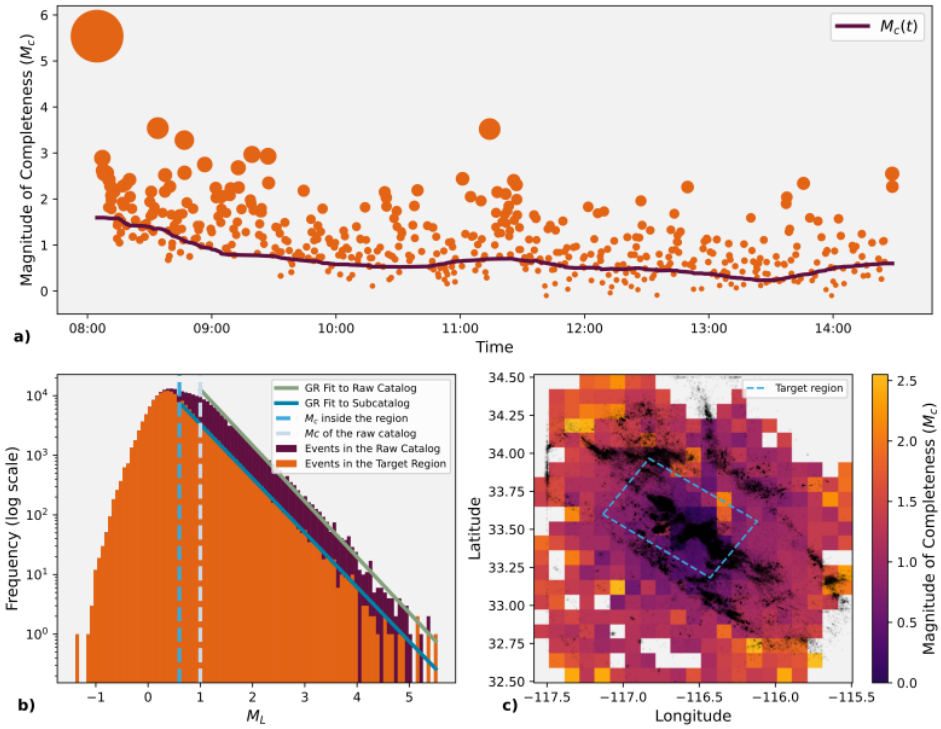


Figure C.2: **a)** the June 10, 2016 Mw5.2 Borrego Springs earthquake and aftershocks, which occurred on the San Jacinto fault zone and is recorded in the WHITE catalog. An estimate of the magnitude of completeness $M_c(t)$ over time using the Maximum Curvature method reveals more incompleteness immediately following the large earthquake. **b)** magnitude-frequency histograms reveal that truncating the raw WHITE catalog to inside the target region decreases M_c. Each histogram is fit to the Gutenberg-Richter (GR) law and an estimate of M_c for each catalog occurs where the histogram deviates from the (GR) line. **c)** An estimate of M_c for gridded regions of the San Jacinto fault zone, using the raw WHITE catalog.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng.
TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.
URL `http://tensorflow.org/`.
Software available from tensorflow.org.

[2] Duncan Carr Agnew.
Equalized plot scales for exploring seismicity data.
Seismological Research Letters, 86(5):1412–1423, 2015.

[3] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang.
Etc: Encoding long and structured inputs in transformers.
arXiv preprint arXiv:2004.08483, 2020.

[4] Rex Allen.
Automatic phase pickers: Their present use and future prospects.
Bulletin of the Seismological Society of America, 72(6B):S225–S242, 1982.

[5] Alessandro Amato and F Mele.
Performance of the ingv national seismic network from 1997 to 2007.
Annals of Geophysics, 2008.

[6] Christoph Bach and Sebastian Hainzl.
Improving empirical aftershock modeling based on additional source information.
Journal of Geophysical Research: Solid Earth, 117(B4), 2012.

[7] Emmanuel Bacry and Jean-Francois Muzy.
Second order statistics characterization of hawkes processes and non-parametric estimation.
arXiv preprint arXiv:1401.0903, 2014.

[8] Emmanuel Bacry, Thibault Jaisson, and Jean-François Muzy.
Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics.
Quantitative Finance, 16(8):1179–1201, 2016.

[9] Mark A Beaumont, Wenyang Zhang, and David J Balding.
Approximate bayesian computation in population genetics.
Genetics, 162(4):2025–2035, 2002.

[10] Mark Bebbington and David Harte.
The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications.
Geophysical Journal International, 154(3):925–946, 2003.

[11] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann.
Neural flows: Efficient alternative to neural odes.
Advances in Neural Information Processing Systems, 34:21325–21337, 2021.

[12] David D Bowman and Charles G Sammis.
Intermittent criticality and the Gutenberg-Richter distribution.
Computational Earthquake Science Part I, pages 1945–1956, 2004.

[13] Emily E Brodsky.
The spatial density of foreshocks.
Geophysical Research Letters, 38(10), 2011.

[14] Emily E Brodsky and Nicholas J van der Elst.
The uses of dynamic earthquake triggering.
Annual Review of Earth and Planetary Sciences, 42:317–339, 2014.

[15] Charles George Broyden.
The convergence of a class of double-rank minimization algorithms 1. general considerations.
IMA Journal of Applied Mathematics, 6(1):76–90, 1970.

[16] Camilla Cattania, Maximilian J Werner, Warner Marzocchi, Sebastian Hainzl, David Rhoades, Matthew Gerstenberger, Maria Liukis, William Savran, Annemarie Christophersen, Agnès Helmstetter, et al.

The forecasting skill of physics-based seismicity models during the 2010–2012 canterbury, new zealand, earthquake sequence.
Seismological Research Letters, 89(4):1238–1250, 2018.

[17] Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel.
Neural spatio-temporal point processes.
In *International Conference on Learning Representations*, 2021.
URL `https://openreview.net/forum?id=XQQA6-So14`.

[18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud.
Neural ordinary differential equations.
Advances in neural information processing systems, 31, 2018.

[19] Ricky TQ Chen, Brandon Amos, and Maximilian Nickel.
Neural spatio-temporal point processes.
arXiv preprint arXiv:2011.04583, 2020.

[20] Lauro Chiaraluce, Raffaele Di Stefano, Elisa Tinti, Laura Scognamiglio, Maddalena Michele, Emanuele Casarotti, Marco Cattaneo, Pasquale De Gori, Claudio Chiarabba, Giancarlo Monachesi, et al.
The 2016 central italy seismic sequence: A first look at the mainshocks, aftershocks, and source models.
Seismological Research Letters, 88(3):757–771, 2017.

[21] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever.
Generating long sequences with sparse transformers.
arXiv preprint arXiv:1904.10509, 2019.

[22] Pawel Chilinski and Ricardo Silva.
Neural likelihoods via cumulative distribution functions.
In *Conference on Uncertainty in Artificial Intelligence*, pages 420–429. PMLR, 2020.

[23] A Christophersen, DA Rhoades, MC Gerstenberger, S Bannister, J Becker, SH Potter, and S McBride.
Progress and challenges in operational earthquake forecasting in new zealand.
In *New Zealand society for earthquake engineering annual technical conference*, 2017.

[24] William S Cleveland.
Robust locally weighted regression and smoothing scatterplots.
Journal of the American statistical association, 74(368):829–836, 1979.

[25] Álvaro Corral and Álvaro González.
Power law size distributions in geoscience revisited.

Earth and Space Science, 6(5):673–697, 2019.

[26] Kyle Cranmer, Johann Brehmer, and Gilles Louppe.
The frontier of simulation-based inference.
Proceedings of the National Academy of Sciences, 117(48):30055–30062, 2020.

[27] Daryl J Daley and David Vere-Jones.
Basic properties of the poisson process.
An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, pages 19–40, 2003.

[28] Daryl J Daley and David Vere-Jones.
Scoring probability forecasts for point processes: The entropy score and information gain.
Journal of Applied Probability, 41(A):297–312, 2004.

[29] Daryl J Daley, David Vere-Jones, et al.
An introduction to the theory of point processes: volume I: elementary theory and methods.
Springer, 2003.

[30] Kelian Dascher-Cousineau, Emily E Brodsky, Thorne Lay, and Thomas HW Goebel.
What controls variations in aftershock productivity?
Journal of Geophysical Research: Solid Earth, 125(2):e2019JB018111, 2020.

[31] Kelian Dascher-Cousineau, Oleksandr Shchur, Emily E. Brodsky, and Stephan Günnemann.
Using deep learning for flexible and scalable earthquake forecasting.
Geophysical Research Letters, 50(17):e2023GL103909, 2023.
doi: https://doi.org/10.1029/2023GL103909.

[32] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau.
Gru-ode-bayes: Continuous modeling of sporadically-observed time series.
Advances in neural information processing systems, 32, 2019.

[33] Isabella Deutsch and Gordon J. Ross.
Abc learning of hawkes processes with missing or noisy event times, 2021.

[34] Jacob Devlin.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.

[35] James Dieterich.
A constitutive law for rate of earthquake production and its application to earthquake clustering.

Journal of Geophysical Research: Solid Earth, 99(B2):2601–2618, 1994.

[36] Peter Diggle.
A kernel method for smoothing point process data.
Journal of the Royal Statistical Society: Series C (Applied Statistics), 34(2):138–147, 1985.

[37] Philip Dixon.
Ripley's k function.
2001.

[38] Zheng Dong, Xiuyuan Cheng, and Yao Xie.
Spatio-temporal point processes with deep non-stationary kernels.
In *The Eleventh International Conference on Learning Representations*, 2023.
URL https://openreview.net/forum?id=PsIk0kO3hKd.

[39] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song.
Recurrent marked temporal point processes: Embedding event history to vector.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, 2016.

[40] Hossein Ebrahimian and Fatemeh Jalayer.
Robust seismicity forecasting based on bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models.
Scientific reports, 7(1):1–15, 2017.

[41] Hossein Ebrahimian, Fatemeh Jalayer, Domenico Asprone, Anna Maria Lombardi, Warner Marzocchi, Andrea Prota, and Gaetano Manfredi.
Adaptive daily forecasting of seismic aftershock hazard.
Bulletin of the Seismological Society of America, 104(1):145–161, 2014.

[42] Şeyda Ertekin, Cynthia Rudin, and Tyler H McCormick.
Reactive point processes: A new approach to predicting power failures in underground electrical systems.
2015.

[43] Karen Felzer, Rachel Abercrombie, and Goran Ekstrom.
A common origin for aftershocks, foreshocks, and multiplets.
Bulletin of The Seismological Society of America - BULL SEISMOL SOC AMER, 94: 88–98, 02 2004.
doi: 10.1785/0120030069.

[44] Karen R Felzer, Thorsten W Becker, Rachel E Abercrombie, Göran Ekström, and James R Rice.
Triggering of the 1999 mw 7.1 hector mine earthquake by aftershocks of the 1992 mw 7.3 landers earthquake.
Journal of Geophysical Research: Solid Earth, 107(B9):ESE–6, 2002.

[45] Edward H Field.
Overview of the working group for the development of regional earthquake likelihood models (relm).
Seismological Research Letters, 78(1):7–16, 2007.

[46] Edward H Field, Kevin R Milner, Jeanne L Hardebeck, Morgan T Page, Nicholas van der Elst, Thomas H Jordan, Andrew J Michael, Bruce E Shaw, and Maximilian J Werner.
A spatiotemporal clustering model for the third uniform california earthquake rupture forecast (ucerf3-etas): Toward an operational earthquake forecast.
Bulletin of the Seismological Society of America, 107(3):1049–1081, 2017.

[47] Edward H Field, Kevin R Milner, Morgan T Page, William H Savran, and Nicholas van der Elst.
Improvements to the third uniform california earthquake rupture forecast etas model (ucerf3-etas).
The Seismic Record, 1(2):117–125, 2021.

[48] Reeves Fletcher and Colin M Reeves.
Function minimization by conjugate gradients.
The computer journal, 7(2):149–154, 1964.

[49] Roger Fletcher.
A new approach to variable metric algorithms.
The computer journal, 13(3):317–322, 1970.

[50] Andrew M Freed.
Earthquake triggering by static, dynamic, and postseismic stress transfer.
Annu. Rev. Earth Planet. Sci., 33:335–367, 2005.

[51] Tomas Geffner, George Papamakarios, and Andriy Mnih.
Score modeling for simulation-based inference.
In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[52] Matt Gerstenberger, Stefan Wiemer, and Lucile M Jones.
Real-time forecasts of tomorrow's earthquakes in California: A new mapping tool.
US Geological Survey, 2004.

[53] Matthew Gerstenberger, Graeme McVerry, David Rhoades, and Mark Stirling.
Seismic hazard modeling for the recovery of christchurch.
Earthquake Spectra, 30(1):17–29, 2014.

[54] Federico Girosi, Michael Jones, and Tomaso Poggio.
Regularization theory and neural networks architectures.
Neural computation, 7(2):219–269, 1995.

[55] Tilmann Gneiting and Matthias Katzfuss.
Probabilistic forecasting.
Annual Review of Statistics and Its Application, 1(1):125–151, 2014.

[56] Tilmann Gneiting and Adrian E Raftery.
Strictly proper scoring rules, prediction, and estimation.
Journal of the American statistical Association, 102(477):359–378, 2007.

[57] Donald Goldfarb.
A family of variable-metric methods derived by variational means.
Mathematics of computation, 24(109):23–26, 1970.

[58] Joan Gomberg.
Unsettled earthquake nucleation.
Nature Geoscience, 11(7):463–464, 2018.

[59] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola.
A kernel two-sample test.
The Journal of Machine Learning Research, 13(1):723–773, 2012.

[60] Laura Gulia and Stefan Wiemer.
The influence of tectonic regimes on the earthquake size distribution: A case study for italy.
Geophysical Research Letters, 37(10), 2010.

[61] Laura Gulia, Antonio Pio Rinaldi, Thessa Tormann, Gianfranco Vannucci, Bogdan Enescu, and Stefan Wiemer.
The effect of a mainshock on the size distribution of the aftershocks.
Geophysical Research Letters, 45(24):13–277, 2018.

[62] Yicun Guo, Jiancang Zhuang, and Shiyong Zhou.
An improved space-time etas model for inverting the rupture geometry from seismicity triggering.
Journal of Geophysical Research: Solid Earth, 120(5):3309–3323, 2015.

[63] Yicun Guo, Jiancang Zhuang, and Yosihiko Ogata.
Modeling and forecasting aftershocks can be improved by incorporating rupture geometry in the etas model.
Geophysical Research Letters, 46(22):12881–12889, 2019.

[64] Beno Gutenberg and Charles Francis Richter.
Magnitude and energy of earthquakes.
Science, 83(2147):183–185, 1936.

[65] Sebastian Hainzl.
Apparent triggering function of aftershocks resulting from rate-dependent incompleteness of earthquake catalogs.
Journal of Geophysical Research: Solid Earth, 121(9):6499–6509, 2016.

[66] Sebastian Hainzl.
Rate-dependent incompleteness of earthquake catalogs.
Seismological Research Letters, 87(2A):337–344, 2016.

[67] Sebastian Hainzl.
Etas-approach accounting for short-term incompleteness of earthquake catalogs.
Bulletin of the Seismological Society of America, 112(1):494–507, 2022.

[68] Sebastian Hainzl and Yosihiko Ogata.
Detecting fluid signals in seismicity data through statistical earthquake modeling.
Journal of Geophysical Research: Solid Earth, 110(B5), 2005.

[69] Sebastian Hainzl, A Christophersen, and B Enescu.
Impact of earthquake rupture extensions on parameter estimations of point-process models.
Bulletin of the Seismological Society of America, 98(4):2066–2072, 2008.

[70] Sebastian Hainzl, Olga Zakharova, and David Marsan.
Impact of aseismic transients on the estimation of aftershock productivity parameters.
Bulletin of the Seismological Society of America, 103(3):1723–1732, 2013.

[71] Thomas C Hanks and Hiroo Kanamori.
A moment magnitude scale.
Journal of Geophysical Research: Solid Earth, 84(B5):2348–2350, 1979.

[72] Jeanne L. Hardebeck.
Appendix s — constraining epidemic type aftershock sequence (etas) parameters from the uniform california earthquake rupture forecast , version 3 catalog and validating the etas model for magnitude 6 . 5 or greater earthquakes.

2013.
URL https://api.semanticscholar.org/CorpusID:34402680.

[73] Jeanne L Hardebeck.
Physical properties of the crust influence aftershock locations.
Journal of Geophysical Research: Solid Earth, 127(10):e2022JB024727, 2022.

[74] DS Harte.
An Etas model with varying productivity rates.
Geophysical Journal International, 198(1):270–284, 2014.

[75] DS Harte.
Log-likelihood of earthquake models: evaluation of models and forecasts.
Geophysical Journal International, 201(2):711–723, 2015.

[76] Alan G Hawkes.
Spectra of some self-exciting and mutually exciting point processes.
Biometrika, 58(1):83–90, 1971.

[77] Agnes Helmstetter and Didier Sornette.
Importance of direct and indirect triggered seismicity in the etas model of seismicity.
Geophysical Research Letters, 30(11), 2003.

[78] Agnès Helmstetter and Maximilian J Werner.
Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for california.
Bulletin of the Seismological Society of America, 104(2):809–822, 2014.

[79] Agnes Helmstetter, Yan Y Kagan, and David D Jackson.
Comparison of short-term and time-independent earthquake forecast models for southern california.
Bulletin of the Seismological Society of America, 96(1):90–106, 2006.

[80] Agnes Helmstetter, Yan Kagan, and David Jackson.
High-resolution time-independent grid-based forecast for m¿= 5 earthquakes in california.
Seismological Research Letters, 78(1):78–86, 2007.

[81] Agnès Helmstetter, Yan Y. Kagan, and David D. Jackson.
Importance of small earthquakes for stress transfers and earthquake triggering.
Journal of Geophysical Research: Solid Earth, 110(B5), 2005.
doi: https://doi.org/10.1029/2004JB003286.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JB003286.

[82] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe.
A trust crisis in simulation-based inference? your posterior approximations can be unfaithful.
arXiv preprint arXiv:2110.06581, 2021.

[83] Marcus Herrmann and Warner Marzocchi.
Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high-resolution earthquake catalogs.
Seismological Research Letters, 92(2A):909–922, 2021.

[84] David P Hill.
A model for earthquake swarms.
Journal of Geophysical Research, 82(8):1347–1352, 1977.

[85] Sepp Hochreiter.
The vanishing gradient problem during learning recurrent neural nets and problem solutions.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02): 107–116, 1998.

[86] Timothy O Hodson.
Root mean square error (rmse) or mean absolute error (mae): When to use them or not.
Geoscientific Model Development Discussions, 2022:1–10, 2022.

[87] Hengguan Huang, Hao Wang, and Brian Mak.
Recurrent poisson process unit for speech recognition.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6538–6545, 2019.

[88] Kate Hutton, Jochen Woessner, and Egill Hauksson.
Earthquake monitoring in southern california for seventy-seven years (1932–2008).
Bulletin of the Seismological Society of America, 100(2):423–446, 2010.

[89] Pablo Iturrieta, José A Bayona, Maximilian J Werner, Danijel Schorlemmer, Matteo Taroni, Giuseppe Falcone, Fabrice Cotton, Asim M Khawaja, William H Savran, and Warner Marzocchi.
Evaluation of a decade-long prospective earthquake forecasting experiment in italy.
Seismological Research Letters, 2024.

[90] Rafael Izbicki, Ann Lee, and Chad Schafer.
High-dimensional density ratio estimation with extensions to approximate likelihood computation.

In *Artificial intelligence and statistics*, pages 420–429. PMLR, 2014.

[91] Junteng Jia and Austin R Benson.
Neural jump stochastic differential equations.
Advances in Neural Information Processing Systems, 32, 2019.

[92] Alan L Kafka and Shoshana Z Levin.
Does the spatial distribution of smaller earthquakes delineate areas where larger earthquakes are likely to occur?
Bulletin of the Seismological Society of America, 90(3):724–738, 2000.

[93] Alan L Kafka and Jessica R Walcott.
How well does the spatial distribution of smaller earthquakes forecast the locations of larger earthquakes in the northeastern united states?
Seismological Research Letters, 69(5):428–440, 1998.

[94] Yan Y Kagan.
Likelihood analysis of earthquake catalogues.
Geophysical journal international, 106(1):135–148, 1991.

[95] Yan Y Kagan.
Short-term properties of earthquake catalogs and models of earthquake source.
Bulletin of the Seismological Society of America, 94(4):1207–1228, 2004.

[96] Yan Y Kagan and L Knopoff.
Statistical short-term earthquake prediction.
Science, 236(4808):1563–1567, 1987.

[97] YY Kagan and DD Jackson.
Long-term probabilistic forecasting of earthquakes.
Journal of Geophysical Research: Solid Earth, 99(B7):13685–13700, 1994.

[98] Diederik P Kingma and Jimmy Ba.
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980, 2014.

[99] Qingkai Kong, Daniel T Trugman, Zachary E Ross, Michael J Bianco, Brendan J Meade, and Peter Gerstoft.
Machine learning in seismology: Turning data into insights.
Seismological Research Letters, 90(1):3–14, 2019.

[100] Takao Kumazawa and Yosihiko Ogata.
Quantitative description of induced seismic activity before and after the 2011 tohoku-oki earthquake by nonstationary etas models.

Journal of Geophysical Research: Solid Earth, 118(12):6165–6182, 2013.

[101] Takao Kumazawa and Yosihiko Ogata.
Nonstationary ETAS models for nonstandard earthquakes.
The Annals of Applied Statistics, 8(3):1825 – 1852, 2014.
doi: 10.1214/14-AOAS759.
URL https://doi.org/10.1214/14-AOAS759.

[102] Sacha Lapins, Berhe Goitom, J-Michael Kendall, Maximilian J Werner, Katharine V Cashman, and James OS Hammond.
A little data goes a long way: Automating seismic phase arrival picking at nabro volcano with transfer learning.
Journal of Geophysical Research: Solid Earth, 126(7):e2021JB021910, 2021.

[103] Steve Lawrence, C Lee Giles, and Ah Chung Tsoi.
Lessons in neural network training: Overfitting may be harder than expected.
In *AAAI/IAAI*, pages 540–545. Citeseer, 1997.

[104] E. L. Lehmann and Joseph P. Romano.
Testing statistical hypotheses.
Springer Texts in Statistics. Springer, New York, third edition, 2005.
ISBN 0-387-98864-5.

[105] Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song.
Learning temporal point processes via reinforcement learning.
arXiv preprint arXiv:1811.05016, 2018.

[106] Zichong Li, Qunzhi Xu, Zhenghao Xu, Yajun Mei, Tuo Zhao, and Hongyuan Zha.
Beyond point prediction: Score matching-based pseudolikelihood estimation of neural marked spatio-temporal point process.
In *Forty-first International Conference on Machine Learning*.

[107] Zichong Li, Qunzhi Xu, Zhenghao Xu, Yajun Mei, Tuo Zhao, and Hongyuan Zha.
Score matching-based pseudolikelihood estimation of neural marked spatio-temporal point process with uncertainty quantification.
arXiv preprint arXiv:2310.16310, 2023.

[108] Eugenio Lippiello, Warner Marzocchi, L De Arcangelis, and C Godano.
Spatial organization of foreshocks as a tool to forecast large earthquakes.
Scientific reports, 2(1):1–6, 2012.

[109] Eugenio Lippiello, F Giacco, L de Arcangelis, W Marzocchi, and Cataldo Godano.
Parameter estimation in the etas model: Approximations and novel methods.

Bulletin of the Seismological Society of America, 104(2):985–994, 2014.

[110] Andrea L Llenos and Nicholas J van der Elst.
Improving earthquake forecasts during swarms with a duration model.
Bulletin of the Seismological Society of America, 109(3):1148–1155, 2019.

[111] Andrea L Llenos, Jeffrey J McGuire, and Yosihiko Ogata.
Modeling seismic swarms triggered by aseismic transients.
Earth and Planetary Science Letters, 281(1-2):59–69, 2009.

[112] Anthony Lomax, Jean Virieux, Philippe Volant, and Catherine Berge-Thierry.
Probabilistic earthquake location in 3d and layered models: Introduction of a metropolis-gibbs method and comparison with linear locations.
Advances in seismic event location, pages 101–134, 2000.

[113] Anna Maria Lombardi.
Estimation of the parameters of etas models by simulated annealing.
Scientific reports, 5(1):8417, 2015.

[114] Anna Maria Lombardi, Warner Marzocchi, and Jacopo Selva.
Exploring the evolution of a volcanic seismic swarm: The case of the 2000 izu islands swarm.
Geophysical research letters, 33(7), 2006.

[115] David Lopez-Paz and Maxime Oquab.
Revisiting classifier two-sample tests.
arXiv preprint arXiv:1610.06545, 2016.

[116] Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke.
Flexible statistical inference for mechanistic models of neural dynamics.
Advances in neural information processing systems, 30, 2017.

[117] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke.
Benchmarking simulation-based inference.
In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

[118] S Mancini, M Segou, MJ Werner, and C Cattania.
Improving physics-based aftershock forecasts during the 2016–2017 central italy earthquake cascade.
Journal of Geophysical Research: Solid Earth, 124(8):8626–8643, 2019.

[119] Simone Mancini, Margarita Segou, Maximilian Jonas Werner, and Tom Parsons.
The predictive skills of elastic coulomb rate-and-state aftershock forecasts during the 2019 ridgecrest, california, earthquake sequence.
Bulletin of the Seismological Society of America, 110(4):1736–1751, 2020.

[120] Simone Mancini, Margarita Segou, MJ Werner, Tom Parsons, Gregory Beroza, and Lauro Chiaraluce.
On the use of high-resolution and deep-learning seismic catalogs for short-term earthquake forecasts: Potential benefits and current limitations.
Journal of Geophysical Research: Solid Earth, 127(11):e2022JB025202, 2022.

[121] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré.
Markov chain monte carlo without likelihoods.
Proceedings of the National Academy of Sciences, 100(26):15324–15328, 2003.

[122] David Marsan.
The role of small earthquakes in redistributing crustal elastic stress.
Geophysical Journal International, 163(1):141–151, 2005.

[123] David Marsan and Olivier Lengline.
Extending earthquakes' reach through cascading.
Science, 319(5866):1076–1079, 2008.

[124] Warner Marzocchi, Anna Maria Lombardi, and Emanuele Casarotti.
The establishment of an operational earthquake forecasting system in italy.
Seismological Research Letters, 85(5):961–969, 2014.

[125] Warner Marzocchi, Matteo Taroni, and Giuseppe Falcone.
Earthquake forecasting during the complex amatrice-norcia seismic sequence.
Science Advances, 3(9):e1701239, 2017.

[126] Jeffrey J McGuire, Margaret S Boettcher, and Thomas H Jordan.
Foreshock sequences and short-term earthquake predictability on east pacific rise transform faults.
Nature, 434(7032):457–461, 2005.

[127] Hongyuan Mei and Jason M Eisner.
The neural hawkes process: A neurally self-modulating multivariate point process.
Advances in neural information processing systems, 30, 2017.

[128] Andrew J Michael and Maximilian J Werner.
Preface to the focus section on the collaboratory for the study of earthquake predictability (csep): New results and future directions.

Seismological Research Letters, 89(4):1226–1228, 2018.

[129] A Mignan, MJ Werner, S Wiemer, C-C Chen, and Y-M Wu.
Bayesian estimation of the spatially varying completeness magnitude of earthquake catalogs.
Bulletin of the Seismological Society of America, 101(3):1371–1385, 2011.

[130] Arnaud Mignan and Jochen Woessner.
Theme iv—understanding seismicity catalogs and their problems.
Community online resource for statistical seismicity analysis, 2012.

[131] Kevin R Milner, Edward H Field, William H Savran, Morgan T Page, and Thomas H Jordan.
Operational earthquake forecasting during the 2019 ridgecrest, california, earthquake sequence with the ucerf3-etas model.
Seismological Research Letters, 91(3):1567–1578, 2020.

[132] Leila Mizrahi, Shyam Nandan, and Stefan Wiemer.
The effect of declustering on the size distribution of mainshocks.
Seismological Society of America, 92(4):2333–2342, 2021.

[133] Leila Mizrahi, Shyam Nandan, and Stefan Wiemer.
Embracing data incompleteness for better earthquake forecasting.
Journal of Geophysical Research: Solid Earth, 126(12):e2021JB022379, 2021.

[134] Leila Mizrahi, Nicolas Schmid, and Marta Han.
lmizrahi/etas, 2022.
URL `https://doi.org/10.5281/zenodo.6583992`.

[135] Leila Mizrahi, Shyam Nandan, Banu Mena Cabrera, and Stefan Wiemer.
suiETAS: Developing and Testing ETAS-Based Earthquake Forecasting Models for Switzerland.
Bulletin of the Seismological Society of America, 05 2024.
doi: 10.1785/0120240007.

[136] Christian Molkenthin, Christian Donner, Sebastian Reich, Gert Zöller, Sebastian Hainzl, Matthias Holschneider, and Manfred Opper.
Gp-etas: semiparametric bayesian inference for the spatio-temporal epidemic type aftershock sequence model.
Statistics and computing, 32(2):29, 2022.

[137] Jesper Møller and Rasmus P Waagepetersen.
An introduction to simulation-based inference for spatial point processes.

In *Spatial statistics and computational methods*, pages 143–198. Springer, 2003.

[138] Nobuhito Mori, Tomoyuki Takahashi, Tomohiro Yasuda, and Hideaki Yanagisawa.
Survey of 2011 tohoku earthquake tsunami inundation and run-up.
Geophysical research letters, 38(7), 2011.

[139] S Mostafa Mousavi and Gregory C Beroza.
Machine learning in earthquake seismology.
Annual Review of Earth and Planetary Sciences, 51:105–129, 2023.

[140] S Mostafa Mousavi, William L Ellsworth, Weiqiang Zhu, Lindsay Y Chuang, and Gregory C Beroza.
Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking.
Nature communications, 11(1):3952, 2020.

[141] Shyam Nandan, Guy Ouillon, Jochen Woessner, Didier Sornette, and Stefan Wiemer.
Systematic assessment of the static stress triggering hypothesis using interearthquake time statistics.
Journal of Geophysical Research: Solid Earth, 121(3):1890–1909, 2016.

[142] Shyam Nandan, Guy Ouillon, Stefan Wiemer, and Didier Sornette.
Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to california.
Journal of Geophysical Research: Solid Earth, 122(7):5118–5143, 2017.

[143] Shyam Nandan, Sumit Kumar Ram, Guy Ouillon, and Didier Sornette.
Is seismicity operating at a critical point?
Physical Review Letters, 126(12):128501, 2021.

[144] Shyam Nandan, Guy Ouillon, and Didier Sornette.
Are large earthquakes preferentially triggered by other large events?
Journal of Geophysical Research: Solid Earth, 127(8):e2022JB024380, 2022.

[145] Radford M Neal.
Slice sampling.
The annals of statistics, 31(3):705–767, 2003.

[146] John A Nelder and Roger Mead.
A simplex method for function minimization.
The computer journal, 7(4):308–313, 1965.

[147] Paul Nyffenegger and Cliff Frohlich.

Recommendations for determining p values for aftershock sequences and catalogs.
Bulletin of the Seismological Society of America, 88(5):1144–1154, 1998.

[148] Yosihiko Ogata.
Estimators for stationary point processes.
Ann. Inst. Statist. Math, 30(Part A):243–261, 1978.

[149] Yosihiko Ogata.
On lewis' simulation method for point processes.
IEEE transactions on information theory, 27(1):23–31, 1981.

[150] Yosihiko Ogata.
Statistical models for earthquake occurrences and residual analysis for point processes.
Journal of the American Statistical association, 83(401):9–27, 1988.

[151] Yosihiko Ogata.
Space-time point-process models for earthquake occurrences.
Annals of the Institute of Statistical Mathematics, 50(2):379–402, 1998.

[152] Yosihiko Ogata and Koichi Katsura.
Comparing foreshock characteristics and foreshock forecasting in observed and simulated earthquake catalogs.
Journal of Geophysical Research: Solid Earth, 119(11):8457–8477, 2014.

[153] Yosihiko Ogata and Jiancang Zhuang.
Space–time etas models and an improved extension.
Tectonophysics, 413(1-2):13–23, 2006.

[154] Takahiro Omi, Yosihiko Ogata, Yoshito Hirata, and Kazuyuki Aihara.
Estimating the etas model from an early aftershock sequence.
Geophysical Research Letters, 41(3):850–857, 2014.

[155] Takahiro Omi, Yosihiko Ogata, Yoshito Hirata, and Kazuyuki Aihara.
Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches.
Journal of Geophysical Research: Solid Earth, 120(4):2561–2578, 2015.

[156] Takahiro Omi, Yosihiko Ogata, Katsuhiko Shiomi, Bogdan Enescu, Kaoru Sawazaki, and Kazuyuki Aihara.
Implementation of a real-time system for automatic aftershock forecasting in japan.
Seismological Research Letters, 90(1):242–250, 2019.

[157] Takahiro Omi, naonori ueda, and Kazuyuki Aihara.

Fully neural network based model for general temporal point processes.
In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 2122–2132. Curran Associates, Inc., 2019.
URL https://proceedings.neurips.cc/paper/2019/file/39e4973ba3321b80f37d9b55f63ed8b8-Paper.pdf.

[158] Morgan T Page and Nicholas J van der Elst.
Turing-style tests for ucerf3 synthetic catalogs.
Bulletin of the Seismological Society of America, 108(2):729–741, 2018.

[159] Morgan T Page, Nicholas van der Elst, Jeanne Hardebeck, Karen Felzer, and Andrew J Michael.
Three ingredients for improved global aftershock forecasts: Tectonic region, time-dependent catalog incompleteness, and intersequence variability.
Bulletin of the Seismological Society of America, 106(5):2290–2301, 2016.

[160] George Papamakarios and Iain Murray.
Fast ε-free inference of simulation models with bayesian conditional density estimation.
Advances in neural information processing systems, 29, 2016.

[161] George Papamakarios, Theo Pavlakou, and Iain Murray.
Masked autoregressive flow for density estimation.
Advances in neural information processing systems, 30, 2017.

[162] George Papamakarios, David Sterratt, and Iain Murray.
Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[163] Fredos Papangelou.
Integrability of expected increments of point processes and a related random change of scale.
Transactions of the American Mathematical Society, 165:483–506, 1972.

[164] Sally H Potter, Julia S Becker, David M Johnston, and Katelyn P Rossiter.
An overview of the impacts of the 2010-2011 canterbury earthquakes.
International Journal of Disaster Risk Reduction, 14:6–14, 2015.

[165] Dennis Prangle, Michael GB Blum, Gordana Popovic, and SA Sisson.
Diagnostic tools for approximate bayesian computation using the coverage property.
Australian & New Zealand Journal of Statistics, 56(4):309–329, 2014.

[166] Dennis Prangle, Paul Fearnhead, Murray P Cox, Patrick J Biggs, and Nigel P French.
Semi-automatic selection of summary statistics for abc model choice.
Statistical applications in genetics and molecular biology, 13(1):67–82, 2014.

[167] Jakob Gulddahl Rasmussen.
Bayesian inference for hawkes processes.
Methodology and Computing in Applied Probability, 15:623–642, 2013.

[168] Jakob Gulddahl Rasmussen.
Lecture notes: Temporal point processes and the conditional intensity function.
arXiv preprint arXiv:1806.00221, 2018.

[169] Stephen L Rathbun.
Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes.
Journal of Statistical Planning and Inference, 51(1):55–74, 1996.

[170] Danilo Rezende and Shakir Mohamed.
Variational inference with normalizing flows.
In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[171] David A Rhoades, Annemarie Christophersen, Matthew C Gerstenberger, Maria Liukis, Fabio Silva, Warner Marzocchi, Maximilian J Werner, and Thomas H Jordan.
Highlights from the first ten years of the new zealand earthquake forecast testing center.
Seismological Research Letters, 89(4):1229–1237, 2018.

[172] Charles F Richter.
An instrumental earthquake magnitude scale.
Bulletin of the seismological society of America, 25(1):1–32, 1935.

[173] Brian D Ripley.
Modelling spatial patterns.
Journal of the Royal Statistical Society: Series B (Methodological), 39(2):172–192, 1977.

[174] Gordon J Ross.
Bayesian estimation of the etas model for earthquake occurrences.
Bulletin of the Seismological Society of America, 111(3):1473–1480, 2021.

[175] Zachary E Ross, Daniel T Trugman, Egill Hauksson, and Peter M Shearer.
Searching for hidden earthquakes in southern california.
Science, 364(6442):767–771, 2019.

[176] William H Savran, Maximilian J Werner, Warner Marzocchi, David A Rhoades, David D Jackson, Kevin Milner, Edward Field, and Andrew Michael.
Pseudoprospective evaluation of ucerf3-etas forecasts during the 2019 ridgecrest sequence.
Bulletin of the Seismological Society of America, 110(4):1799–1817, 2020.

[177] William H Savran, José A Bayona, Pablo Iturrieta, Khawaja M Asim, Han Bao, Kirsty Bayliss, Marcus Herrmann, Danijel Schorlemmer, Philip J Maechling, and Maximilian J Werner.
pycsep: a python toolkit for earthquake forecast developers.
Seismological Society of America, 93(5):2858–2870, 2022.

[178] Danijel Schorlemmer and MC Gerstenberger.
Relm testing center.
Seismological Research Letters, 78(1):30–36, 2007.

[179] Danijel Schorlemmer and Jochen Woessner.
Probability of detecting an earthquake.
Bulletin of the Seismological Society of America, 98(5):2103–2117, 2008.

[180] Danijel Schorlemmer, Stefan Wiemer, and Max Wyss.
Variations in earthquake-size distribution across different stress regimes.
Nature, 437(7058):539–542, 2005.

[181] Danijel Schorlemmer, Maximilian J Werner, Warner Marzocchi, Thomas H Jordan, Yosihiko Ogata, David D Jackson, Sum Mak, David A Rhoades, Matthew C Gerstenberger, Naoshi Hirata, et al.
The collaboratory for the study of earthquake predictability: Achievements and priorities.
Seismological Research Letters, 89(4):1305–1313, 2018.

[182] Salvatore Scudero, Carlo Marcocci, and Antonino D'Alessandro.
Insights on the italian seismic network from location uncertainties.
Journal of Seismology, 25(4):1061–1076, 2021.

[183] Stefanie Seif, Arnaud Mignan, Jeremy Douglas Zechar, Maximilian Jonas Werner, and Stefan Wiemer.
Estimating etas: The effects of truncation, missing data, and model assumptions.
Journal of Geophysical Research: Solid Earth, 122(1):449–469, 2017.

[184] Francesco Serafini, Finn Lindgren, and Mark Naylor.
Approximation of bayesian hawkes process with inlabru.
Environmetrics, page e2798, 2023.

[185] David F Shanno.
Conditioning of quasi-newton methods for function minimization.
Mathematics of computation, 24(111):647–656, 1970.

[186] Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont.
Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models.
arXiv preprint arXiv:2210.04872, 2022.

[187] Robert Shcherbakov.
Statistics and forecasting of aftershocks during the 2019 Ridgecrest, California, earthquake sequence.
Journal of Geophysical Research: Solid Earth, 126(2):e2020JB020887, 2021.

[188] Robert Shcherbakov, Jiancang Zhuang, Gert Zöller, and Yosihiko Ogata.
Forecasting the magnitude of the largest expected earthquake.
Nature communications, 10(1):4051, 2019.

[189] Oleksandr Shchur.
Modeling Continuous-time Event Data with Neural Temporal Point Processes.
PhD thesis, Technische Universität München, 2022.

[190] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann.
Intensity-free learning of temporal point processes.
arXiv preprint arXiv:1909.12127, 2019.

[191] Oleksandr Shchur, Nicholas Gao, Marin Biloš, and Stephan Günnemann.
Fast and flexible temporal point processes with triangular maps.
Advances in neural information processing systems, 33:73–84, 2020.

[192] Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann.
Neural temporal point processes: A review.
arXiv preprint arXiv:2104.03528, 2021.

[193] Peter M Shearer.
Introduction to seismology.
Cambridge university press, 2019.

[194] David R Shelly.
A 15 year catalog of more than 1 million low-frequency earthquakes: Tracking tremor and slip along the deep san andreas fault.
Journal of Geophysical Research: Solid Earth, 122(5):3739–3753, 2017.

[195] Didier Sornette and Maximilian J Werner.
Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences.
Journal of Geophysical Research: Solid Earth, 110(B8), 2005.

[196] Didier Sornette and Maximillian J Werner.
Apparent clustering and apparent background earthquakes biased by undetected seismicity.
Journal of Geophysical Research: Solid Earth, 110(B9), 2005.

[197] Ilaria Spassiani, Giuseppe Falcone, Maura Murru, and Warner Marzocchi.
Operational earthquake forecasting in italy: validation after 10 yr of operativity.
Geophysical Journal International, 234(3):2501–2518, 2023.

[198] Ross S Stein.
The role of stress transfer in earthquake occurrence.
Nature, 402(6762):605–609, 1999.

[199] Seth Stein and Michael Wysession.
An introduction to seismology, earthquakes, and earth structure.
John Wiley & Sons, 2009.

[200] Samuel Stockman.
ss15859/neural-point-process, 2023.

[201] Samuel Stockman, Daniel J. Lawson, and Maximilian J. Werner.
Forecasting the 2016–2017 central apennines earthquake sequence with a neural point process.
Earth's Future, 11(9):e2023EF003777, 2023.
doi: https://doi.org/10.1029/2023EF003777.
e2023EF003777 2023EF003777.

[202] Samuel Stockman, Daniel J. Lawson, and Maximilian J. Werner.
Sb-etas: using simulation based inference for scalable, likelihood-free inference for the etas model of earthquake occurrences.
Statistics and Computing, 34(5):174, 2024.
ISSN 1573-1375.
doi: 10.1007/s11222-024-10486-6.
URL `https://doi.org/10.1007/s11222-024-10486-6`.

[203] Richard Styron and Marco Pagani.
The gem global active faults database.
Earthquake Spectra, 36(1_suppl):160–180, 2020.

[204] Tetsuo Takanami, Genshiro Kitagawa, and Kazushige Obara.
Hi-net: High sensitivity seismograph network, japan.
Methods and applications of signal processing in seismic network operations, pages 79–88, 2003.

[205] Yen Joe Tan, Felix Waldhauser, William L Ellsworth, Miao Zhang, Weiqiang Zhu, Maddalena Michele, Lauro Chiaraluce, Gregory C Beroza, and Margarita Segou.
Machine-learning-based high-resolution earthquake catalog for the 2016–2017 central italy sequence, 2021.
URL `https://doi.org/10.5281/zenodo.4736089`.

[206] Yen Joe Tan, Felix Waldhauser, William L Ellsworth, Miao Zhang, Weiqiang Zhu, Maddalena Michele, Lauro Chiaraluce, Gregory C Beroza, and Margarita Segou.
Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 central italy sequence.
The Seismic Record, 1(1):11–19, 2021.

[207] Matteo Taroni, Warner Marzocchi, Danijel Schorlemmer, Maximilian Jonas Werner, Stefan Wiemer, Jeremy Douglas Zechar, Lukas Heiniger, and Fabian Euchner.
Prospective csep evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for italy.
Seismological Research Letters, 89(4):1251–1261, 2018.

[208] Clifford H Thurber.
Nonlinear earthquake location: theory and examples.
Bulletin of the Seismological Society of America, 75(3):779–790, 1985.

[209] Daniel T Trugman and Yehuda Ben-Zion.
Coherent spatial variations in the productivity of earthquake sequences in california and nevada.
The Seismic Record, 3(4):322–331, 2023.

[210] Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez.
Deep reinforcement learning of marked temporal point processes.
arXiv preprint arXiv:1805.09360, 2018.

[211] Tokuji Utsu.
A relation between the area of after-shock region and the energy of main-shock.
J. Seismol. Soc. Jpn., 2, 7:233–240, 1955.

[212] Tokuji Utsu.
Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations.

Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics, 3(3): 129–195, 1970.

[213] Tokuji Utsu.
Aftershocks and earthquake statistics (2): further investigation of aftershocks and other earthquake sequences based on a new classification of earthquake sequences.
Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics, 3(4): 197–266, 1971.

[214] Tokuji Utsu and Akira Seki.
A relation between the area of after-shock region and the energy of main-shock.
Journal of the Seismological Society of Japan, 7:233–240, 1955.
URL `https://api.semanticscholar.org/CorpusID:133541209`.

[215] Tokuji Utsu, Yosihiko Ogata, et al.
The centenary of the omori formula for a decay law of aftershock activity.
Journal of Physics of the Earth, 43(1):1–33, 1995.

[216] Nicholas J van der Elst.
B-positive: A robust estimator of aftershock magnitude distribution in transiently incomplete catalogs.
Journal of Geophysical Research: Solid Earth, 126(2):e2020JB021027, 2021.

[217] Nicholas J van der Elst, Jeanne L Hardebeck, Andrew J Michael, Sara K McBride, and Elizabeth Vanacore.
Prospective and retrospective evaluation of the us geological survey public aftershock forecast for the 2019–2021 southwest puerto rico earthquake and aftershocks.
Seismological Society of America, 93(2A):620–640, 2022.

[218] Bart Van Merriënboer, Olivier Breuleux, Arnaud Bergeron, and Pascal Lamblin.
Automatic differentiation in ml: Where we are and where we should be going.
Advances in neural information processing systems, 31, 2018.

[219] N Vargas and Tilmann Gneiting.
Bayesian point process modelling of earthquake occurrences.
Technical report, Technical Report, Ruprecht-Karls University Heidelberg, Heidelberg, Germany . . . , 2012.

[220] A Vaswani.
Attention is all you need.
Advances in Neural Information Processing Systems, 2017.

[221] Alejandro Veen and Frederic P Schoenberg.
Estimation of space–time branching process models in seismology using an em–type algorithm.
Journal of the American Statistical Association, 103(482):614–624, 2008.

[222] Qi Wang, Frederic Paik Schoenberg, and David D Jackson.
Standard errors of parameter estimates in the etas model.
Bulletin of the Seismological Society of America, 100(5A):1989–2001, 2010.

[223] Yuan Wang, Zhipeng Gui, Huayi Wu, Dehua Peng, Jinghang Wu, and Zousen Cui.
Optimizing and accelerating space–time ripley's k function based on apache spark for distributed spatiotemporal point pattern analysis.
Future Generation Computer Systems, 105:96–118, 2020.

[224] Maximilian J Werner, Agnès Helmstetter, David D Jackson, and Yan Y Kagan.
High-resolution long-term and short-term earthquake forecasts for california.
Bulletin of the Seismological Society of America, 101(4):1630–1648, 2011.

[225] Malcolm CA White, Yehuda Ben-Zion, and Frank L Vernon.
A detailed earthquake catalog for the san jacinto fault-zone region in southern california.
Journal of Geophysical Research: Solid Earth, 124(7):6908–6930, 2019.

[226] Stefan Wiemer and Max Wyss.
Minimum magnitude of completeness in earthquake catalogs: Examples from alaska, the western united states, and japan.
Bulletin of the Seismological Society of America, 90(4):859–869, 2000.

[227] J Woessner, JL Hardebeck, and E Haukkson.
What is an instrumental seismicity catalog, community online resource for statistical seismicity analysis, doi: 10.5078/corssa-38784307, 2010.

[228] J Woessner, Sebastian Hainzl, W Marzocchi, MJ Werner, AM Lombardi, F Catalli, B Enescu, M Cocco, MC Gerstenberger, and S Wiemer.
A retrospective comparative forecast test on the 1992 landers sequence.
Journal of Geophysical Research: Solid Earth, 116(B5), 2011.

[229] J Woessner, J Hardebeck, and E Hauksson.
Theme iv—understanding seismicity catalogs and their problems, 2022.

[230] Jochen Woessner and Stefan Wiemer.
Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty.
Bulletin of the Seismological Society of America, 95(2):684–698, 2005.

[231] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen Chu.
Modeling the intensity function of point process via recurrent neural networks.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[232] Zheng Xiaogu and David Vere-Jones.
Further applications of the stochastic stress release model to historical earthquake data.
Tectonophysics, 229(1-2):101–121, 1994.

[233] Hanwen Xing, Geoff Nicholls, and Jeong Lee.
Calibrated approximate bayesian inference.
In *International Conference on Machine Learning*, pages 6912–6920. PMLR, 2019.

[234] Xue Ying.
An overview of overfitting and its solutions.
In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

[235] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li.
Spatio-temporal diffusion point processes.
In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3173–3184, 2023.

[236] J Douglas Zechar, Matthew C Gerstenberger, and David A Rhoades.
Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts.
Bulletin of the Seismological Society of America, 100(3):1184–1195, 2010.

[237] J Douglas Zechar, Danijel Schorlemmer, Maximilian J Werner, Matthew C Gerstenberger, David A Rhoades, and Thomas H Jordan.
Regional earthquake likelihood models i: First-order results.
Bulletin of the Seismological Society of America, 103(2A):787–798, 2013.

[238] Xiao-Gu Zheng and David Vere-Jones.
Application of stress release models to historical earthquakes from north china.
Pure and Applied Geophysics, 135(4):559–576, 1991.

[239] Zihao Zhou and Rose Yu.
Automatic integration for spatiotemporal neural point processes.
Advances in Neural Information Processing Systems, 36, 2024.

[240] Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu.
Neural point process for learning spatiotemporal event dynamics.
In *Learning for Dynamics and Control Conference*, pages 777–789. PMLR, 2022.

[241] Shixiang Zhu, Haoyun Wang, Xiuyuan Cheng, and Yao Xie.
Neural spectral marked point processes.
In *International Conference on Learning Representations*, 2022.
URL `https://openreview.net/forum?id=0rcbOaoBXbg`.

[242] Weiqiang Zhu and Gregory C Beroza.
Phasenet: a deep-neural-network-based seismic arrival-time picking method.
Geophysical Journal International, 216(1):261–273, 2019.

[243] Jiancang Zhuang, David S Harte, Maximilian J Werner, Sebastian Hainzl, and Shiyong Zhou.
Basic models of seismicity: Temporal models.
2012.

[244] Jiancang Zhuang, Maximilian J Werner, Sebastian Hainzl, David Harte, and Shiyong Zhou.
Basic models of seismicity: temporal models.
Community Online Resource for Statistical Seismicity Analysis, 2012.

[245] Jiancang Zhuang, Yosihiko Ogata, and Ting Wang.
Data completeness of the kumamoto earthquake sequence in the jma catalog and its influence on the estimation of the etas parameters.
Earth, Planets and Space, 69(1):1–12, 2017.

[246] Mark D Zoback, Mary Lou Zoback, Van S Mount, John Suppe, Jerry P Eaton, John H Healy, David Oppenheimer, Paul Reasenberg, Lucile Jones, C Barry Raleigh, et al.
New evidence on the state of stress of the san andreas fault system.
Science, 238(4830):1105–1111, 1987.

[247] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha.
Transformer hawkes process.
In *International conference on machine learning*, pages 11692–11702. PMLR, 2020.

PGR_submission_Stockman_Sam_1503332.pdf

ORIGINALITY REPORT

56%	21%	52%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Samuel Stockman, Daniel J. Lawson, Maximilian J. Werner. "Forecasting the 2016–2017 Central Apennines Earthquake Sequence With a Neural Point Process", Earth's Future, 2023 Publication	20%
2	Samuel Stockman, Daniel J. Lawson, Maximilian J. Werner. "SB-ETAS: using simulation based inference for scalable, likelihood-free inference for the ETAS model of earthquake occurrences", Statistics and Computing, 2024 Publication	18%
3	era.ed.ac.uk Internet Source	1%
4	research-information.bris.ac.uk Internet Source	1%
5	agupubs.onlinelibrary.wiley.com Internet Source	1%
6	arxiv.org Internet Source	1%

7	export.arxiv.org Internet Source	1 %
8	central.scec.org Internet Source	1 %
9	hdl.handle.net Internet Source	1 %
10	Submitted to University of Bristol Student Paper	<1 %
11	www.researchgate.net Internet Source	<1 %
12	Leonhard Held, Niel Hens, Philip O'Neill, Jacco Wallinga. "Handbook of Infectious Disease Data Analysis", CRC Press, 2019 Publication	<1 %
13	escholarship.org Internet Source	<1 %
14	Alba Bernabeu, Jiancang Zhuang, Jorge Mateu. "Spatio-Temporal Hawkes Point Processes: A Review", Journal of Agricultural, Biological and Environmental Statistics, 2024 Publication	<1 %
15	Submitted to University of Birmingham Student Paper	<1 %
16	d-nb.info Internet Source	<1 %

17	files.scec.org Internet Source	<1 %
18	proceedings.mlr.press Internet Source	<1 %
19	earthquake.usgs.gov Internet Source	<1 %
20	www.usgs.gov Internet Source	<1 %
21	Submitted to University College London Student Paper	<1 %
22	www.research-collection.ethz.ch Internet Source	<1 %
23	Chenlong Li, Zhanjie Song, Xu Wang. "Nonparametric method for modelling clustering phenomena in emergency calls under spatial-temporal self-exciting point processes", IEEE Access, 2019 Publication	<1 %
24	proceedings.neurips.cc Internet Source	<1 %
25	openreview.net Internet Source	<1 %
26	nnr.co.za Internet Source	<1 %
	edoc.ub.uni-muenchen.de	

27	Internet Source	<1%
28	scholar.deep-time.org Internet Source	<1%
29	u-pad.unimc.it Internet Source	<1%
30	Guo, Yicun, Jiancang Zhuang, and Shiyong Zhou. "An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering : Inverting rupture geometry", Journal of Geophysical Research Solid Earth, 2015. Publication	<1%
31	Thystere Matondo Bantidi, Takeo Ishibe, Georges Mavonga Tuluka, Bogdan Enescu. "Estimating spatio-temporal variable parameters of Epidemic Type Aftershock Sequence model in a region with limited seismic network coverage: a case study of the East African Rift System", Geophysical Journal International, 2024 Publication	<1%
32	scholars.wlu.ca Internet Source	<1%
33	stat.paperswithcode.com Internet Source	<1%

34	"ECAI 2020", IOS Press, 2020 Publication	<1 %
35	pdffox.com Internet Source	<1 %
36	repository.kulib.kyoto-u.ac.jp Internet Source	<1 %
37	edoc.hu-berlin.de Internet Source	<1 %
38	static.seismo.ethz.ch Internet Source	<1 %
39	Fan Zhang, Xiao-Zhong Yang, Feng-Zhi Cui. "Earthquake detection probabilities and completeness magnitude in the northern margin of the Ordos Block", Applied Geophysics, 2024 Publication	<1 %
40	Submitted to University of Sydney Student Paper	<1 %
41	discovery.ucl.ac.uk Internet Source	<1 %
42	dr.ntu.edu.sg Internet Source	<1 %
43	tel.archives-ouvertes.fr Internet Source	<1 %

44	Encyclopedia of Earthquake Engineering, 2015. Publication	<1 %
45	Frederic Paik Schoenberg, Marc Hoffmann, Ryan J. Harrigan. "A recursive point process model for infectious diseases", Annals of the Institute of Statistical Mathematics, 2018 Publication	<1 %
46	dspace.mit.edu Internet Source	<1 %
47	Submitted to Associatie K.U.Leuven Student Paper	<1 %
48	dokumen.pub Internet Source	<1 %
49	papyrus.bib.umontreal.ca Internet Source	<1 %
50	Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, Ashok N. Srivastava. "Advances in Machine Learning and Data Mining for Astronomy", Chapman and Hall/CRC, 2019 Publication	<1 %
51	edocs.fu-berlin.de Internet Source	<1 %
52	Submitted to University of Edinburgh Student Paper	<1 %

53	dblp.org Internet Source	<1 %
54	theses.ncl.ac.uk Internet Source	<1 %
55	Submitted to UC, San Diego Student Paper	<1 %
56	Submitted to University of Southampton Student Paper	<1 %
57	ethz.ch Internet Source	<1 %
58	www.mdpi.com Internet Source	<1 %
59	Hernando Ombao, Martin Lindquist, Wesley Thompson, John Aston. "Handbook of Neuroimaging Data Analysis", Chapman and Hall/CRC, 2019 Publication	<1 %
60	Submitted to Ulsan National Institute of Science and Technology Student Paper	<1 %
61	par.nsf.gov Internet Source	<1 %
62	Submitted to UCL Student Paper	<1 %
	impa.usc.edu	

63	Internet Source	<1%
64	Chen, Ricky Tian Qi. "Generative Modeling with Differentiable Dynamics", University of Toronto (Canada), 2023 Publication	<1%
65	digital.lib.washington.edu Internet Source	<1%
66	publikationen.sulb.uni-saarland.de Internet Source	<1%
67	cloud.tencent.com Internet Source	<1%
68	orca.cardiff.ac.uk Internet Source	<1%
69	www.ism.ac.jp Internet Source	<1%
70	api.repository.cam.ac.uk Internet Source	<1%
71	Submitted to Skolkovo Institute of Science and Technology (Skoltech) Student Paper	<1%
72	Anna Maria Lombardi, Warner Marzocchi. "The double branching model for earthquake forecast applied to the Japanese seismicity", Earth, Planets and Space, 2011 Publication	<1%

73	Haoyuan Zhang, Shuya Ke, Wenqi Liu, Yongwen Zhang. "A combining earthquake forecasting model between deep learning and Epidemic-Type Aftershock Sequence (ETAS) model", Geophysical Journal International, 2024 Publication	<1 %
74	Submitted to Queen Mary and Westfield College Student Paper	<1 %
75	Siwei Yu, Jianwei Ma. "Deep Learning for Geophysics: Current and Future Trends", Reviews of Geophysics, 2021 Publication	<1 %
76	Submitted to University of Lancaster Student Paper	<1 %
77	geolib.geo.auth.gr Internet Source	<1 %
78	ir.canterbury.ac.nz Internet Source	<1 %
79	Shang, Jin. "Predictive Modeling of Asynchronous Event Sequence Data", Louisiana State University and Agricultural & Mechanical College, 2023 Publication	<1 %
80	web.math.ku.dk Internet Source	

		<1%
81	"Earthquakes and Multi-hazards Around the Pacific Rim, Vol. II", Springer Science and Business Media LLC, 2019 Publication	<1%
82	Submitted to Coventry University Student Paper	<1%
83	Sean R. Ford, Peter Labak. "An Explosion Aftershock Model with Application to On-Site Inspection", Pure and Applied Geophysics, 2015 Publication	<1%
84	Thystere Matondo Bantidi, Takeshi Nishimura. "Spatio-temporal clustering of successive earthquakes as inferred from analyses of global CMT and NIED F-net catalogs", Earth, Planets and Space, 2022 Publication	<1%
85	Submitted to University of Hong Kong Student Paper	<1%
86	s3-eu-west-1.amazonaws.com Internet Source	<1%
87	seismica.library.mcgill.ca Internet Source	<1%
88	www.science.org	

Internet Source

<1%

89	Submitted to Indian Institute of Technology Student Paper	<1%
90	Khorshidi, Samira. "Adversarial Attacks and Defense Mechanisms to Improve Robustness of Deep Temporal Point Processes.", Purdue University, 2023 Publication	<1%
91	Kresin, Conor Joseph. "Applications and Properties of Point Processes", University of California, Los Angeles, 2023 Publication	<1%
92	bora.uib.no Internet Source	<1%
93	wrap.warwick.ac.uk Internet Source	<1%
94	Geerlings, Elmar. "Origin of Power-Law Behaviour in the Size Distribution of Extreme Events of Gross Primary Productivity", Universidade de Lisboa (Portugal), 2022 Publication	<1%
95	Gourab Saha, Md Toki Tahmid, Md. Shamsuzzoha Bayzid. "EmbedSimScore: Advancing Protein Similarity Analysis with	<1%

Structural and Contextual Embeddings", Cold Spring Harbor Laboratory, 2024
Publication

96	Malcolm C. A. White, Yehuda Ben-Zion, Frank L. Vernon. "A Detailed Earthquake Catalog for the San Jacinto Fault-Zone Region in Southern California", Journal of Geophysical Research: Solid Earth, 2019 Publication	<1 %
97	Wahyu Triyoso. "Applying the Akaike Information Criterion (AIC) in earthquake spatial forecasting: a case study on probabilistic seismic hazard function (PSHF) estimation in the Sumatra subduction zone", Geomatics, Natural Hazards and Risk, 2024 Publication	<1 %
98	d197for5662m48.cloudfront.net Internet Source	<1 %
99	pubs.geoscienceworld.org Internet Source	<1 %
100	www.jstatsoft.org Internet Source	<1 %
101	Molyneux, James. "Estimation of Spatial-Temporal Hawkes Models for Earthquake Occurrences.", University of California, Los Angeles, 2018 Publication	<1 %

102	Olivier C. Pasche, Valérie Chavez-Demoulin, Anthony C. Davison. "Causal modelling of heavy-tailed variables and confounders with application to river flow", Extremes, 2022 Publication	<1 %
103	gfzpublic.gfz-potsdam.de Internet Source	<1 %
104	pangea.stanford.edu Internet Source	<1 %
105	portal.research.lu.se Internet Source	<1 %
106	www.reaktproject.eu Internet Source	<1 %
107	Submitted to Oklahoma State University Student Paper	<1 %
108	Qiang Zhang, Aldo Lipani, Emine Yilmaz. "Learning Neural Point Processes with Latent Graphs", Proceedings of the Web Conference 2021, 2021 Publication	<1 %
109	Submitted to UM, University College Student Paper	<1 %
110	Submitted to University of Maryland, University College Student Paper	<1 %

111	icml.cc Internet Source	<1%
112	Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin. "Bayesian Data Analysis", Chapman and Hall/CRC, 2019 Publication	<1%
113	Gossett, Derreck John. "Prevalence and Characteristics of Seismic Magnitude Clustering", Miami University, 2023 Publication	<1%
114	www.nature.com Internet Source	<1%
115	Haoting Zhang, Donglin Zhan, James Anderson, Rhonda Righter, Zeyu Zheng. "Clustering Then Estimation of Spatio-Temporal Self-Exciting Processes", INFORMS Journal on Computing, 2024 Publication	<1%
116	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
117	J. Woessner. "A retrospective comparative forecast test on the 1992 Landers sequence", Journal of Geophysical Research, 05/26/2011 Publication	<1%

118	Xiaoting Li, Christian Genest, Jonathan Jalbert. "A self-exciting marked point process model fordrought analysis", Environmetrics, 2021 Publication	<1 %
119	studylib.net Internet Source	<1 %
120	www.pge.com Internet Source	<1 %
121	Encyclopedia of Earth Sciences Series, 2011. Publication	<1 %
122	Junhyeon Kwon, Yingcai Zheng, Mikyoung Jun. "Flexible spatio-temporal Hawkes process models for earthquake occurrences", Spatial Statistics, 2023 Publication	<1 %
123	Submitted to Nexford Learning Solutions Student Paper	<1 %
124	Trugman, Daniel Taylor. "Deviant Earthquakes: Data-driven Constraints on the Variability in Earthquake Source Properties and Seismic Hazard.", University of California, San Diego, 2018 Publication	<1 %
125	dblp.dagstuhl.de Internet Source	<1 %
	rise-eu.org	

126	Internet Source	<1%
127	worldwidescience.org Internet Source	<1%
128	"Seismological Society of America San Francisco, California 100th Anniversary Earthquake Conference 18-22 April", Seismological Research Letters, 2006 Publication	<1%
129	Christian Molkenthin, Christian Donner, Sebastian Reich, Gert Zöller, Sebastian Hainzl, Matthias Holschneider, Manfred Opper. "GP-ETAS: semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model", Statistics and Computing, 2022 Publication	<1%
130	Nandan, Shyam, Guy Ouillon, Jochen Woessner, Didier Sornette, and Stefan Wiemer. "Systematic Assessment of the Static Stress-Triggering Hypothesis using Inter-earthquake Time Statistics : Assessing Static Triggering Hypothesis", Journal of Geophysical Research Solid Earth, 2016. Publication	<1%
131	Shi, Hui. "Neural-Symbolic Methods for Neural Architecture Design", University of	<1%

	California, San Diego, 2024 Publication	
132	ddescholar.acemap.info Internet Source	<1 %
133	www.eurecom.fr Internet Source	<1 %
134	Nanda, S. J., K. F. Tiampo, G. Panda, L. Mansinha, N. Cho, and A. Mignan. "A tri-stage cluster identification model for accurate analysis of seismic catalogs", Nonlinear Processes in Geophysics, 2013. Publication	<1 %
135	Wang, Yuexi. "Deep Approximate Bayesian Inference", The University of Chicago, 2023 Publication	<1 %
136	os.zhdk.cloud.switch.ch Internet Source	<1 %
137	papers.neurips.cc Internet Source	<1 %
138	people.math.aau.dk Internet Source	<1 %
139	riskcenter.ethz.ch Internet Source	<1 %
140	summit.sfu.ca Internet Source	<1 %

141	uu.diva-portal.org Internet Source	<1 %
142	Alireza Azarbakht, Hossein Ebrahimian, Fatemeh Jalayer, John Douglas. "Variations in hazard during earthquake sequences between 1995 and 2018 in western Greece as evaluated by a Bayesian ETAS model", Geophysical Journal International, 2022 Publication	<1 %
143	Barbel Finkenstadt, Leonhard Held, Valerie Isham. "Statistical Methods for Spatio-Temporal Systems", Chapman and Hall/CRC, 2019 Publication	<1 %
144	Bhattacharjee, Robi. "Theoretical Foundations of Trustworthy Machine Learning", University of California, San Diego, 2023 Publication	<1 %
145	Contessi, Silvio. "Aggregate implications of firm heterogeneity in open economies", Proquest, 20111004 Publication	<1 %
146	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023) - Proceedings of the International Conference	<1 %

on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024
Publication

---

147 Donald B. Owen. "Statistics of Quality", CRC Press, 2020 <1%
Publication

---

148 Edward H. Field, Kevin R. Milner, Morgan T. Page, William H. Savran, Nicholas van der Elst. "Improvements to the Third Uniform California Earthquake Rupture Forecast ETAS Model (UCERF3-ETAS)", The Seismic Record, 2021 <1%
Publication

---

149 Frederic Paik Schoenberg. "On the relationship between lower magnitude thresholds and bias in epidemic-type aftershock sequence parameter estimates", Journal of Geophysical Research, 04/08/2010 <1%
Publication

---

150 Giuseppe Petrillo, Jiancang Zhuang. "Bayesian earthquake forecasting approach based on the epidemic type aftershock sequence model", Earth, Planets and Space, 2024 <1%
Publication

---

151 Min Liu, Miao Zhang, Weiqiang Zhu, William L. Ellsworth, Hongyi Li. "Rapid Characterization of the July 2019 Ridgecrest, California, <1%

Earthquake Sequence From Raw Seismic Data Using Machine-Learning Phase Picker", Geophysical Research Letters, 2020
Publication

152	P. K. Srijith, Michal Lukasik, Kalina Bontcheva, Trevor Cohn. "Longitudinal Modeling of Social Media with Hawkes Process Based on Users and Networks", Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17, 2017 Publication	<1 %
153	Submitted to University of Warwick Student Paper	<1 %
154	Zachary E Ross, Yisong Yue, Men-Andrin Meier, Egill Hauksson, Thomas H Heaton. "PhaseLink: A Deep Learning Approach to Seismic Phase Association", Journal of Geophysical Research: Solid Earth, 2019 Publication	<1 %
155	Zhou, Xinyu. "Statistical Depth in Point Process and Its Applications", The Florida State University, 2024 Publication	<1 %
156	ebin.pub Internet Source	<1 %
157	eprints.lancs.ac.uk Internet Source	

		<1%
158	lib.buet.ac.bd:8080 Internet Source	<1%
159	publishup.uni-potsdam.de Internet Source	<1%
160	reports-archive.adm.cs.cmu.edu Internet Source	<1%
161	www.drquigs.com Internet Source	<1%
162	www.gordonjross.co.uk Internet Source	<1%
163	"Machine Learning and Knowledge Discovery in Databases: Research Track", Springer Science and Business Media LLC, 2023 Publication	<1%
164	"Seismicity Patterns, their Statistical Significance and Physical Meaning", Springer Science and Business Media LLC, 1999 Publication	<1%
165	Amr S. Elnashai, Luigi Di Sarno. "Fundamentals of Earthquake Engineering", Wiley, 2008 Publication	<1%

166	Andrew B. Lawson, David G.T. Denison. "Spatial Cluster Modelling", Chapman and Hall/CRC, 2019 Publication	<1 %
167	Atila Abdulkadiroğlu, Parag A. Pathak, Christopher R. Walters. "Free to Choose: Can School Choice Reduce Student Achievement?", American Economic Journal: Applied Economics, 2018 Publication	<1 %
168	Chenlong Li, Zhanjie Song, Wenjun Wang. "Space–time inhomogeneous background intensity estimators for semi-parametric space–time self-exciting point process models", Annals of the Institute of Statistical Mathematics, 2019 Publication	<1 %
169	David Vere-Jones. "Some models and procedures for space-time point processes", Environmental and Ecological Statistics, 06/2009 Publication	<1 %
170	Dong, Elisa. "Testing Aftershock Forecasts Using Bayesian Methods", The University of Western Ontario (Canada), 2022 Publication	<1 %

171	Eweis-LaBolle, Jonathan Tammer. "A Manifold Learning Approach for Inverse Problems", University of California, Irvine, 2024 Publication	<1 %
172	James R. Holliday, Donald L. Turcotte, John B. Rundle. "Self-similar branching of aftershock sequences", Physica A: Statistical Mechanics and its Applications, 2008 Publication	<1 %
173	Jesper Møller, Mohammad Ghorbani, Ege Rubak. "Mechanistic spatio-temporal point process models for marked point processes, with a view to forest stand data", Biometrics, 2016 Publication	<1 %
174	Olhede, Sofia. "Statistical Applications of Wavelets", Encyclopedia of Complexity and Systems Science, 2009. Publication	<1 %
175	Qu, Helen. "Towards Precision Photometric Type Ia Supernova Cosmology With Machine Learning", University of Pennsylvania, 2024 Publication	<1 %
176	Rodolfo Console, Maura Murru, Giuseppe Falcone. "Earthquake Occurrence", Wiley, 2017 Publication	<1 %

177	Stefanie Seif, Arnaud Mignan, Jeremy Douglas Zechar, Maximilian Jonas Werner, Stefan Wiemer. "Estimating ETAS: The effects of truncation, missing data, and model assumptions", Journal of Geophysical Research: Solid Earth, 2017 Publication	<1 %
178	Steven Bradley Lowen, Malvin Carl Teich. "Fractal-Based Point Processes", Wiley, 2005 Publication	<1 %
179	Thiele, Leander F.. "Getting Ready for New Data: Approaches to Some Challenges in Cosmology", Princeton University, 2024 Publication	<1 %
180	Webb, . "Density Estimation - Bayesian", Statistical Pattern Recognition Webb/Statistical Pattern Recognition, 2011. Publication	<1 %
181	Zheng Xiaogu, David Vere-Jones. "Further applications of the stochastic stress release model to historical earthquake data", Tectonophysics, 1994 Publication	<1 %
182	argon.ess.washington.edu Internet Source	<1 %
183	docplayer.net Internet Source	<1 %

184	eartharxiv.org Internet Source	<1 %
185	eprints.soton.ac.uk Internet Source	<1 %
186	epub.ub.uni-muenchen.de Internet Source	<1 %
187	hal.archives-ouvertes.fr Internet Source	<1 %
188	hugepdf.com Internet Source	<1 %
189	munin.uit.no Internet Source	<1 %
190	ndl.ethernet.edu.et Internet Source	<1 %
191	open.library.ubc.ca Internet Source	<1 %
192	qmro.qmul.ac.uk Internet Source	<1 %
193	vdoc.pub Internet Source	<1 %
194	web.archive.org Internet Source	<1 %
195	www.arxiv-vanity.com Internet Source	<1 %

196	www.duo.uio.no Internet Source	<1 %
197	www.fedoa.unina.it Internet Source	<1 %
198	www.ijcai.org Internet Source	<1 %
199	www.scribd.com Internet Source	<1 %
200	Advanced Disassembly Planning, 2014. Publication	<1 %
201	Djorno, Alexandra. "Extension of the Hawkes Process for Modeling Crowdfunding Platform Dynamics", Yale University, 2024 Publication	<1 %
202	Govind Waghmare, Ankur Debnath, Siddhartha Asthana, Aakarsh Malhotra. "Modeling Inter-Dependence Between Time and Mark in Multivariate Temporal Point Processes", Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022 Publication	<1 %
203	Haoran Meng, Jeffrey J. McGuire, Yehuda Ben-Zion. "Semiautomated Estimates of Directivity and Related Source Properties of Small to Moderate Southern California Earthquakes	<1 %

Using Second Seismic Moments", Journal of Geophysical Research: Solid Earth, 2020
Publication

---

204 Kadambi, Pradyumna. "Robust Networks: Neural Networks Robust to Quantization Noise and Analog Computation Noise Based on Natural Gradient", Arizona State University, 2020 <1%
Publication

---

205 Kelian Dascher-Cousineau, Oleksandr Shchur, Emily E. Brodsky, Stephan Günnemann. "Using Deep Learning for Flexible and Scalable Earthquake Forecasting", Geophysical Research Letters, 2023 <1%
Publication

---

206 Patrick J. Laub, Young Lee, Thomas Taimre. "The Elements of Hawkes Processes", Springer Science and Business Media LLC, 2021 <1%
Publication

---

207 Rubanova, Yulia. "Continuous-Time Latent-Variable Models for Time Series", University of Toronto (Canada), 2020 <1%
Publication

---

208 Shyam Nandan, Guy Ouillon, Didier Sornette. "Are Large Earthquakes Preferentially Triggered by Other Large Events?", Journal of Geophysical Research: Solid Earth, 2022 <1%

Publication

209	Tao, Long. "Contributions to Statistical Analysis Methods for Neural Spiking Activity.", Boston University, 2018 Publication	<1 %
210	"Machine Learning, Optimization, and Data Science", Springer Science and Business Media LLC, 2022 Publication	<1 %
211	Encyclopedia of Complexity and Systems Science, 2009. Publication	<1 %
212	Hong, Chengkuan. "Deep Neyman-Scott Processes", University of California, Riverside, 2022 Publication	<1 %
213	Jacobsen, Christian S.. "Enhancing Physical Modeling with Interpretable Physics-Aware Machine Learning", University of Michigan, 2024 Publication	<1 %
214	Jiancang Zhuang, Maura Murru, Giuseppe Falcone, Yicun Guo. "An extensive study of clustering features of seismicity in Italy from 2005 to 2016", Geophysical Journal International, 2018 Publication	<1 %

215	Jingfang Fan, Jun Meng, Josef Ludescher, Xiaosong Chen et al. "Statistical physics approaches to the complex Earth system", Physics Reports, 2020 Publication	<1 %
216	Jingfang Fan, Jun Meng, Josef Ludescher, Xiaosong Chen et al. "Statistical physics approaches to the complex Earth system", Physics Reports, 2021 Publication	<1 %
217	Junru Ren, Shaomin Wu. "Prediction of user temporal interactions with online course platforms using deep learning algorithms", Computers and Education: Artificial Intelligence, 2023 Publication	<1 %
218	Manisha Dubey, Ragja Palakkadavath, P. K. Srijith. "Bayesian neural hawkes process for event uncertainty prediction", International Journal of Data Science and Analytics, 2023 Publication	<1 %
219	Mohammadamin Sedghizadeh, Robert Shcherbakov. "The Analysis of the Aftershock Sequences of the Recent Mainshocks in Alaska", Applied Sciences, 2022 Publication	<1 %

220	Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, Yong Li. "Spatio-temporal Diffusion Point Processes", Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023 Publication	<1 %
221	core.ac.uk Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		