TESI DI DOTTORATO DI RICERCA

IN

SCIENZE DELLA TERRA, DELL'AMBIENTE E DELLE RISORSE

XXXV Ciclo

Materia della Tesi:

Geofisica

Titolo

Enhancing earthquake forecasting through innovative stochastic modelling

Candidato

Ester Manganiello

DR993806

Relatore

Prof. Warner Marzocchi

Correlatore

Dr. Marcus Herrmann

Coordinatore del Corso di Dottorato di Ricerca

Prof.ssa Rosa Di Maio

“Bisogna avere ancora caos
dentro di sé per partorire
una stella danzante.”

“One must still have chaos in
oneself to be able to give
birth to a dancing star.”

“Man muss noch Chaos in sich
haben, um einen tanzenden
Stern gebären zu können.”

(Friedrich Wilhelm
Nietzsche, Also sprach
Zarathustra, 1967–77)

Abstract

The possibility to predict earthquakes has always been a hotly debated topic in the scientific community due to the destructive power of these events.

Although we are not (yet?) able to predict deterministically with high precision where, when, and how big the next earthquake will struck, currently, several models are able to state, in probabilistic terms, that a seismic event can occur in a certain place, time interval and with a certain magnitude. These models have a big limitation, which consists of the inability to forecast large earthquakes with high probabilities. In fact, these models, such as ETAS model (Epidemic Type Aftershock Sequence), assume the independence of the magnitude: the magnitude of the next earthquake is randomly sampled from an exponential-like distribution, bounding the probability of large earthquake to the sampling of the right tail of such distribution.

In this thesis, we aim to provide new insights on this problem that can be useful to improve the probabilistic forecasting models.

Specifically, we analyze foreshocks, i.e., events that precede an event of higher magnitude: the goal is to understand if they have characteristics such as to be recognized a-*priori* (i.e., before the occurrence of a large seismic event), thus distinguishing them from the rest of seismicity. If so, their *predictive* power could be used to improve the probability of larger events.

In this thesis, we use ETAS model as null hypothesis, i.e., able to represent all seismicity, and we compare the foreshock sequences of synthetic and real catalogues. The results highlight discrepancies between what the model is able to predict and what we observe in reality. However, these discrepancies do not concern all foreshock sequences, but only seismic sequences with peculiar

characteristics: they are located mainly in areas of high heat flow, with a small/medium mainshock magnitude, and characterized by a high number of foreshocks. These sequences are commonly referred in the literature as swarms. Foreshock sequences occurring in low heat flux zone conform well with what expected by the ETAS model.

To further investigate on the difference between earthquake sequences, we use a scalar parameter called ALD (Average Leaf Depth). This parameter is able to quantitatively represent the topological structure of the sequences, i.e., *swarm* type (with high ALD values) or burst type (the earthquake sequence explained by the ETAS model, with low ALD values). We test the null hypothesis that ETAS represents well the reality and therefore, that it is also able to reproduce seismic sequences with high ALD values (i.e., *swarm*). Comparing the ALD values of real sequences and synthetic sequences we obtained highly variable results depending on the specific ETAS model used and the type of source. This makes ALD a parameter to be used with extreme care and its variability makes it not so useful for the discretization of seismic sequences (i.e., between *swarm* and aftershock-type sequences).

As in the literature there is no criterion to objectively differentiate the types of seismicity, we present a new methodology based on machine learning. First of all, we collect several information that characterize the seismic sequences in terms of space, time, and magnitude. We apply a Cluster Analysis to the collected dataset and identify the optimal number of clusters into which to divide the sequences. There are two predominant clusters that differ from a physical and statistical point of view: one group is mainly located in areas of high heat flow, is characterized by a high productivity and rate of foreshocks, therefore comparable to *swarm*-type sequences. Conversely, the other group has characteristics attributable to aftershock-type sequences.

In conclusion, the above-mentioned method could be useful for recognizing the type of a seismic sequence (i.e., *swarm* or aftershock type), collecting all the

information available on it (ALD, heat flow value corresponding to its location, productivity, etc.). If we were able to apply these methodologies in the short term and identify *swarm*-type sequences, we could reduce the probability that a large seismic event could occur imminently, thus improving the probability estimates of current models.

Sommario

La possibilità di prevedere i terremoti è sempre stato un argomento ampiamente dibattuto nella comunità scientifica a causa del loro potenziale distruttivo.

Allo stato attuale non siamo (ancora?) in grado di prevedere *deterministicamente* e con elevata precisione dove, quando e quanto sarà grande il prossimo terremoto; tuttavia, esistono diversi modelli in grado di affermare, in termini *probabilistici*, che un evento sismico possa verificarsi in un determinato luogo, intervallo di tempo e con una certa grandezza. Purtroppo, questi modelli, come quello ETAS (Epidemic Type Aftershock Sequence), hanno un grande limite che consiste nell'impossibilità di prevedere grandi terremoti con alte probabilità. Le ragioni di questa limitazione vanno ricercate nell'assunzione di indipendenza della magnitudo: la magnitudo di un evento non dipende da quella degli eventi passati. In questi modelli, infatti, la magnitudo di un terremoto futuro viene campionata casualmente da una distribuzione di tipo esponenziale, limitando la probabilità di un grande terremoto al campionamento della coda destra di tale distribuzione.

In questo lavoro di tesi ci poniamo l'obiettivo di ovviare a questo problema, fornendo delle skills che possano essere utili per migliorare la previsione probabilistica dei modelli.

A tal proposito si è pensato di analizzare i foreshocks, ovvero gli eventi sismici che precedono una grande magnitudo: l'obiettivo è quello di capire se essi abbiano delle caratteristiche tali da essere riconosciuti a-*priori* (ovvero, prima dell'accadimento di un grande evento sismico), discostandosi in tal modo dal resto della sismicità. Se così fosse, il loro potenziale *premonitore* potrebbe essere utilizzato per migliorare la stima delle probabilità di accadimento di eventi più grandi.

In questa tesi abbiamo utilizzato il modello ETAS come ipotesi nulla, ovvero in grado di rappresentare tutta la sismicità, e abbiamo confrontato le sequenze di foreshocks di cataloghi sintetici e del catalogo reale. I risultati evidenziano delle discrepanze tra ciò che il modello è in grado di predire e ciò che osserviamo nella realtà. Tuttavia, tali discrepanze non concernono tutte le sequenze di foreshocks, ma soltanto sequenze sismiche con delle caratteristiche peculiari: esse si localizzano principalmente nelle zone di alto flusso di calore, hanno una magnitudo del mainshock medio-bassa, e sono caratterizzate da un elevato numero di foreshocks. Tali sequenze vengono comunemente definite in letteratura come *swarm*. Le sequenze di foreshocks che avvengono nelle zone di basso flusso sono conformi a quanto previsto dal modello ETAS.

Per approfondire ulteriormente le differenze tra le tipologie di sequenze sismiche, utilizziamo un parametro scalare definito ALD (Average Leaf Depth). Tale parametro è in grado di rappresentare quantitativamente la struttura topologica delle sequenze, se di tipo *swarm* (con alti valori di ALD) o di tipo *burst* (ovvero le sequenze sismiche spiegate dal modello ETAS, con bassi valori di ALD). Noi testiamo l'ipotesi nulla che ETAS rappresenti bene la realtà e dunque, che sia anche in grado di riprodurre sequenze sismiche con alti valori di ALD (cioè di tipo *swarm*). Confrontando i valori di ALD di sequenze reali e di sequenze sintetiche abbiamo ottenuto risultati molto variabili a seconda dello specifico modello ETAS utilizzato e del tipo di sorgente. Questo rende ALD un parametro da utilizzare con estrema cautela e la sua variabilità lo rende poco utile per la discretizzazione delle sequenze sismiche (ovvero tra *swarm* e sequenze di tipo aftershock).

Partendo dal presupposto che in letteratura non esiste un criterio per differenziare i tipi di sismicità in maniera oggettiva, presentiamo una nuova metodologia basata sul machine learning. In primis raccogliamo diverse informazioni che caratterizzino le sequenze sismiche in termini di spazio, tempo e magnitudo. Al dataset collezionato applichiamo una *Cluster Analysis* e

individuiamo il numero ottimale di clusters in cui dividere le sequenze. Due sono i clusters predominanti che si differenziano da un punto di vista fisico e statistico: un gruppo si localizza principalmente in zone di alto flusso di calore, è caratterizzato da un elevata produttività e rate dei foreshocks, dunque assimilabile a sequenze di tipo *swarm*. Viceversa, l'altro gruppo presenta caratteristiche ascrivibile alle sequenze di tipo aftershocks.

In conclusione, il metodo sopra-citato potrebbe essere utile per riconoscere la tipologia di una sequenza sismica (cioè se di tipo *swarm* o di tipo aftershocks), raccogliendo tutte le informazioni che si hanno su di essa (ALD, valore di heat flow corrispondente alla sua localizzazione, produttività, ecc.).

Se riuscissimo ad applicare tali metodologie nel breve termine e a identificare sequenze di tipo *swarm* potremmo ridurre la probabilità che un grande evento sismico possa verificarsi in maniera imminente, dunque migliorando le stime di probabilità degli attuali modelli.

Acknowledgements

First of all, I thank Warner Marzocchi, my supervisor, for having believed in me from the first moment. Many thanks for everything you taught me in these years, both personally and professionally. I really appreciate your patience and effort in helping me, always super-available for any problem that arises. Thank you for all the opportunities you gave me, the projects we've done together, the congress I've had the honor of participating in, and all the scientists I've had the pleasure of meeting. Thank you because your teachings go beyond this PhD thesis, scientific papers, posters, and so on; you know how to spread the love you have for your work and inspire young students towards science and being curious. I'm really happy to have learned more than what I expected.

I thank my super co-tutor, Marcus Herrmann. First, you are not only one of the best teachers I could ever have, an amazing researcher, and a heedful hearted scientist, but you are also a great friend. The patience you have had with me is something incredible. This thesis would never have been what it is now without your constant help and support. I owe you a lot! I know, I have already thanked you many times, but I think it will never be enough. I'm really happy to have you by my side. Thanks also for our teas, walks, stimulating discussion, endless calls to solve problems, and the laughs with Fabio. Thank you for your advice on technology, science, cooking, delicious pizzerias, and fantastic places. Thanks also for introducing me to Antonella, a hurricane of joy and happiness.

I want to thank Ian Main, my supervisor in Edinburgh, for giving me the opportunity to study there. Thank you for your kindness, sweetness, and helpfulness. Besides to having learned a lot from a scientific point of view, I also had the opportunity to live in a wonderful place, Scotland, always in my heart. Thanks for introducing me to fantastic people like Mark and Francesco,

they are amazing colleagues! Thank you also for always having a comforting word for me!

I also want to thank my best friend from Edinburgh, Farnaz. You welcomed me and accompanied me on this journey like a great friend. Thank you for sharing your Iranian culture with me. I'm back home with the heart full of joy, hoping to see you soon!

I thank the two reviewers of this thesis that contribute to improve it a lot! Thanks for your suggestion, comments, and food for thought.

Switching to Italian…voglio ringraziare papà e mamma che insieme mi danno una forza incredibile. Sono un ancora di salvezza, un porto sicuro. (Partendo dal presupposto che non vorrei mai essere genitore di una figlia come me) vi sono immensamente riconoscente e grata. Sono anni che siete al mio fianco, sempre pronti a darmi un consiglio, una parola di conforto nei momenti bui, e a gioire delle mie piccole conquiste. Come ben sapete, non sarà il dottorato a far cambiare il mio essere; dunque, continuate ad essere armati di sopportazione e non perdete mai il vostro entusiasmo, perché per me è fonte di salvezza. Vi voglio bene… Grazie!

Ringrazio tutti i mie amici del gruppo squilibrio (Marika (la miss capobanda), Duty (perchè io così ti chiamerò per sempre,o moglie), Anto (terribile dolcezza), Giovi (l'amore di sempre), Titti (la mia uno-due storica, non solo nel calcio), Robi (la mia irpina prefe), David (l'ancora di salvezza delle notti brave). Non so quale congiunzione astrale ci abbia unito, ma non c'è niente al mondo che mi renda più felice e spensierata dello stare insieme a voi.

Ringrazio le famosissime, Ale & Debora. Sono felice di avervi al mio fianco da tutto questo tempo, e spero di continuare a rincorrervi sempre intorno al mondo, senza perderci mai di vita, così come facciamo da più di 20 anni. Vi voglio bene.

Ringrazio l'amica e super-collegue Paola, il mio spiraglio di luce del Distar. Con te si è chiuso un cerchio di tre anni fantastici. Grazie per le nostre infinite

chiacchierate, il tuo costante supporto, la tua schiettezza e dolcezza. Looking to see Corrado & Manganiello su qualche pagina in giro nel mondo. Ti voglio bene.

In fine, ringrazio Mario, come se un semplice ringraziamento potesse bastare. Conosci tutte le lacrime e tutti i sorrisi racchiusi in questo libricino; sempre orgoglioso di me, nonostante tutto. Questo lavoro non si sarebbe mai concluso se non ci fossi stato tu al mio fianco, costantemente. Grazie perché tutto è iniziato da te e con te. Questo lavoro, il più lungo, difficile e bello che io abbia mai fatto, è tutto dedicato a te.

Contents

Abstract I

Sommario IV

Acknowledgements VII

1.Introduction 1-4

1.1 Earthquake forecasting models 1-6

1.1.1 ETAS: Epidemic type aftershock sequence model 1-6

1.1.2 ETAS parameters and their estimation 1-9

1.2 Do forecasting models need to be improved? 1-11

1.3 Thesis outline 1-12

2.New physical implications from revisiting foreshock activity in southern California 2-15

2.1 Introduction 2-16

2.2 Data and Methods 2-18

2.3 Results 2-22

2.3.1 Testing for anomalous foreshock activity 2-22

2.3.2 Correlating foreshock sequences with the heat flow 2-25

2.4 Discussion & Conclusion 2-28

3.A useful scalar measure for improving earthquake forecasting: the Average Leaf Depth 3-31

3.1 Introduction 3-31

3.2 Method 3-33

3.3 Results 3-35

3.4 Discussion and Conclusion 3-39

4.What makes an earthquake sequence a swarm? Using Cluster analysis to discriminate earthquake sequences in southern California 4-41

4.1 Introduction 4-42

4.2 Data & Method 4-45

4.2.1 Collected dataset 4-45

4.2.2 k-means Cluster Analysis 4-46

4.2.3 Statistical properties of identified clusters 4-47

4.2.4 Leave one out method 4-49

4.3. Results 4-52

4.3.1 Results of cluster analysis using the complete dataset 4-52

4.3.2 Physical interpretation of identified clusters 4-55

4.3.3 Results of cluster analysis using the most influential subset 4-59

4.3.4 Physical interpretation of identified clusters using most important features 4-65

4.4 Discussion and Conclusion 4-68

5.Conclusion 5-72

5.1 Outlook 5-74

SI.1 Supporting Information for Chapter 2 5-76

SI.1.1 Used ETAS models 5-76

SI.1.2 Verifying the reliability of the ETAS models 5-77

SI.1.3 Alternative analysis of TEST1 5-78

SI.1.4 Results for a significance level of 95% 5-78

SI.2 Supporting Information for Chapter 4 5-98

List of Figures 5-109

List of Tables 5-117

Bibliography 5-118

1

Introduction

The word *science* is derived from the Latin *scientia* which means *knowledge*. It is used in the motto *Scientia Potentia Est*, a Latin aphorism meaning *knowledge is power*, currently attributed to the philosopher Thomas Hobbes. According to his feeling, the true knowledge is based on causes, so that to understand a phenomenon means to be able to explain how it is produced (Jesseph, 2010). Knowledge being power is understood as *success from ability* (Greco, 2010): *"acting from a position of knowledge – versus ignorance – permits us to successfully navigate the perils of the external world"* (Farrelly, 2021).

From this point of view, the main goal of doing science is to *learn* how the physical and natural world works. We have always been surrounded by incredible and mysterious natural phenomena that arouse the curiosity of scientists and philosophers. Among these phenomena, earthquakes are a natural force that fascinates the scientific community, mainly for their power and for the enormous impact these phenomena have on our lives; in fact, earthquakes are the cause of many damages and countless victims. Therefore, the reason why we need to study this natural phenomenon is twofold: on the one hand, we are driven by the curiosity to understand what exactly happens under our feet, and, on the other hand, we want to use this information to limit the damage that they can cause. Understanding a phenomenon requires to observe it, study it, capture principal forces, and reproduce it through physical and statistical laws; in this spirit, *knowing* also means *predicting*, which in this case is synonymous with *saving*.

In the past, scientists tried to predict, in a deterministic way, exactly the hour, the time, and the magnitude of very destructive earthquakes, analyzing several precursors, i.e., signals that are observable before earthquakes and indicate an impending event (Jordan et al., 2011). The grim presence of such information has led to a long series of failures; for this reason, the idea of being able to predict earthquakes through this approach remains an unrealistic dream for the current knowledge.

Seismologists had the brilliant idea of collecting laws and relationships that characterize earthquakes to develop models that make probabilistic forecasts. The *probabilistic* approach of these mathematical models is the key word that differs considerably from the deterministic approach. Such probabilistic *earthquake forecasting models* (see next section) provide probabilities for the occurrence of earthquakes in time and space for a given magnitude (Jordan et al., 2011). Such models can be tested against the observation to achieve important insights about the veracity of the used relationships.

Making reliable forecasts requires understanding as much as possible about the spatiotemporal behavior of earthquakes and the underlying processes that drive them. Despite many years of research, methods for characterizing the dynamics of seismicity are lacking. Moreover, one of the answered questions is whether every earthquake follows the same laws or whether particular patterns exist that reflect different governing properties (Seif et al., 2019).

In the present PhD thesis, I will address these problems by comparing what we observe in reality with the output of earthquake forecasting models. I briefly described relevant concepts in the following Section 1.1. The main motivation is to find new insights that can be useful to enhance earthquake forecasting, as for avoiding earthquake losses and improving community preparedness (Jordan et al., 2011). I also define several methods to explore different types of seismicity and explore the physical reason behind their occurrences.

1.1 Earthquake forecasting models

An earthquake forecasting model is a systematic method for calculating the probabilities of event occurrence within a space-time-magnitude domain; forecasting intervals can range from long-term (several years to many decades), medium-term (months to years), and short-term (few months or less) (Jordan et al., 2011). They can be set up for certain regions (Felzer et al., 2002; Seif et al., 2017) or globally (e.g., Bird et al. 2015) and can be based on statistical (e.g., Ogata, 1988, Reasenberg and Jones 1989; see next section) or physical (e.g., Mancini et al., 2019) approaches. Currently, the most useful information to create such models comes from real observation of events in different domains (time, space, magnitude, tectonic).

Concerning a statistical approach, several models exist that differ according to the assumptions made on the observed seismicity. Among them, earthquake triggering models attempt to represent the properties of earthquake clustering through empirical relationship (Jordan et al., 2011). In these models, any earthquake can trigger one or more other earthquakes, generating its own aftershocks. This subclass includes Epidemic-Type Aftershock Sequence (ETAS) whose structure is detailed in the following subsection.

1.1.1 ETAS: Epidemic type aftershock sequence model

ETAS is one of the most commonly used *probabilistic* and *statistical* earthquake forecasting models. It was introduced by Ogata (1988) and belongs to a class of branching point process models known in the statistical literature as *Hawkes* or *self-exciting* point processes: an independent event (*background*) may *excite* the other events (called *triggered* events), starting their own branch. An important aspect of this model is that any event can trigger a larger event, i.e., a small-

magnitude event (like M2.0) can trigger a M7.0. All linked earthquakes create an earthquake sequence (also called seismic cluster) as shown in Figure 1.1.

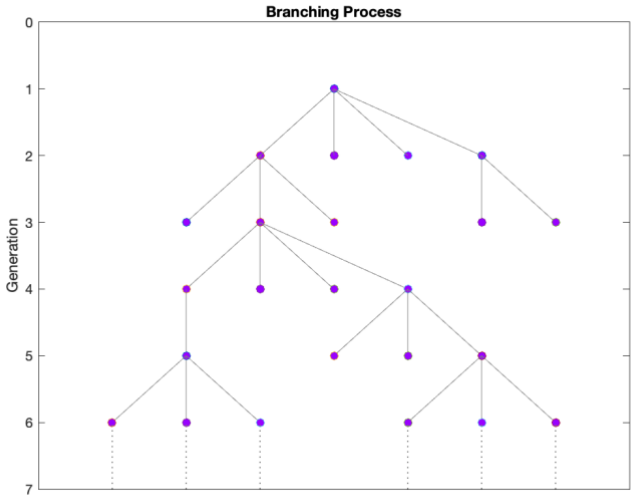


Figure 1-1 Simple representation of self-exciting point processes

From a mathematical point of view, the seismicity of ETAS can be described by the conditional intensity function that quantifies the earthquake occurrence rate λ at time t and location x, y with mainshock M, given the history of past events H_t (eq. 1, Seif et al., 2019). It is formed by three main components: the *background* seismicity (independent events), the *triggered* seismicity (*dependent* events) and the magnitude frequency distribution of events.

$$\lambda(t,x,y,M|H_t) = [\mu(x,y) + \sum_{j:t_j<t} g\big(t-t_j, x-x_j, y-y_j;\ M_j\big)] * f(m) \quad (1)$$

Background rate — Triggered function — Magnitude frequency distribution

1) The first part is the spatially heterogenous background rate: it can be described by a Poisson process that is stationary in time:

$$\mu\,(x,y) = \nu\,\rho(x,y) \quad (2)$$

where ν is a positive-valued parameter.

Events occur independently and may be interpreted as caused by the same underlying process (e.g., tectonic movements). The term *spatially heterogenous* means that background events are best represented by a heterogeneous distribution, since earthquakes cluster in space and occur preferably on or near faults (Seif et al., 2017).

A preliminary step to compute the background rate is to identify events as background using a declustering procedure (Seif et al., 2017; Zhuang, 2011; Van Stiphout et al., 2012).

2) The second part describes the triggering effect: aftershocks can be triggered by the background events or previous aftershocks. The triggered function $g(t,x,y|H_t)$ can be represented by three mains statistical relationships (Seif et al., 2019):

$$g(t,x,y|H_t) = \underbrace{K_0 e^{\alpha(M-M_{cut})}}_{\text{Productivity law}} * \underbrace{c^{p-1}(t+c)^{-p}(p-1)}_{\text{Omori (modified) laws}} * \underbrace{f(x,y|M)}_{\text{Power law spatial decay}} \quad (3)$$

i) the Productivity law describes the number of events triggered by a specific magnitude event M (i.e., his aftershocks), ii) the Omori law describes the decreasing temporal occurrence of aftershocks in time, and iii) the Power law spatial decay describes the rate of aftershocks (number of earthquakes per unit area) with distance from the triggering earthquake with magnitude M. This last relationship can be determined as follows (Seif et al., 2019):

$$f(x,y|M) = \frac{q-1}{\pi D^2 e^{\gamma(m-m_0)}}\left(1+\frac{x^2+y^2}{D^2 e^{\gamma(m-m_0)}}\right)^{-q} \quad (4)$$

Eq. (3) and (4) are characterized by seven parameters ($K, \alpha, c, p, q, D, \gamma$; see next subsection) that have to be estimated from seismicity in the region under consideration (more in the next section).

3) The third part describes the magnitude distribution of earthquakes, which is considered separately because the magnitude of an event is not influenced by the history (i.e., past events). This assumption implies magnitude independence, stating that the magnitude is randomly drawn from the same magnitude-frequency distribution, no matter the triggering event magnitude (Stallone & Marzocchi, 2019). It can be described by the Gutenberg-Richter relation:

$$f(m) = \beta e^{(-\beta(m-m_{min}))} \quad (5)$$

where β is related to the b-value of the Gutenberg-Richter relation with $\beta = b * \log(10)$ and m_{min} is the lower cutoff magnitude.

These processes combined represent the spatiotemporal clustering of earthquakes. An important aspect in the "ETAS world" is that all earthquakes are treated in the same way because they are realizations of the same physical process. In fact, there is no feature that distinguishes events and any discrimination (i.e., foreshocks, mainshocks, and aftershocks) can only be made *a-posteriori*.

1.1.2 ETAS parameters and their estimation

An ETAS model is parameterized, i.e., fitted, to the earthquake catalog used for the analysis. Given the set of eight parameters $\theta = (\nu, K, \alpha, c, p, q, D, \gamma)$, I here describe what they represent and how to determine them.

The first parameter ν is related to the background seismicity. Once background events have been identified, the integral of $\mu(x, y)$ over the considered area represents the number of background events N_{GB} (Seif et al., 2017), defining ν as:

$$\nu = \frac{N_{GB}}{\int_{x,y} \rho(x,y) dxdy} \quad (6)$$

The parameters related to the triggered seismicity are $K, \alpha, c, p, q, D, \gamma$.

- K, α are related to the productivity law: K quantifies the capability of triggering new earthquakes, no matter what the magnitude of the triggering event is; α represents the scaling for the capacity to trigger earthquakes as function of magnitude (Seif et al., 2017).
- c, p describe the parameters of the Omori law, i.e., the decay of aftershock occurrence with time: c accomplishes a smoothed start of the power law decay, to avoid uniqueness of mainshock at $t = 0$; p describes the "speed" of the decay, which becomes faster as p increases.
- q, D are spatial kernel parameters and γ is the aftershock scaling factor; all of them control the spatial decay of aftershocks.

All parameters $\theta = (\nu, K, \alpha, c, p, q, D, \gamma)$ have to be estimated jointly, but such procedure is a hassle because they can be strongly biased, correlated, and unstable (Seif et al., 2017; Harte, 2015). Several alternative approaches have been proposed to try to deal with this issue, such as the commonly used Maximum Likelihood Estimation (MLE) method, which determines the set of parameters θ that maximize the log-likelihood function:

$$logL(\theta) = \sum_{k=1}^{N} log\lambda_\theta(t_k, x_k, y_k \,|H_{t_k}) - \int_0^T \iint_s \lambda_\theta(t, x, y \,|H_t) dx\, dy\, dt \quad (7)$$

where the subscript k runs over all events occurring in the study region s and the study time interval $[0, T]$ (Zhuang et al., 2002).

Other studies propose the use of Bayesian inference for parameter estimation (i.e., using prior of θ parameters), as the work of Ross (2021): this strategy reduces the correlation among parameters (through a latent variable) and considers their uncertainty when forecasting future earthquakes. Petrillo &

Zhuang (2022) also propose a Bayesian approach for parameter optimization that can be used for spatio-temporal ETAS model as well.

1.2 Do forecasting models need to be improved?

ETAS has shown good forecasting performance, making it an appropriate null, or reference, model, especially for aftershock sequences (Page and van der Elst, 2022). It is able to provide probabilistically where, when, and how earthquakes will occur on the short-term.

To validate forecasting methods, they need to be tested against the data prospectively (after the fact, i.e., data that is not available when creating a model), such as done in CSEP (Zechar et al., 2010; Schorlemmer et al., 2018). Testing assesses their reliability and skill, which provides valuable feedback for improving our understanding of involved processes. In these prospective tests, ETAS was found to be the best-performing model (Nanjo et al., 2012; Taroni et al., 2018). Specifically for the occurrence of foreshocks (i.e., events that occur before large events), several studies have retrospectively tested (before the fact, i.e., using existing data) the expectation of the ETAS model (Seif et al., 2019; Petrillo and Lippiello, 2021; Moutote et al., 2021). They assert that, although ETAS is describing seismicity generally well, there are still discrepancies.

These findings indicate that seismicity is influenced by processes that are not yet considered in the (classical) ETAS models. For example, such discrepancy is also proposed by the pre-slip model, in which foreshocks are diagnostic precursors as a byproduct of an aseismic slip that precedes large earthquakes. Such an anomalous behavior—if it can be detected in real time—may indicate an impending mainshock, and consequently increase the probability for a large earthquake. Gulia & Wiemer (2019) explore the possibility to recognize an on-going sequence as *aftershock*-sequence (i.e., the mainshock has already

happened) or foreshock-sequence of a future large event. Their findings suggest that the probability of a future larger event increases – after $M \geq 6$ – if the b-values are constant or falling sharply, in a well-defined box that mimics the source of the magnitude $M \geq 6$ event. However, there is considerable skepticism regarding the predictability using foreshocks, and no improvement has yet been made in this direction (Marzocchi & Zhuang, 2011).

Other studies (Zaliapin & Ben Zion, 2013, Chen et al., 2012, Enescu et al., 2009) interpreted these discrepancies as swarm-like sequences, although there is no accepted objective (quantitative) definition of what makes a sequence swarm-like. Only qualitative definitions exist, based on their "atypical" behavior compared to aftershock-type sequences (Mogi, 1963): no dominant mainshock, as the largest events are of similar magnitude (Ross et al., 2021); they predominantly occur in particular areas, such as in zones of high heat flow (Zaliapin and Ben Zion, 2013), volcanic areas, or reservoirs (Kato et al, 2010); and they are considered to be driven by movements of fluids in the subsoil (Vidale & Shearer, 2006), aseismic slip (Lohman and McGuire, 2007), or presence of gas. If there are statistical differences between the seismic sequences, we should be able to characterize them quantitatively. If so, recognizing an on-going sequence as a swarm, rather than as a foreshock-mainshock-aftershock-sequence, could greatly improve earthquake forecasting (decreased likelihood for a large mainshock).

All these findings suggest that there is potential for improving earthquake forecasting models and to better understand the dynamics of earthquake occurrence. With this thesis, I seek to better characterize the real seismicity and to contribute to enhancing the earthquake forecasting skill. I will address these potential advances in different ways in the following three chapters.

1.3 Thesis outline

This thesis consists of three main chapters; one is already published and the other two are in preparation for submission.

In Chapter 2, together with my supervisors, I evaluate the expectation of ETAS (i.e., synthetic catalogs) with seismicity data of southern California with a specific focus on foreshock sequences. Based on previous results (Seif et al., 2019; Petrillo and Lippiello, 2021; Moutote et al., 2021), we wondered why ETAS is unable to reproduce some foreshock sequences, and if this discrepancy can be explained with a different trigger mechanism of foreshocks. This chapter was recently published in GRL (Manganiello et al., 2023).

In Chapter 3, I come back to ETAS and again compare its synthetic catalogs with the reality, but from a different point of view: considering the entire sequences and using a scalar measure – the Average Leaf Depth. This measure is able to represent the topological structure of sequences. According to Zaliapin and Ben Zion (2013), which proposed this measure for cluster analysis, it should allow discriminating two different kinds of seismicity: burst-like (i.e., due to a mainshock) and swarm-like sequences. We explore which characteristics of the average leaf depth as observed in real catalogs ETAS is able to reproduce. Differences will hint at how burst-like sequences differ from *swarm-like* sequences.

The results of Chapter 2 and 3 will indicate systematic discrepancies between ETAS and real seismicity, presumably due to the occurrence of swarm-like sequences. However, a univocal definition of "swarms" is still missing. For this reason, we decided to elaborate a new method that characterizes and discriminates seismicity in a more objective and quantitative way, which I present in Chapter 4. We collect several earthquake properties that characterize sequences in different domains (space, time, and magnitude). We perform a *k-means* cluster analysis of the collected dataset to extract the most characteristic types of sequence, which we expect to be related to burst-like and swarm-like

sequences. Since not all properties will be equally important for a discrimination, we will identify the most important properties to discriminate sequences by types. In the end, we analyze the physical properties of each identified cluster by performing various statistical analyses and tests.

Finally, I will conclude my thesis providing new knowledge and findings that may help to make more performing and powerful forecasting models; I also hope that this material can lead the basis for further studies and insights into the study of different seismicity types.

2

New physical implications from revisiting foreshock activity in southern California

This chapter is published as:

Manganiello, E., Herrmann, M., & Marzocchi, W. (2023). New physical implications from revisiting foreshock activity in southern California. *Geophysical Research Letters*, *50*, e2022GL098737. https://doi.org/10.1029/2022GL098737

Abstract

Foreshock analysis promises new insights into the earthquake nucleation process and could potentially improve earthquake forecasting. Well-performing clustering models like the Epidemic-Type Aftershock Sequence (ETAS) model assume that foreshocks and general seismicity are generated by the same physical process, implying that foreshocks can be identified only in retrospect. However, several studies have recently found higher foreshock activity than predicted by ETAS. Here, we revisit the foreshock activity in southern California using different statistical methods and find anomalous foreshock sequences, i.e., those unexplained by ETAS, mostly for mainshock magnitudes below 5.5. The spatial distribution of these anomalies reveals a preferential

occurrence in zones of high heat flow, which are known to host swarm-like seismicity. Outside these zones, the foreshocks generally behave as expected by ETAS. These findings show that anomalous foreshock sequences in southern California do not indicate a pre-slip nucleation process, but swarm-like behavior driven by heat flow.

2.1 Introduction

It is well known that many large earthquakes are preceded by smaller events, e.g., 1999 M7.6 Izmit, Turkey (Bouchon et al., 2011; Ellsworth & Bulut, 2018), 2009 M6.1 L'Aquila, Italy (Chiaraluce et al., 2011), 2011 M9.0 Tohoku, Japan (Kato et al., 2012), 2019 M7.1 Ridgecrest, USA (Meng & Fan, 2021), which are (a posteriori) called foreshocks. The role of foreshocks in earthquake predictability can be epitomized by two still debated conceptual hypotheses about earthquake nucleation: the "pre-slip model" versus the "cascade model" (Ellsworth & Beroza, 1995; Gomberg, 2018). According to the former, foreshocks are diagnostic precursors, because they are triggered by an aseismic slip that precedes large earthquakes; in the latter model, foreshocks are like any other earthquake, which trigger one another, with one of them eventually becoming exceedingly larger (the mainshock).

Notwithstanding the still active debate on these hypotheses, seismologists are not yet able to recognize foreshocks in real-time, tacitly implying that foreshocks are not different from the rest of seismicity, indirectly supporting the cascade model. This view is further supported by the fact that the current best-performing class of short-term earthquake forecasting models (e.g., Nanjo et al. 2012, Taroni et al., 2018)—the Epidemic-Type Aftershock Sequences (ETAS; Ogata, 1988) model—assumes that foreshocks, mainshocks, and aftershocks are undistinguishable and governed by the same process. ETAS describes a branching point process also known as Hawkes or self-exciting point process:

every earthquake can trigger other earthquakes according to established empirical relations, with their magnitudes being independent from past seismicity. In essence, ETAS implicitly acknowledges the cascade model and its good forecasting performance makes ETAS an appropriate null hypothesis.

Instead, if foreshocks are dominated by other mechanisms than earthquake self-triggering, as the pre-slip model expects, foreshocks could be distinguished from general seismicity and potentially anticipate a larger earthquake. Several studies recently investigated foreshock sequences of southern California and found that they deviate from expectations of the classical ETAS model with spatially invariant triggering parameters. For example, Seif et al. (2019), Petrillo and Lippiello (2021), and Moutote et al. (2021) find, albeit at varying degrees, a higher foreshock activity in real seismicity than in synthetic catalogs simulated with ETAS. Hence, ETAS appears unable to completely reproduce observed seismicity, suggesting that foreshocks are distinct from general seismicity and governed by different mechanisms. These findings provide hope that identifying foreshocks as such could significantly improve earthquake predictability.

Here we reexamine foreshock activity in southern California; we investigate the existence and main characteristics of foreshock sequences that cannot be explained by ETAS, i.e., anomalous foreshock sequences. In other words, we look for new insights on the evidence against the cascade model. To be comparable to previous analyses, we use ETAS with spatially invariant triggering parameters. We perform two statistical tests and consider the potential influence of subjective choices by using two different existing ETAS models and two different methods to identify mainshocks and their foreshocks. To fathom the main characteristics of possible anomalous foreshock sequences, we investigate different magnitude classes and analyze the spatial correlation with heat flow as a physical parameter.

2.2 Data and Methods

We use the relocated earthquake catalog for southern California catalog (Hauksson et al., 2012), selecting all earthquakes with $M \geq 2.5$ from 1-1-1981 to 31-12-2019 except nuclear events (i.e., at the Nevada Test site), totaling 47'574 events.

Analyzing a catalog and distinguishing mainshocks, foreshocks, and aftershocks is unavoidably subjective because no absolute and precise procedure exists for this task (Molchan & Dmitrieva, 1992; Zaliapin et al., 2008). To reduce subjectivity, we analyze the catalog using two different techniques: the Nearest-Neighbor (NN) clustering analysis proposed by Baiesi and Paczuski (2004) and elaborated by Zaliapin et al. (2008), and the spatiotemporal windows (STW) method.

The NN method involves calculating distances in the space-time-magnitude domain between every event *j* and all previous events *i*. The event *i* with the shortest distance to *j* is called parent event, or NN; this distance is the NN distance η_j. By assigning a NN to each event *j*, all events become associated with another. To identify individual families (i.e., sequences) or single events, we use the same threshold $\eta_0 = 10^{-5}$ as Zaliapin et al. (2008), which removes event associations with too large η_j. For each sequence, we refer to the event with the largest magnitude as the mainshock and all associated preceding events as its foreshocks. We only consider sequences with foreshocks and ignore those that have no foreshocks.

For the STW method, we initially consider all events with magnitude $M \geq 4$ as possible mainshocks; to select foreshocks, we consider any preceding event within a spatiotemporal window of 10 km and 3 days from each mainshock (Agnew and Jones, 1991; Marzocchi and Zhuang, 2011; Seif et al., 2019). We exclude mainshocks that (i) do not contain any events within this window (i.e., sequences without foreshocks); (ii) are preceded by a larger event within this

window (Seif et al., 2019); and (iii) are preceded by an event with $M > 5$ within 100 km and 180 days (similar to Seif et al., 2019).

To simulate synthetic catalogs, we use two different parametrizations of the ETAS model that are already calibrated to southern California: (i) of K. Felzer (Felzer et al., 2002, see Supporting Information SI1 Text S1 and Table S1), based on parameter values of Hardebeck et al. (2008), see Table S2; and (ii) of S. Seif (Seif et al., 2017, see Text S1 and Table S3). In the results section, we only show results based on the former ETAS model and refer to results of the second model in the supporting information SI1. We prefer using existing models for four reasons explained in Text S1 (e.g., reduce degrees of freedom and confirm stability of our results). We verified the reliability of both ETAS models to produce realistic earthquake rates for different magnitude ranges using a Turing-style test (Page & van der Elst, 2018), see Text S2 and Figures S1–S4.

Once the mainshocks and their foreshocks have been identified in both the real and 10'000 synthetic catalogs (1000 for the second ETAS model), we compare their foreshock statistics using two tests named TEST1 and TEST2. Both are described in detail below; they both use ETAS as null hypothesis—which implies a cascade model as null hypothesis—but they examine different aspects of the problem. TEST1 involves the average number of observed foreshocks per sequence, whereas TEST2, which has been inspired by the work of Seif et al. (2019), involves the frequency of observing a certain number of foreshocks per sequence. We apply both tests to various mainshock magnitude classes $C_{\mathrm{M}} = \{4.0 \leq m_{\mathrm{M}} < 4.5, 4.5 \leq m_{\mathrm{M}} < 5.0, 5.0 \leq m_{\mathrm{M}} < 5.5, 5.5 \leq m_{\mathrm{M}} < 6.0, m_{\mathrm{M}} \geq 6.0\}$ and foreshock magnitude thresholds $T_{\mathrm{F}} = \{m_{\mathrm{F}} \geq 2.5,\ m_{\mathrm{F}} \geq 3.0,\ m_{\mathrm{F}} \geq 3.5,\ m_{\mathrm{F}} \geq 4.0\}$; these choices are based on Seif et al. (2019), but we added the class $4.0 \leq m_{\mathrm{M}} < 4.5$ to C_{M}. Although we report statistical test results, we do not formally account for applying the tests multiple times; the results are

therefore meant to indicate possible patterns of (apparently) anomalous foreshock activity.

In TEST1, the null hypothesis under test $H_0^{(1)}$ is that the average number of foreshocks (among sequences with foreshocks) in the real catalog is not larger than the corresponding quantity in the synthetic catalogs. For each mainshock magnitude class $\mathrm{c} \in C_\mathrm{M}$ and each foreshock magnitude threshold $\mathrm{t} \in T_\mathrm{F}$, we count the number of mainshocks (with foreshocks), $N_\mathrm{M}^\mathrm{real}$, and the number of foreshocks $N_\mathrm{F}^\mathrm{real}$ in the real catalog; $N_\mathrm{F}^\mathrm{real}$ is normalized by $N_\mathrm{M}^\mathrm{real}$ to obtain $\widehat{N}_\mathrm{F}^\mathrm{real}$. We calculate the same quantity for each synthetic catalog and build its empirical cumulative distribution function (eCDF); if $\widehat{N}_\mathrm{F}^\mathrm{real}$ is above the 99$^\mathrm{th}$ percentile of the eCDF, we reject $H_0^{(1)}$ at a significance level of 0.01.

In TEST2, the null hypothesis under test $H_0^{(2)}$ is that for each number of foreshocks, $N_\mathrm{F} > 0$, the frequency of observed cases is not larger than the frequency in synthetic catalogs. For each $c \in C_\mathrm{M}$ and each $t \in T_\mathrm{F}$, we count the number of mainshocks that have a certain N_F and normalize it by $N_\mathrm{M}^\mathrm{real}$. In this way, we obtain the probability mass function (PMF) for the real catalog as a function of N_F. Then, we apply the same procedure to each synthetic catalog and obtain 10'000 synthetic PMFs (1000 PMFs for the second ETAS model), for which we calculate the 99$^\mathrm{th}$ percentile at each N_F. Finally, at each N_F, we reject $H_0^{(2)}$ at a significance level of 0.01 if the corresponding PMF value of the real catalog is larger than the 99$^\mathrm{th}$ percentile (i.e., when the real catalog contains more foreshock sequences with this specific N_F than expected by ETAS). In essence, TEST2 seeks anomalies at every N_F, whereas TEST1 could be seen as a cumulative version of TEST2.

Based on the results of the tests, we can label each foreshock sequence as 'anomalous' or 'normal' using an intuitive approach: for TEST1, if the null hypothesis is rejected for a certain class, all foreshock sequences with a N_F

larger than the 99th percentile of the eCDF in that class are labeled as 'anomalous' (and 'normal' otherwise); for TEST2, if the null hypothesis is rejected for a specific N_{F}, all sequences with this N_{F} are labeled as 'anomalous' (and 'normal' otherwise). Effectively, a foreshock sequence in $c \in C_{\mathrm{M}}$ is labeled 'anomalous' if it is 'anomalous' in at least one class $t \in T_{\mathrm{F}}$. For TEST1, we argue that the approach is conservative due to comparing individual sequences with the average behavior of foreshock sequences; we therefore perform an alternative analysis (but not a test) without normalizing the number of foreshocks by N_{M}, see Text S3 and Figure S6.

To approach a physical interpretation of possible anomalous foreshock sequences in the real catalog, we analyze their spatial distribution. Specifically, taking inspiration from Zaliapin and Ben-Zion (2013), we interpolate heat flow measurements with a radial smoothing approach ($r = 20$ km) to acknowledge areas without data. We associate each foreshock sequence with the interpolated heat flow value at the mainshock location; if measurements within r are unavailable, we discard the sequence. Then we test if the distribution of associated heat flow values is significantly different for normal and anomalous foreshock sequences. We employ two statistical tests: the two-sample Kolmogorov-Smirnov test (null hypothesis: the two distributions have the same parent distribution) and the paired Wilcoxon test (null hypothesis: the two distributions have the same median). The Kolmogorov-Smirnov test is sensitive to any kind of difference between both distributions, whereas the Wilcoxon test is sensitive to one distribution having higher values than the other.

2.3 Results

2.3.1 Testing for anomalous foreshock activity

Figure 2.1 shows the results of TEST1 for each class in C_{M} and T_{F} using the ETAS model of K. Felzer and NN to identify mainshocks and their foreshocks; the results using STW are reported in Figure S5. Each subplot compares the eCDF (based on synthetic catalogs) with the observed value from the real catalog. As shown in Figure 2.1 and Figure S5, TEST1 rejects $H_0^{(1)}$, i.e., identifies anomalous foreshock sequences, exclusively for mainshock magnitudes $m_{\mathrm{M}} < 5.5$. Of a total of 152 foreshock sequences, we find 61 (40%) to be anomalous; with the STW method we find 143 foreshock sequences of which 34 (23%) are anomalous. Using instead the alternative analysis without normalizing by N_{M} (Figure S6), we find 19 (13%) to be anomalous, which suggests that TEST1 overestimated the number of anomalies due to using averages, as anticipated in Text S4. Applying TEST1 to the second ETAS model, we find 47 anomalous foreshock sequences for both NN and STW methods (31% and 33%, respectively, see Figures S7 and S8).

Figure 2.2 shows the results of TEST2 for each class in C_{M} and T_{F} using the NN method; the results using the STW method are reported in Figure S9. Most PMF values of the real catalog are not anomalous because they are below the 99$^{\mathrm{th}}$ percentile of synthetic PMF values. We find 22 of 152 (14%) foreshock sequences to be anomalous, most of which are again associated with $m_{\mathrm{M}} < 5.5$ (only three have larger m_{M}). Using the STW method we find 13 of 143 (9%) to be anomalous. Applying TEST2 to the second ETAS model, we identify 34 (22%) using NN method and 14 (10%, using STW method) to be anomalous (Figures S10 and S11, respectively).

For comparison, Figures 2.2 and S9–S11 also report the results using the approach of Seif et al. (2019), which tests a similar yet different null hypothesis than TEST2. Specifically, they treat all synthetic catalogs as one single

compound catalog. In this way, the PMF is normalized by a much larger number of mainshocks than contained in an individual synthetic catalog; for an increasing number of synthetic catalogs, the PMF decreases progressively (i.e., lowering the detectable minimum frequency) and moves further away from the real observation. In other words, our TEST2 honors that a finite earthquake catalog must have a lower detectable frequency of foreshocks in the PMF; this lower frequency depends on the number of mainshocks that have foreshocks, which in turn depends on the length of the earthquake catalog (the lowest frequency is one out of the number of mainshocks that have foreshocks). In addition, the approach of Seif et al. (2019) normalizes the PMF by the total number of mainshocks that have foreshocks (N_{M}, as we do in TEST2) or *no* foreshocks, which further reduces the PMF by another 0.5–1 order of magnitude depending on $c \in C_{\mathrm{M}}$.

We repeated TEST1 and TEST2 at a 0.05 significance level (i.e., 95$^{\mathrm{th}}$ percentile), which was originally used by Seif et al. (2019), see Text S4 and Figures S12 and S13.

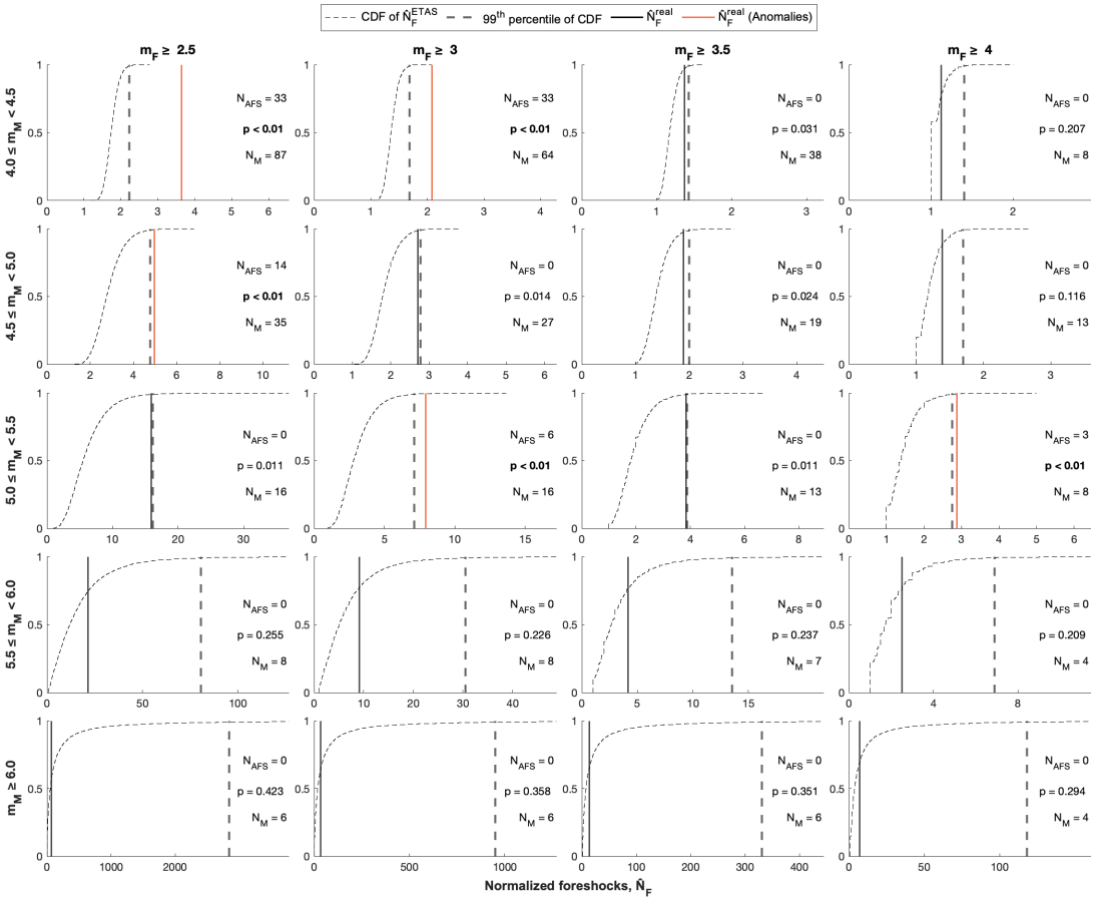


Figure 2-1 Results of TEST1 for various classes of the mainshock magnitude m_M (rows) and thresholds for the foreshock magnitude m_F (columns). Each subplot displays the number of normalized foreshocks $\hat{N}_F$ for the real catalog (vertical line; red if anomalous, black otherwise) and the empirical Cumulative Distribution Function (eCDF, dashed curve) with its 99th percentile (dashed vertical line) for 10'000 synthetic catalogs. Each subplot also reports the number of anomalous foreshock sequences, N_{AFS}, the p-value for TEST1, and the number of mainshocks, N_M. The results are based on K. Felzer's ETAS model and the NN method; Figure S5 shows results using the STW method, and Figure S7 and S8 using the second ETAS model. Note that each subplot uses a different N_F-axis range to account for the varying data range.

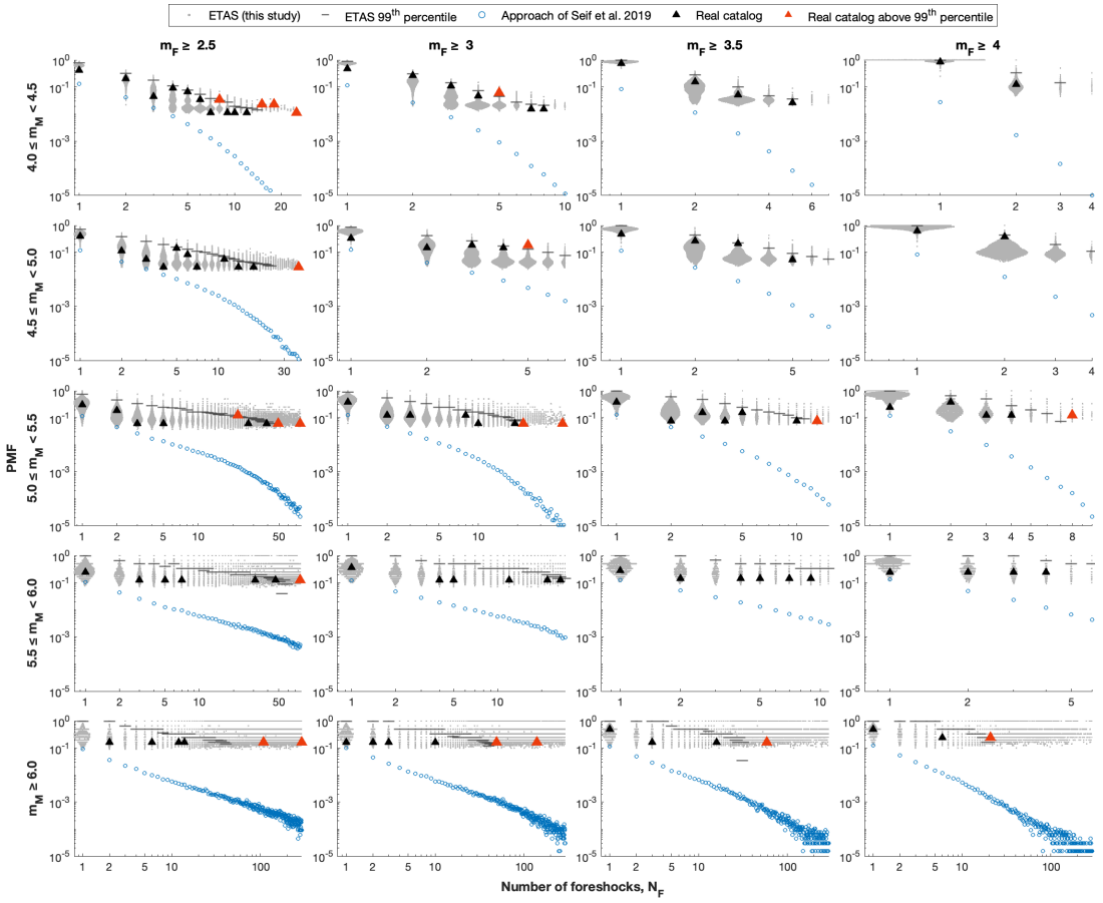


Figure 2-2 Results of TEST2 showing probability mass functions (PMFs) of the number of foreshocks N_F for various classes of m_M (rows) and m_F (columns). The PMFs are shown for (i) the real catalog (triangles), (ii) all synthetic catalogs (small gray dots as swarm distributions) with their 99th percentile (gray horizontal bars), and (iii) when considering all synthetic catalogs as a single compound catalog (blue open circles, using the approach of Seif et al., 2019). Triangles become red when they are located above the 99th percentile of (ii). The results are based on K. Felzer's ETAS model and the NN method; Figure S9 shows results using the STW method, and Figures S10 and S11 using the second ETAS model. Note that each subplot uses a different N_F-axis range.

2.3.2 Correlating foreshock sequences with the heat flow

To investigate the physical cause of anomalous foreshock sequences, we inspect their correlation with the local heat flow. We choose this property because previous studies suggested a relation between heat flow and statistical properties of earthquake sequences (e.g., Enescu et al., 2009, Chen & Shearer, 2016; Ross et al., 2021; Zaliapin & Ben-Zion, 2013).

Figures 2.3a and 2.4a overlay the locations of normal and anomalous foreshock sequences identified by TEST1 and TEST2, respectively, on a heat flow map. Figures 2.3b and 2.4b show the corresponding eCDFs of the heat flow interpolated at the locations of normal and anomalous foreshock sequences. In both cases, anomalous foreshock sequences tend to occur more frequently at locations of higher heat flow than normal foreshock sequences. This trend is confirmed by the *p*-values of the two-sample Kolmogorov-Smirnov and paired Wilcoxon tests (see annotations in Figures 2.3b and 2.4b): Being below 0.05, they indicate that the two samples come from different parent distributions with different medians. Figures 2.3 and 2.4 are based on the NN method to identify mainshocks and their foreshocks; the results based on the STW method confirm our findings (see Figures S14 and S15), as do the results based on the second ETAS model (see Figures S16–S19) and on the 0.05 significance level (Figures S20 and S21). Moreover, TEST1-based results are stable even if we use the alternative analysis without normalizing by N_{M} (see Figure S22). We verified the stability of our results using an independent modeling of foreshock sequences by Petrillo and Lippiello (2021): the authors provided us locations of their identified normal and anomalous foreshock sequences, letting us apply our analysis on a dataset that is completely independent from our assumptions and modeling choices (see Figure S23). It confirms our findings of a preferential occurrence of foreshock anomalies in high heat flow zones. We summarize the *p*-values of all the different analyses in Table S4.

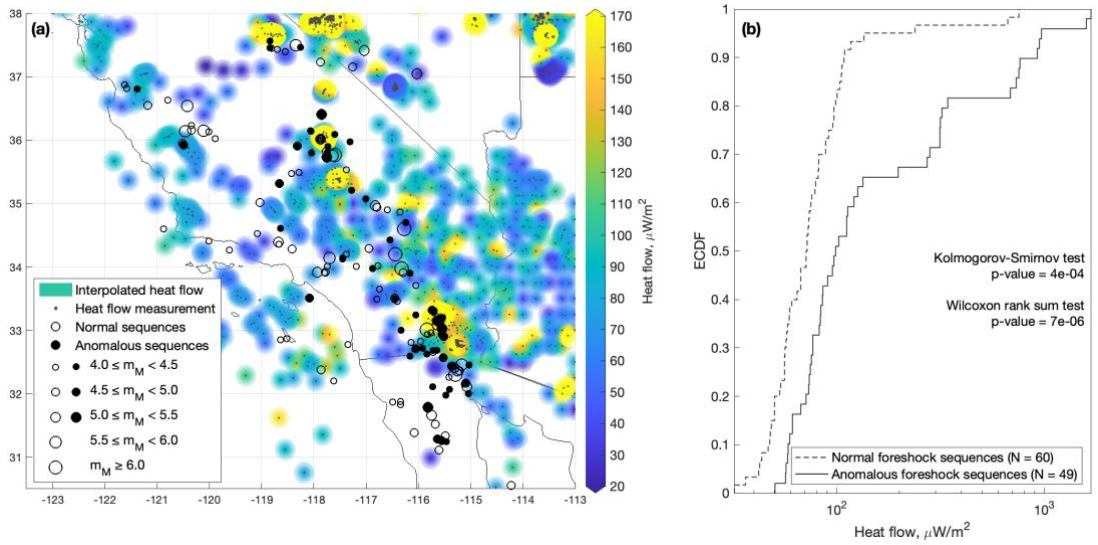


Figure 2-3 Correlating foreshock sequences with the heat flow. (a) Locations of normal (empty circles) and anomalous foreshock sequences (filled circles) identified with TEST1 overlayed on a heat flow map. The circles sizes scales with m_M (see legend). The interpolated heat flow map is based on sampled heat flow measurements (small gray dots, see Data and Methods section); (b) eCDFs of heat flow values at locations of normal (dashed curve) and anomalous foreshock sequences (solid curve); both eCDFs are compared using two statistical tests (see annotation with corresponding p-values). The results are based on K. Felzer's ETAS model and the NN method; Figure S14 shows results using the STW method, and Figures S16 and S18 using the second ETAS model.

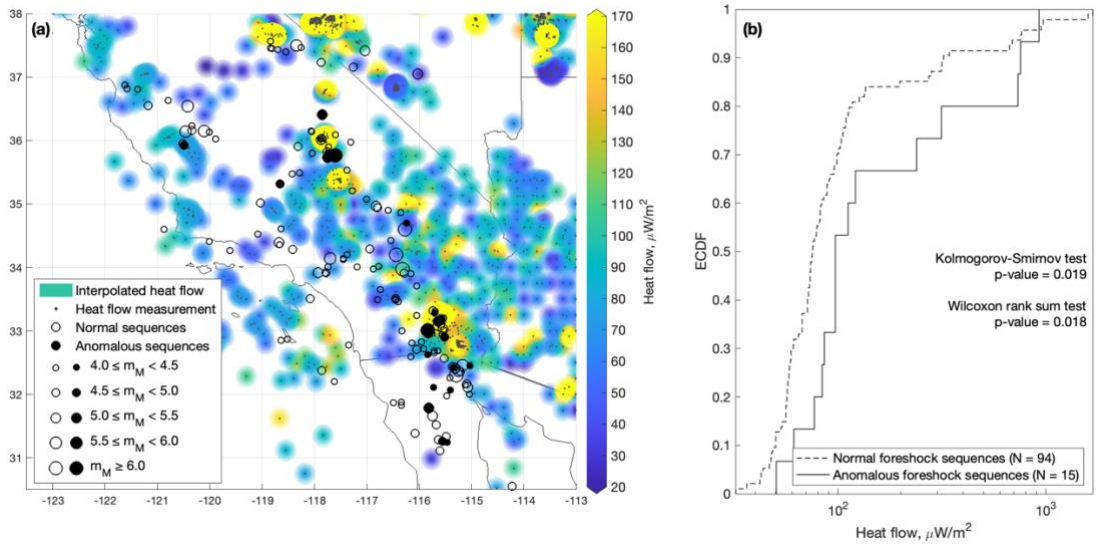


Figure 2-4 Like Figure 2.3 but with foreshock sequences labeled as 'anomalous' or 'normal' using TEST2. Figure S9 shows results using the STW method, and Figures S17 and S19 using the second ETAS model.

Finally, we add a word of caution on the interpretation of the results, that is, the spatial coverage of heat flow data compared to the earthquake activity is rather incomplete in northern Mexico. For instance, several anomalous foreshock sequences occur in this area but cannot be included in the heat flow analysis due to the lack of data. In addition, the available heat flow measurements in northern Mexico are not consistent with the Geothermal map of North America (Blackwell & Richards, 2004), which indicates a generally high heat flow (> 100 $\mu W/m^2$) in this area along the San Andreas Fault.

2.4 Discussion & Conclusion

We have found that foreshocks have the same characteristics of general seismicity as expected by ETAS, except for some cases. Our finding is in general agreement with previous studies of foreshock activity, all of which found (with some important differences not discussed here) higher foreshock activity than expected (Chen & Shearer, 2016; Moutote et al., 2021; Petrillo & Lippiello, 2021; Seif et al., 2019). However, our results additionally show that foreshock anomalies are mostly associated with mainshock magnitudes below 5.5—independently from the two tests, the two ETAS models, the two procedures to identify mainshocks and their foreshocks, and an independent set of foreshock anomalies. Moreover, these anomalies are located preferentially (and statistically significant) in zones of high heat flow. The combination of these two findings suggests that sequences with anomalous foreshock activity behave more like seismic swarms. In fact, independent studies (e.g., Enescu et al., 2009, Chen & Shearer, 2016; Ross et al., 2021; Zaliapin & Ben-Zion, 2013) have shown that swarm-like seismicity is common in those areas where we have found anomalous foreshock sequences.

Our results do not allow us to further elucidate why foreshock anomalies correlate with high heat flow. The anomalies may be driven by specific physical

mechanisms (e.g., actual seismic swarms mostly driven by fluids) or still relate to a cascade model that is not spatially uniform. The latter may be better described by an ETAS model with spatially varying triggering parameters. Indeed, Enescu et al. (2009) and Nandan et al. (2017) show that some parameters of a spatially varying ETAS model (which mostly depend on the more abundant aftershocks) correlate with the heat flow in southern California. Their more elaborated clustering model implies more active foreshock sequences where the heat flow is high, which agrees with our empirical findings using (less abundant) foreshocks.

Conversely, foreshock sequences located in zones of lower heat flow predominantly behave as expected, i.e., in agreement with the null hypothesis given by the ETAS model (which mimics the cascade model). If we interpret the difference in foreshock activity as evidence of the pre-slip model, it must have a minor effect in zones of lower heat flow, but it may become more important in zones of high heat flow. In other words, our results are inconsistent with pre-slip as a general nucleation process; pre-slip may become only relevant under specific tectonic conditions, such as in high heat flow. Our results do not prove the cascade model as the truth, but neither do they bring any evidence against it nor in favor of the pre-slip model. Perhaps alternative hypotheses open up a middle ground: recent studies proposed that both processes can coexist and relate to each other (McLaskey 2019; Cattania & Segall 2021) or that nucleation follows a different process (Kato and Ben-Zion, 2021). But like the pre-slip model, these conjectures remain to be tested.

Our results also highlight the importance of analyzing earthquake sequences in zones of high heat flow in more detail, especially to understand the physical reasons of anomalous foreshock sequences: Are they related to seismic swarms with an implicit limitation to the mainshock magnitude? Or are they related to different clustering processes than those driving tectonic sequences? The difference is crucial, in particular regarding the forecasting of large earthquakes.

Our findings raise an urgent need to find (quasi-)real-time methods to discriminate swarm-like from ETAS-like sequences. Such a discrimination could lead to significant improvements in earthquake forecasting, because being able to identify a swarm-like sequence as such could markedly reduce the forecast probability for a large earthquake. A promising indicator could be the background rate component of ETAS, which has been found to increase during swarm-like seismicity (Hainzl and Ogata 2005; Lombardi et al., 2006; Llenos et al. 2009; Kumazawa et al., 2016). Another possibility was raised by Zaliapin and Ben-Zion (2013) demonstrating that swarm-like sequences have a different topologic tree structure (i.e., an internal clustering hierarchy, which connects background and triggered earthquakes). Unfortunately, this approach can currently only be used retrospectively, limiting its applicability in earthquake forecasting. We envision other possible parameterizations of the topologic tree structure that may facilitate its use for earthquake forecasting.

3

A useful scalar measure for improving earthquake forecasting: the Average Leaf Depth

Abstract

Recent studies have shown significant differences between observations of earthquake sequences and the expectation by ETAS (Epidemic-Type Aftershock Sequence) models - a well-performing class of earthquake clustering models. To better understand why these discrepancies exist, we discriminate different types of seismicity using the average leaf depth (ALD), a scalar measure that represents the topological structure of a sequence. Zaliapin and Ben Zion (2013) showed that high ALD values characterize swarm-type sequences, which the standard ETAS model anticipates poorly. Here we estimate ALD of real and synthetic catalogs using several ETAS models: we note important variation according to the used model.

3.1 Introduction

One of the most addressed issues in seismology is characterizing earthquake sequences and represent them through physical laws and/or empirical relations. These ingredients are the basis of the best-performing class of earthquake forecasting models, ETAS (Epidemic Type Aftershock Sequences). Studying

phenomena of the past (i.e., seismicity) makes it possible to forecast what will happen in the future. The question arises: can we anticipate earthquake sequences well, i.e., reproduce reality? To provide an answer, we need to quantify the discrepancies and understand how well ETAS is able to reproduce the reality.

We already know that there are significant differences between real and synthetically simulated catalogs (Moutote et al., 2021; Lippiello et al., 2017; Petrillo & Lippiello, 2021; Seif et al., 2019; Zaliapin et Ben Zion, 2013). These discrepancies concern sequences that are typically defined as swarm: they do not have a single large earthquake but many of the same size that is typically in the medium magnitude range (M3-5), have many foreshocks before those large events over an extended time period, and are located in particular regions. All those behaviors are not described well by a standard ETAS model, which leads us to believe that different types of seismicity exist, but no objective (i.e., quantitative) way to differentiate between them. This ability would pave the way for improved earthquake forecasting by being able to 1) identifying the different types of seismicity, ii) studying and understanding their differences, and iii) characterizing them from a physical and mathematical point of view. Ultimately, these findings could be implemented in forecasting models.

To address points i) and ii), Zaliapin &Ben Zion (2013) use a simple scalar measure, the average leaf depth (ALD), which quantifies the topological structure of sequences. They use ALD to divide sequences into burst-like (low ALD, dominated by first-generation offspring of the mainshock) and swarm-like sequences (high ALD, composed of many offspring generations). As there is not a specific threshold between these types, Zaliapin and Ben Zion (2013) describe the distribution of ALD as "*a mixture of two exponential distributions with distinct slopes and change point at ALD = 5*". In addition, they compute ALD for sequences of a single synthetic catalog: they do not find ALD > 5, nor

evidence of swarm-like seismicity. This finding confirms that swarm-type sequences follow different laws than ETAS.

Here, we reinvestigate the work Zaliapin & Ben Zion (2013) and explore the characteristics of sequences that ETAS does not describe well. For that, we compute ALD for real and synthetic catalogs reproduced using different ETAS models.

3.2 Method

We use the same real catalog provided in Chapter 2 (see Method section). We reproduce three set of 1'000 synthetic catalogs using: i) K. Felzer model imposing point sources for all earthquakes, ii) K. Felzer model with planar sources for large earthquakes (M > 6.5), and iii) S. Seif model that uses point sources for all earthquakes (more details in Method section of Chapter 2).

In this contest, we use only the NN-declustering method to identify sequences for all catalogs, as explained in Chapter 2 (see Method section).

Once sequences have been identified, we use the *average leaf depth* (ALD) as a scalar measure to characterize them quantitatively. ALD is based on the idea to represent a sequence as a tree structure (see Figure 3.1): the first vertex is the parent event and starts a family; subsequent events that are directly triggered by the parent represent the first generation of event(s) (i.e., one or more child, which each creates a branch); each of those children can trigger a second generation of event(s), and so on. ALD is the sum of the generation level of each leaf (i.e., last event of each branch), $\sum d_g$, divided by the total number of leaves n_g:

$$ALD = \frac{\sum d_g}{n_g}$$

To explain why this measure is useful to characterized sequences, we show the structure of two different sequences (Figure 3.1): the first one with high ALD (Linear structure), and the second one with low ALD (Burst structure). High ALD values represent sequences with many events for each generation that follow one each other ($n_e = 3$), and few number of generations ($n_g = 1$). On the other side, a burst structure has low ALD: the number of generations is too high ($n_g = 4$) with few events for each generation ($n_e = [2, 2, 1, 2]$). Therefore, in sequences with the linear structure, an event tends to trigger only one event; sequences with a burst structure are characterized by an event that triggers several branches of events.

These findings come from the work of Zaliapin and Ben Zion (2013). Although they are not able to identify a sharp boundary between the cluster types, ALD is believed to facilitate the classification of two basic types of earthquake sequences: swarms ($ALD > 5$) and bursts ($ALD < 5$). This classification can only be done for large family size, i.e., $n > 100$ and mainshock magnitude $M_M < 7$ (Zaliapin and Ben Zion, 2013).

In Figures 3.2 & 3.3, we explore ALD values as function of family size and mainshock magnitude for each dataset. We also compare the results obtained in this work with those of Zaliapin and Ben Zion (2013) (Figures 3.4); they use a single synthetic catalog with minimum magnitude $M = 2$ and a different ETAS model with respect to ours. By performing the analysis of ALD for each family in one synthetic catalog, they do not find ALD > 5, defining such values as highly improbable. Such finding supports the hypothesis that ETAS does not forecast swarm-type sequences.

3.3 Results

In this chapter, we want to explore ALD values of real and synthetic catalogs. In Figure 3.1, we provide a simple example of his computation for two different types of sequences: linear and burst.

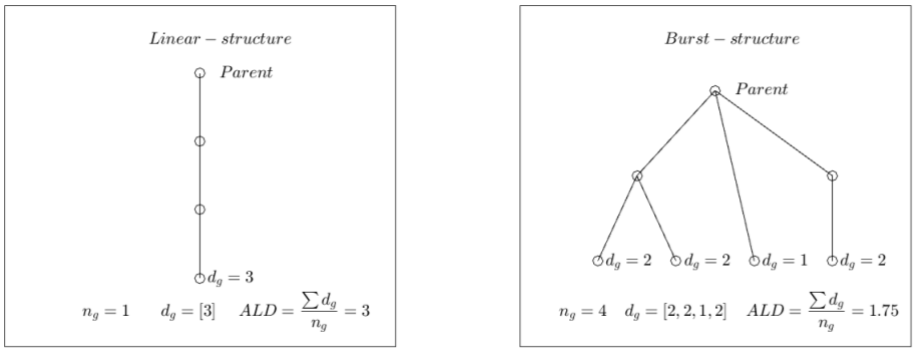


Figure 3-1 Average leaf depth, ALD, of two different types of simplistic sequences; d_g is the sum of events over each leaf, and n_g is the number of leaves; ALD is the ratio of d_g and n_g (figure inspired by the work of Zaliapin & Ben Zion, 2013).

Each subplot of Figure 3.2 represents a specific dataset: we plot ALD as function of the family size (i.e., the logarithm of the number of events in each sequence) colored according to the density of data: the closer one gets to the area of higher density, the more data overlaps.

Figure 3.2 (a) highlights a deviation of ALD distribution in the real catalog respect to the expectation from ETAS (Figures 3.2 b, c, d). In detail, the real seismicity (Figure 3.2 a) is characterized by several trends: i) sequences where ALD increases with the family size (until N < 3), ii) sequences whose ALD is independent from the family size, and iii) apparent increase of ALD for large families (N > 2). The maximum ALD is ALD $= 14$.

With K. Felzer ETAS models, important differences rise according to the used shape-source: imposing point-source, the maximum value is $ALD = 10$ and ALD is independent from the family size (Figure 3.2 b); using planar source for

sequences with mainshock $M_M \geq 6.5$, the maximum ALD is $ALD = 27$ and all of high ALD are related to the sequences with $N \geq 3$ (Figure 3.2 c). Both lack an increasing trend of ALD with the family size for $N < 3$, that instead characterize the real catalog.

In Figure 3.2-d, we show the results using S. Seif point source model: compared to previous models, there are higher ALD values (until $ALD = 37$) also for $N < 3$. Furthermore, it is not possible to recognize an increasing trend of ALD values with family size.

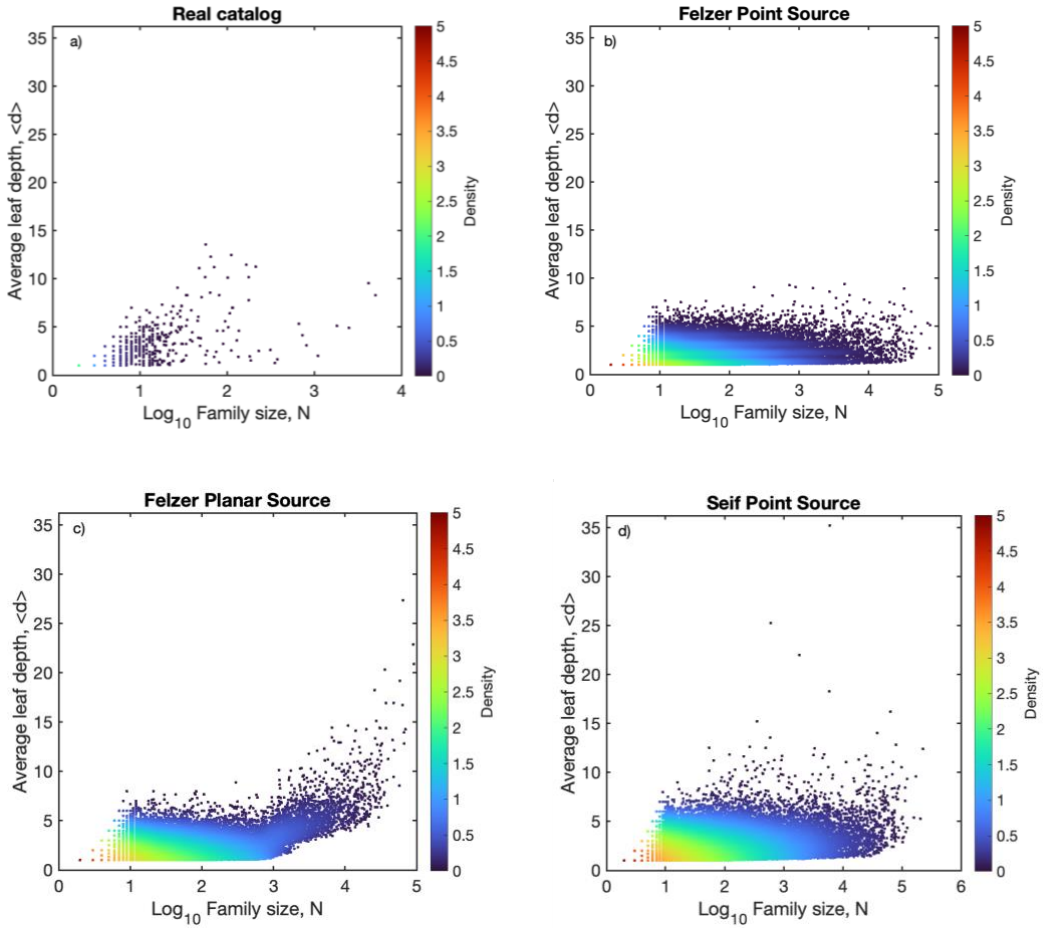


Figure 3-2 Average leaf depth, ALD, vs family size, N, for the real catalog (a) and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). Dots are colored according to the density of the data. Higher density areas have more overlapped data.

To better explore discrepancies among real and ETAS models, in Figure 3.3 we plot ALD as function of the mainshock magnitude M_M of sequences.

We can note the same trends as in Figure 3.2: for both real and synthetic catalogs, most sequences have ALD that is independent from M_M (i.e., constant trend); for large sequences ($M_M \geq 6.5$), ALD increases again with M_M only for real catalog and for model using planar source (Figure 3.3 a, c).

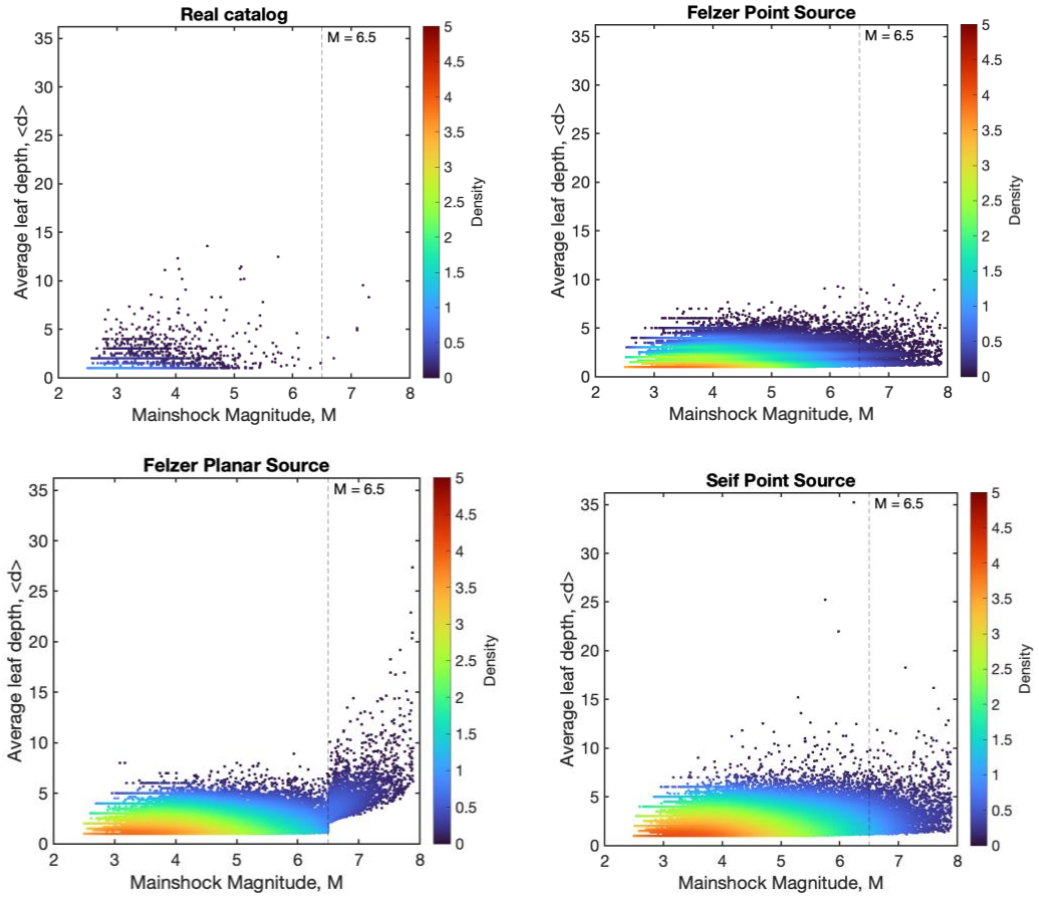


Figure 3-3 Average leaf depth, ALD, vs mainshock magnitude, M_M, for the real catalog (a) and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). Dots are colored according to the density of the data. Vertical line highlights the threshold $M_M = 6.5$ used for Felzer planar source model. Higher density areas have more overlapped data.

Figure 3.4 shows the distribution of ALD for real and synthetic catalogs using histograms. To make a comparison with previous findings by Zaliapin and Ben Zion (2013), we compute the fraction of families with ALD > 5. Considering the complete dataset, 1.95% of real sequences have ALD larger than this threshold; using Felzer model with point source, only 0.03% of the synthetic sequences have $ALD > 5$, and 0.06% using planar source model. We note similar results for point-source model by S. Seif: 0.06% of synthetic sequences with $ALD > 5$.

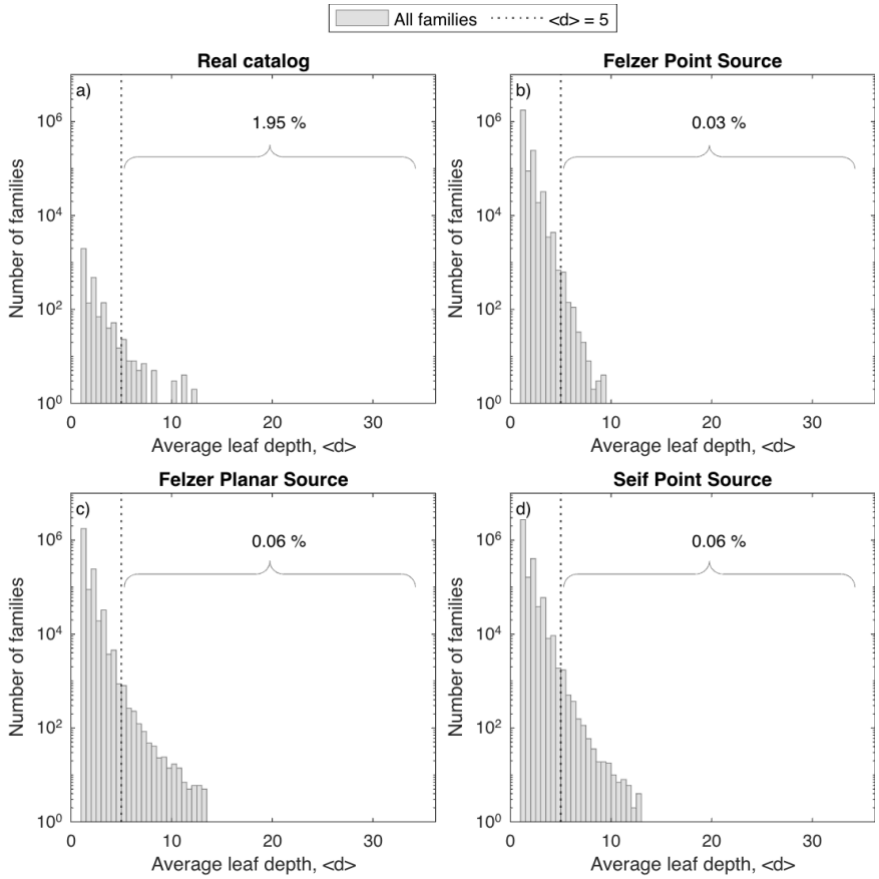


Figure 3-4 Distribution of ALD in terms of histograms for (a) the real catalog and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). The vertical dotted lines represent ALD $= 5$, the maximum ALD found by Zaliapin and Ben Zion (2013) using one synthetic ETAS catalog; curly brackets represent the fraction of families with ALD > 5;

3.4 Discussion and Conclusion

In the current study, we have found that the topological structures of synthetic sequences depending on the used ETAS model. None of the analyzed models fully represents the topological structures of real sequences.

We also highlight important difference using the same ETAS model (by K. Felzer) with different type of earthquake-sources. Using a point source, it is hard to find sequences with very high ALD and no dependence with mainshock magnitude and/or family size can be noted (that instead characterize real seismicity). These results partially agree with the work of Zaliapin and Ben Zion (2013): they argued that it is highly unlikely to produce sequences with $ALD > 5$ using ETAS.

This became true using planar source for sequences with mainshock magnitude $M_M \geq 6.5$: above this threshold, ALD values increase exponentially as function of mainshock magnitude and family size, as apparently happen for real sequences. However, even such planar-source model has some discrepancies with the real seismicity: it is not able to forecast sequences with high *average leaf depth* values ($ALD > 10$) with moderate mainshock magnitude ($3.9 < M_M < 6$) and large number of events inside ($1 < N < 3$). We could associate these unreproducible sequences with *swarm*-type seismicity: although there is no clear definition of a swarm-type seismicity, they typically have a large spatiotemporal extension, large number of foreshocks and aftershocks, and several of the largest events in a comparable magnitude range; all these characteristics agree with the range of ALD, M_M, and N of sequences that ETAS planar-source model of K. Felzer is unable to reproduce. In fact, previous results (Chen and Shearer, 2016; Manganiello et al., 2023; Moutote et al., 2021; Seif et al., 2019; Lippiello et al., 2017; Zaliapin and Ben Zion, 2013) assert, albeit at varying degrees, that ETAS does not reproduce *swarm*-type sequences, which are characterized by large number of foreshocks, with small-to-moderate

mainshock magnitude, and located in a particular region (like high heat flow zones).

Anyway, we cannot assert the same conclusion using the point-source model by S. Seif. We find sequences with very high ALD (ALD > 5) for different order of magnitude (i.e., also for $M_M < 6.5$), which is in contrast to the above and Zaliapin and Ben-Zion (2013) results.

Currently, we are not able to explain why there are differences in ALD-trend using several models and types of sources. Finally, we can only assert that ETAS, in some cases, is able to reproduce sequences with *average leaf depth* values higher than those identified by Zaliapin and Ben Zion (2013), but the percentages of such sequences are still lower than for real seismicity, for all used models. Some ETAS models cannot reproduce sequences which seem to be associate with swarm-type seismicity. Moreover, the used source model could also play a very important role in the calculation of ALD. In this regard, Bayliss et al. (2019) highlights another significant uncertainty in the *average leaf depth* computation by using a probabilistic approach for identifying earthquake clusters (i.e., without using a binary threshold). This aspect can be also a limiting factor when discriminating between burst and swarm-type seismicity.

Although ETAS is one of the best-performing earthquake forecasting models, it still describes seismicity differently from what we observe in reality. However, the variability of the ALD values, among the different models, leads us to think that ALD is not a completely objective method for the discretization of seismic types. On the contrary, its computation can be influenced by various factors, such as the type of used source. Consequently, it is difficult to state whether the observed discrepancies between ETAS and real seismicity are actually associated with swarm-type sequences at all. At least, we cannot assert that, just using average leaf depth.

4

What makes an earthquake sequence a swarm? Using Cluster analysis to discriminate earthquake sequences in southern California

Abstract

Characterizing and discriminating earthquake sequences, especially in southern California, is nothing new (e.g., Zaliapin & Ben Zion, 2013; Chen et Shearer 2016), but it lacks a formal and consistent definition of their main types (i.e., swarm-type and aftershock-type sequences). Here we propose a method to discriminate between sequences quantitatively using a comprehensive collection of their properties. For each sequence we extract various features that characterize them in space, time, and magnitude domain (incl. previously proposed definitions of swarms based on a*verage leaf depth* and *relative timing of the largest event*). We perform a *k-means* cluster analysis of collected features for different *k-values* and using different methods to identify the optimal number of clusters. Finally, we investigate identified clusters in terms of their spatial distribution, their correlation with the heat flow, and statistical properties such as the magnitude frequency distribution, temporal rate behavior, and foreshock/aftershock productivity. We find statistically significant differences among them, which indicates that they represent different types of seismicity, making our discrimination meaningful. One cluster is more common in zones

of high heat flow, with a high productivity of foreshocks for small mainshock magnitude and larger number of events approaching the mainshock (rate), indicating swarm seismicity. Our findings provide insights into the interrelation and joint significance of sequence properties, which allows distinguishing sequences consistently in retrospect. To contribute to improved earthquake forecasting in future work, we will restrict our approach to features related only to foreshock information (i.e., before the largest event); characterizing an on-going sequence could then allow us to recognize the type of sequence in real-time and adjust the probability for a larger earthquake accordingly (e.g., lowering it upon indication of a swarm-type sequence).

4.1 Introduction

Swarm sequences are one of the most debated topics in seismological field (Zaliapin & Ben Zion, 2013; Chen et al., 2012; Enescu et al., 2009; Farrell et al., 2009; Mogi et al., 1963), since there is no objective definition of what they are. There are several evidence of the fact that they have "atypical" behavior compared to normal seismicity, i.e., the best known and already widely studied aftershock-type sequences (Mogi, 1963). Usually, for such seismic sequences, we can recognize a larger-magnitude event (mainshock) that (directly or indirectly) triggers smaller-magnitude events that tend to decrease over time (Mogi, 1963). In the case of swarm-type sequences, the mainshock is not easily distinguishable from other events, as its magnitude is very similar to the others (Ross et al., 2021); external forcing may play an important role, increasing stress and/or reducing strength on seismogenic faults (Shelly et al., 2016). In fact, the engine of these sequences seems to correlate with high heat flow zones (Manganiello et al, 2023; Zaliapin and Ben Zion, 2013), movements of fluids in the subsoil (Vidale & Shearer, 2006), aseismic slip (Lohman and McGuire, 2007), presence of gas, volcanic areas, reservoirs (Kato et al., 2010).

Consequently, there is no clear definition of what a swarm is, and this affects the possibility to perform statistical analysis on them, not understudying their origins. Previous studies attempted to quantify the difference between the types of sequences, in particular:

1. Zaliapin & Ben Zion (2013) use the average leaf depth (ALD) as a simple scalar measure to classify sequences based on their topological structures; they identify swarm-like sequences (with more linear structure) and burst-like sequences (with more branched structure) according to their ALD (see Chapter 3 for more details).

2. Chen et Shearer (2016) use the *relative timing* (which is the time delay of the largest event normalized by the mean time delay within the sequence) to define aftershock-type cluster ($t_{rel} < 0.2$, less prone to spatial migration) and swarm-type ($t_{rel} > 0.2$ more prone to spatial migration).

However, we do not know if these definitions are influenced by the same underlying laws or mechanisms.

The main goal of this chapter is to inspect all southern California sequences and to provide a working definition of the different types of sequences, starting from their characterization in space, time, and magnitude domain; in so doing, we collect a large comprehensive dataset for seismicity, also considering the knowledge we already have (as ALD and *relative timing* as quantitative measurements to classify sequences). We analyze all these characteristics at the same time, and we perform the *k*-means cluster analysis to split the sequences into different groups, if dissimilar, or in the same group if similar. As we do not assume the existence of only two types of seismicity, we explore several groups.

Moreover, we wonder if all collected features have the same weight for the cluster analysis, or if there are some needless ones. To explore that, we use an approach based on the leave one out method, getting a subset with the most

influential features to use for the cluster analysis. We also wonder about the selected characteristics and why they are so useful for the cluster identification.

Finally, we focus on understanding the origin of identified clusters and why they are different, to achieve different governing properties among them. For that, we analyze the spatial correlation of identified clusters with heat flow values (as a physical parameters) and their statistical properties such as the magnitude frequency distribution, temporal rate behavior and foreshock/aftershock productivity. Based on previous studies, we expect to find a cluster (the aftershock-type) which responds to the laws we are already aware of, and another cluster (the swarm-type) whose physical and statistical properties are unusual and in agreement with previous studies, such as localization of mainshocks in particular areas (as already found by Manganiello et al., 2023; Zaliapin & Ben Zion, 2013) and with a huge productivity of foreshocks (Seif et al., 2019; Lippiello et al., 2017).

The usefulness of this procedure lies in being able to discriminate between sequences quantitatively, finding the best separation simultaneously for all properties. This is handier than performing statistical analyses for each property separately. Being able to identify different types of sequences, with specific characteristics, can contribute to improving earthquake forecasting and the understanding of earthquake nucleation processes. If there are statistical differences among seismic sequences, we could i) classify them in different clusters according to their statistics, ii) characterize them from different points of view (Magnitude frequency distribution, rate, productivity), and iii) use these outlooks to improve earthquake forecasting: recognize an ongoing sequence *a-priori* as "swarm" could exclude the hypothesis of an impending large earthquake.

4.2 Data & Method

The method is explained in several steps through the following subparagraphs (from 4.2.1 to 4.2.4) and each of them is graphically showed in Figure 4.1.

4.2.1 Collected dataset

The real catalog used in this chapter and the NN-declustering method for identifying sequences are explained in Chapter 2 (see Method section). We only consider sequences with a number of events $N \geq 10$ and we collect the following features for each of them:

1. *Relative timing* of the largest event
2. Average Leaf Depth (ALD)
3. Depth of the largest event
4. Nearest fault distance of each sequence
5. Relative duration of foreshocks and aftershocks
6. Area of foreshocks and aftershocks
7. Magnitude difference between mainshock and largest foreshock & aftershock
8. Productivity of foreshock and aftershocks
9. Largest magnitude

In the Supporting Information file SI2, we provide all the equations used to calculate the characteristics listed above and we also show several plots

representing them (from Figure S2.1 to S2.11). Before performing the analyses, we normalize the dataset using the following boxcox MATLAB function:

If λ is not = 0, then:

$$data(\lambda) = \frac{data^{\lambda}-1}{\lambda};$$

If λ is = 0, then:

$$data(\lambda) = \ln(data);$$

where λ is the value that maximizes the Log-Likelihood Function (LLF). One of the most important conditions for using the boxcox function is that all variables (i.e., all features computed for all sequences) have non-null (i.e., no-missing values) or positive values. This condition could limit the analysis, as it is not possible to comply with that for various characteristics: for example, it is not possible to calculate the area of the foreshocks for sequences that do not have any, as well as the difference in magnitude with the mainshock, etc. In these cases, we replace null or zero values with the minimum positive value of that specific variable, reduced by one unit.

4.2.2 k-means Cluster Analysis

We perform the cluster analysis exploring the entire collected and normalized dataset. We use this technique because it is able to find clusters that share common features and explore valuable information hidden in the groups.

Considering that there are numerous clustering algorithms proposed in the literature, we decide to use the *k-means* due to its simplicity and efficiency (Li & Wu, 2012): it is able to split data in several *k*-groups, selecting a set of initial centroid locations (according to *k*-value). Each feature of the dataset is assigned

to a centroid using the squared Euclidean distance; in this way clusters start to form. The centroid position is updated every time a feature is added to the cluster; this process ends when all characteristics are assigned to a group and there are no more variations (Kaufman and Rousseeuw, 1990).

The first step of this analysis is to understand how many clusters characterize seismicity. For this purpose, *k-means* cluster analysis comes in handy, as we can decide how many types of clusters we want to explore (i.e., *k-values*); therefore, to achieve our goal, we perform the analysis for *k-values* = [1:10].

However, it is well known that *k-means* cluster analysis can be a double-edged sword: it is always able of identifying different types of clusters (for each *k-value*) even if there is no relationship among data and, consequently, no possibility to interpret the results. To overcome this problem, we look for the optimal number of clusters (N_{opt}) by crossing the results of different methods: i) *Average silhouette mean*, for which the maximum value is the appropriate number of clusters (Kaufman and Rousseeuw, 1990), and ii) the *Elbow method*, for which the knee point of the distribution of the total within-cluster sum of square is the optimal number of cluster.

4.2.3 Statistical properties of identified clusters

To prove the statistically differences among identified clusters, we start to investigate sequences in terms of spatial distribution. For each cluster, we explore the correlation of the mainshock location with heat flow values of southern California as physical quantity, taking inspiration from Manganiello et al., 2023 (Chapter 2). We make eCDFs of interpolated heat flow values and we compare them using two statistical tests: Kolmogorov-Smirnov (KS) and Wilcoxon rank (W) tests.

As we want to discriminate the types of sequences from different points of view, we also perform a statistical analysis in the time and magnitude domain. We plot the MFDs of the mainshocks and MFDs of all events, both for each identified cluster. We test if the magnitude distribution is significantly different among clusters, performing two statistical tests (KS and W tests mentioned above) and we compute the *b-values* and their uncertainties for all event distributions (Marzocchi et al., 2020):

$$b_n = \frac{1}{\bar{M} - M_{min}};$$

$$b = b_n * \frac{N-1}{N};$$

where $\bar{M}$ is the mean magnitude of events above M_{min} ; N is the number of events above M_{min}. We compute the uncertainties using Tinti e Mulargia (1987):

$$\Delta b = \frac{1 - p}{ln10 * \Delta M \sqrt{p * N}}$$

$$\text{with } \Delta M = 0.01,\ p = 1 + \frac{\Delta M}{\bar{M} - M_{min}}$$

Using Z-test, we also test the null hypothesis:

$$H_0: b_1 = b_2$$

$$Z\ test = \frac{b_1 - b_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Where b_1, b_2 (the b-values), N_1, N_2 (the number of events) and σ_1^2, σ_2^2 (the variances) are referred to identified couple of clusters, respectively.

Next, we count the number of foreshocks and aftershocks of each sequence as function of the mainshock magnitude; we fit their logarithmic values with a

regression line, returning the slope and the intercept for each cluster. We also compute the productivity mean values for each magnitude bin (large 0.01).

As last statistical property, we analyze the temporal behavior of each cluster (i.e., the rate). First, we compute the time differences between each mainshock and its foreshocks and aftershocks (T_F, T_A, respectively); second, we use the following formulation to compute the rate:

$$Rate_{for} = \frac{T_{F\_bin}}{diff(bin_F)} * \frac{1}{N_M}; \quad Rate_{aft} = \frac{T_{A\_bin}}{diff(bin_A)} * \frac{1}{N_M};$$

T_{F_bin} are the difference-time values for each logarithmic bin, bin_F, that go from the logarithm of T_F minimum value to the logarithm of T_F maximum value, spaced by 25 (the same for T_A & bin_A).

T_{F_bin} and T_{A_bin} are normalized by $diff(bin_F)$ & $diff(bin_A)$, respectively, (i.e., differences between adjacent elements) and by the number of mainshocks (N_M) of that specific cluster.

We also compute the survival function of the rate for foreshocks ($S_{t_{for}}$) and aftershocks ($S_{t_{aft}}$) for each cluster; survival functions are computed as follows:

$$S_{t_{for}} = 1 - eCDF(T_F);$$

$$S_{t_{aft}} = 1 - eCDF(T_A).$$

4.2.4 Leave one out method

To make the analysis more concise and straightforward, we wonder if all characteristics of the used dataset are needful for the classification, or if there is a subset of them able of achieving the same result. For that, we use a method based on the leave one out strategy: we remove from the dataset one feature at

a time to explore several subsets. For each subset, we repeat the whole declustering procedure, exploring the optimal number of clusters through the average silhouette mean. C_N^i are the *i*nitial clusters identified using the entire dataset (with $N = 1{:}\,N_{opt}$) and C_N^s the clusters identified for a specific *s*ubset. Each initial cluster C_N^i is compared with a new identified cluster C_N^s, the one with which it has the most elements in common. If the difference of elements (i.e., misclassified sequences) between the compared clusters is less than 10%, we go ahead with the analysis by removing another feature from the current subset. Vice-versa, the removed feature has a very strong influence on cluster analysis, so it is reinserted into the subset and another one is removed. In Table 4.1 we figure out the used procedure and, for each explored subset, we specify the removed and (eventually) reinserted feature, the optimal number of clusters (k) and the number of misclassified sequences respect to each initial C_N^i.

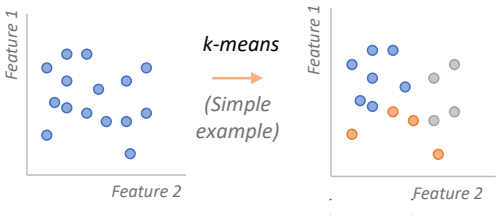
We redo the cluster analysis using the most important features as subset, exploring again the physical properties of new identified clusters. We wonder about the importance of selected features for seismicity classification: we look at their eCDFs for each identified cluster.

1. Collected features (F) of seismic sequences (id)

id	F_1	F_2	F_3
Seq.1	2.3	3.2	2.7
Seq.2	4.5	2.1	4.8
…	….	…	…
$Seq._n$	4.6	7.6	8.8

(Simple example of collected dataset)

2. *k*-means cluster analysis



- Split sequences in *k*-clusters (for *k* = 1:10)
- Use Average silhouette method and Elbow method to identify the optimal number of clusters

3. Statistical properties of identified clusters

For each identified cluster explore:

- Heat flow values of mainshock location
- The Magnitude Frequency Distribution (MFD)
- Foreshock and aftershock productivity
- Foreshock and aftershock rate

4. Leave one out method

	X		
id	F_1	F_2	F_3
Seq.1	2.3	3.2	2.7
Seq.2	4.5	2.1	4.8
…	….	…	…
$Seq._n$	4.6	7.6	8.8

- Remove one feature at time from the dataset
- Redo *k*-means cluster analysis and identify new clusters (i.e., point 2.)
- Compare the new clusters with the previous identified clusters (i.e., using the complete dataset)

✓ If the number of miss-classified sequences is high, the features is important for the seismic discrimination and must be used into the dataset!

Figure 4-1 Simple explanation of the proposed method: 1) Collect information on seismic sequences as explained in section 4.2.1; 2) perform a k-means cluster analysis on collected dataset and identify the optimal number of clusters to split sequences, as explained in section 4.2.2; 3) look at statistical

properties of each identified cluster, as explained in section 4.2.3; 4) perform the leave one out method to identify the most important features from the collected dataset, as explained in section 4.2.4.

4.3. Results

In section 4.3.1 we show the results of the cluster analysis using the entire dataset; next we search for the optimal number of clusters using two different statistic methods. Finally (section 4.3.2), we study sequences of each identified cluster from several points of view.

In section 4.3.3, we show the results of the leave one out method used to identify a subset with the most influential characteristics; moreover, we do an in-depth analysis to understand the reason for their strong influence in the identification of sequence-types. Finally, we use the selected subset to repeat the cluster analysis and explore each identified seismicity group (section 4.3.4).

4.3.1 Results of cluster analysis using the complete dataset

First, we represent the normalized dataset in form of histograms (Figure 4.2). As explained in the method section, some sequences do not have foreshocks, and null or zero values have been replaced with the minimum positive value (rescaled of one unit) of the feature. They correspond to left bars in some histogram (as relative timing, relative duration of foreshocks, area of foreshocks and so on)."

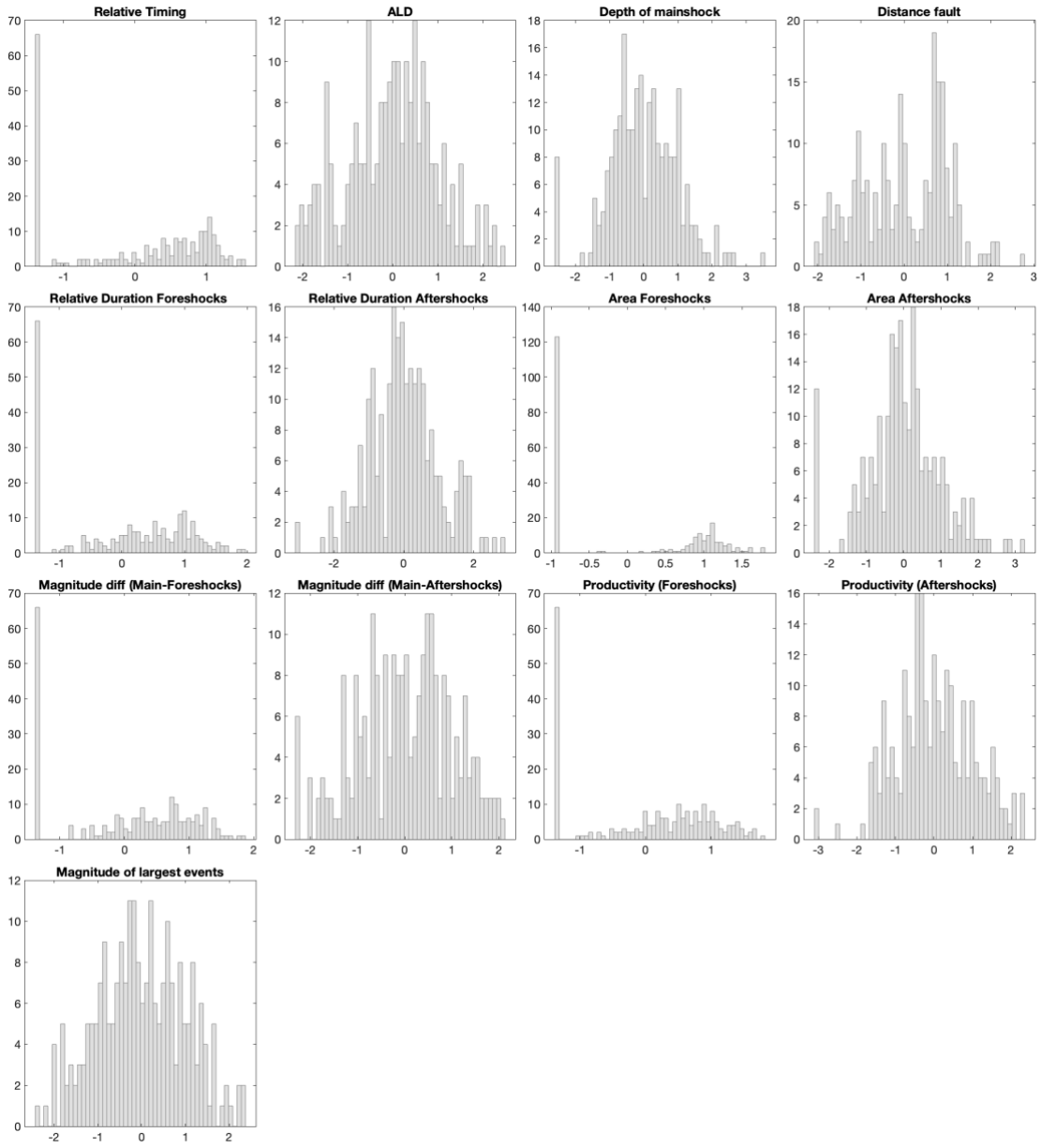


Figure 4-2 Histogram of normalized feature values.

In Figure 4.3 we display the identified clusters for each *k*-value using the silhouette plots (on the left): for each sequence inside the cluster, its silhouette value represents how similar it is to other sequences in the same cluster, respect to sequences in other clusters (Kaufman and Rousseeuw, 1990); values range from –1 (bad matched) to 1 (well matched). We also report the number of sequences belonging to each cluster using stem plots (right plot).

For *k*-values = 2 sequences are better matched to its own cluster having mostly positive silhouette values. This is a useful clue to identify the optimal number of clusters.

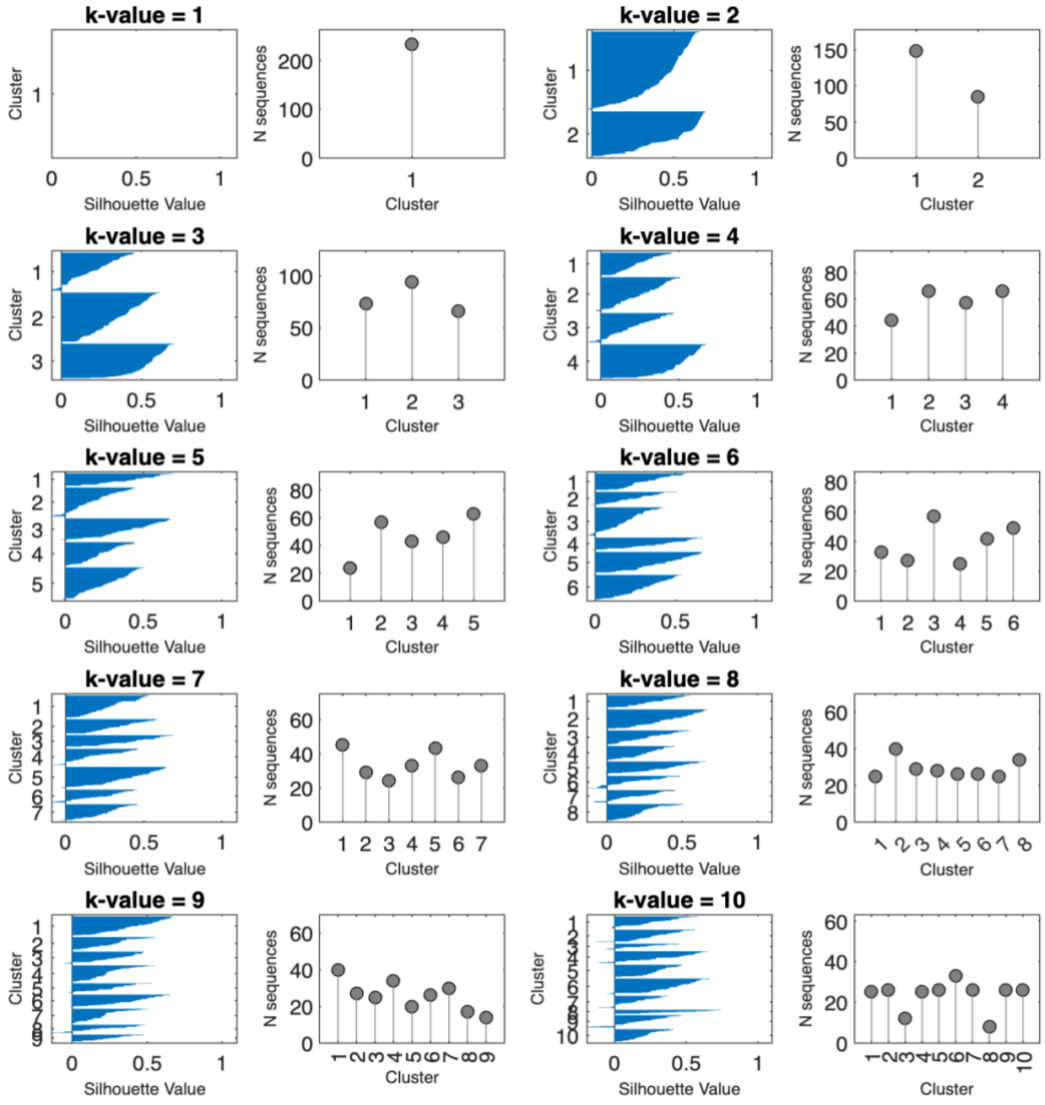


Figure 4-3(On the left) Silhouette plots for each k-value = [1:10]: x-axis corresponds to the silhouette value of each element; y-axis corresponds to the number of identified clusters. (On the right) Corresponding stem plot for each silhouette plot: x-axis represents the number of clusters and y-axis represents the number of sequences inside each cluster.

To properly define how many types of the seismic sequences exist, we present in Figure 4.4 two methods used to identify the optimal number of clusters. The result of *average silhouette mean* is a representation of what we highlighted in

Figure 4.3, i.e., the optimal number of clusters corresponds to *k-value* = 2. However, we decide to also perform the *Elbow method* in order to investigate the existence of only one cluster; this method requires a more subjective interpretation of the results, i.e., selecting the knee of the curve as the optimal number of clusters. However, it seems to agree with the previous results: for N clusters = [2:10] all *wss* values are included in a range between $2*10^3$ and $1*10^3$ with a slightly decreasing rate, while for N cluster = 1 the value rises dramatically to $3*10^3$, making a knee at N cluster = 2.

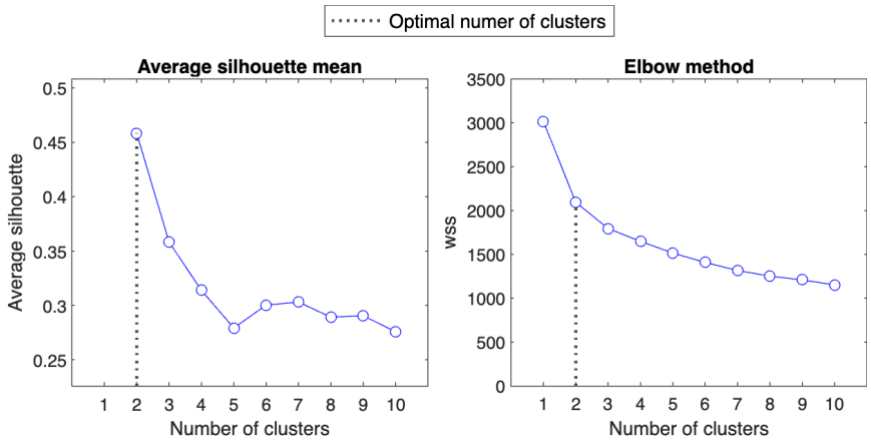


Figure 4-4 Different methods used to identify the optimal number of clusters highlight by the vertical dotted lines.

Selecting *k-value* = 2 as the optimal number, we identify two clusters named C_1^i= 148 sequences and C_2^i = 85 sequences.

4.3.2 Physical interpretation of identified clusters

The main goal of this work is to provide a method for sequence classification, but also, we want to explore the statistical differences among clusters; we analyze the two set of sequences from several points of view, in order to make a final crossing of all information.

First, we deepen the spatial correlation of sequences using heat flow values interpolated at their mainshock location (Figure 4.5); statistical results highlight an important difference: the null hypothesis is rejected for both tests (p-values < 0.05), stating that the heat flow values of the clusters do not belong to the same population.

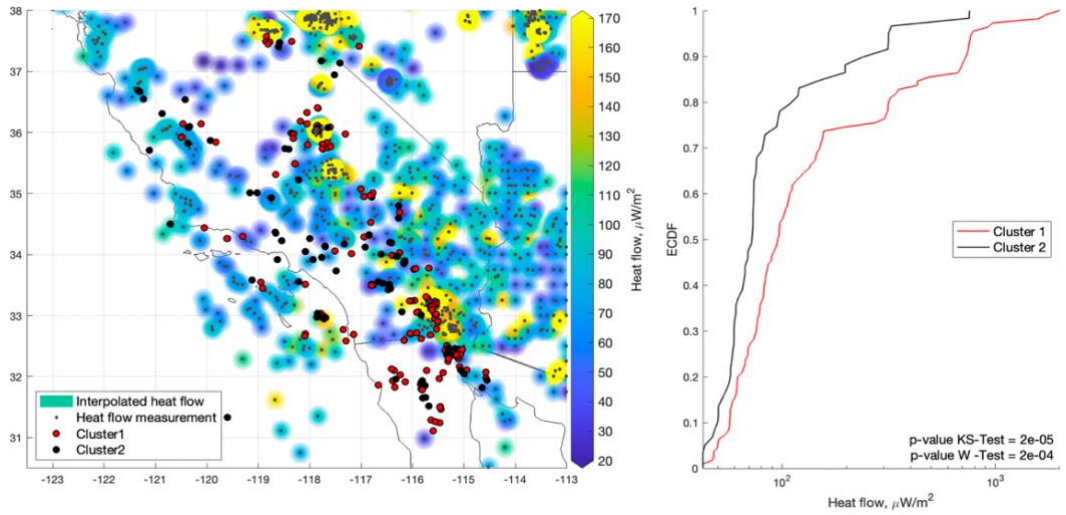


Figure 4-5 On the left, interpolated heat flow map with sampled points and mainshock location of sequences of each cluster; on the right eCDF of interpolated heat flow values of each mainshock of each cluster; eCDF are compared using two statistical tests (see annotation with corresponding p-values).

Second, we compare different clusters in the magnitude domain: Figure 4.6-a shows the Magnitude Frequency Distribution (MFD) of the mainshock magnitudes of sequences; in Figure 4.5-b, we consider the magnitude of all events belonging to each cluster. We test the null hypothesis that the magnitude of events come from the same parent distribution: due to the rejection of the null hypotheses (p-values < 0.05), we can assert that the magnitude of clusters come from different parent distribution (based on the KS-test) with different means (based on the W-test). Testing the b-values using Z-test, the null hypothesis ($H_0: b_1 = b_2$) is rejected as $Zscore \ll -3$ & $p \ll 0.0013$ (Z-score table provided by Davis 2002).

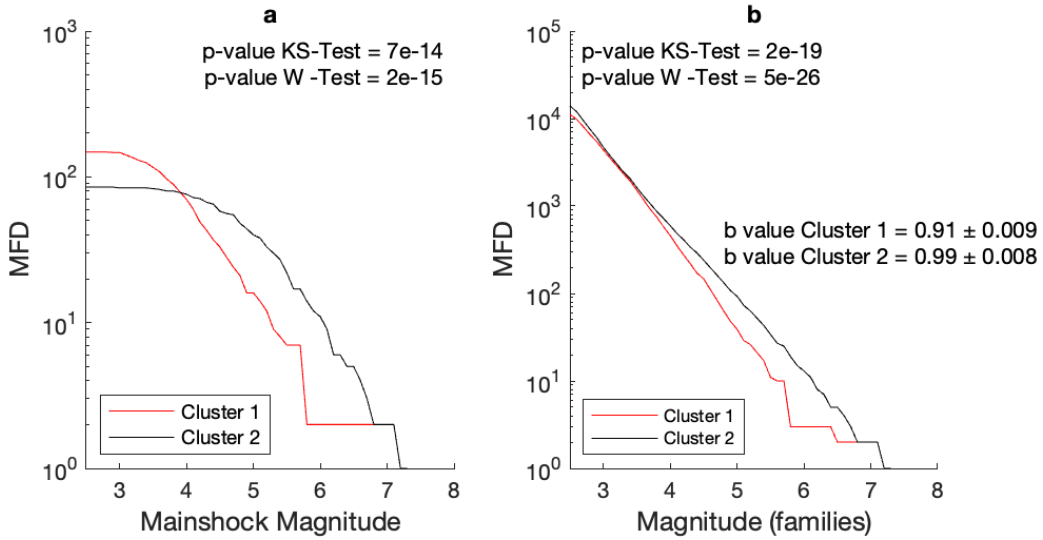


Figure 4-6 a) Magnitude Frequency Distribution of mainshocks for each identified cluster; b) Magnitude Frequency Distribution of all events for each identified cluster; b-value is also computed and reported for each cluster with the uncertainty. MFDs are compared using two statistical tests (see annotation with corresponding p-values).

We show the results of productivity for foreshocks (Figure 4.7, left) and aftershocks (Figure 4.7, right) for each identified cluster. For small mainshock magnitudes ($M_m < 6$), the productivity of foreshocks is generally higher for C_1^i; C_2^i has higher productivity for larger mainshock magnitude (Cluster2 S = 0.4), while C_1^i depends less on the mainshock magnitude (Cluster1 S = 0.2). The aftershock productivity does not differ among the clusters (Fig. 4.7, left).

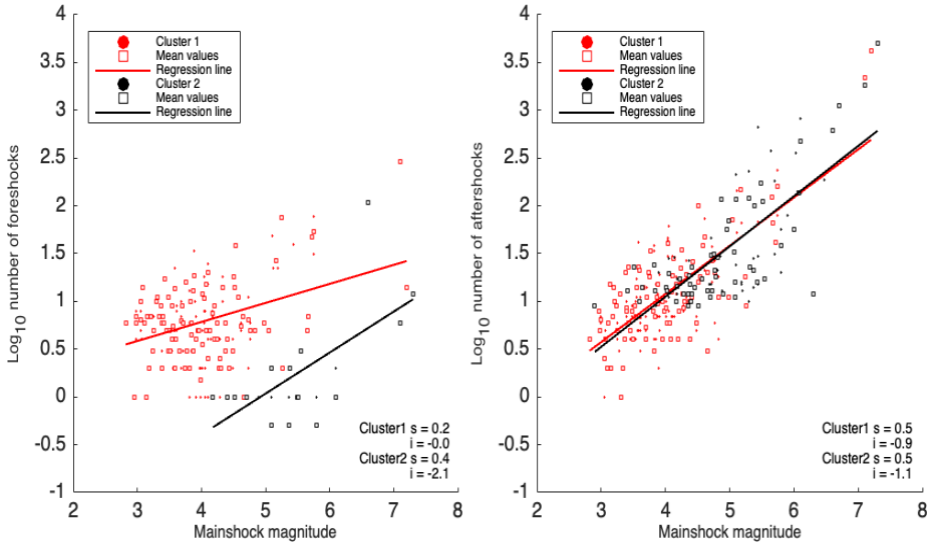


Figure 4-7 Number of foreshock and aftershock for each mainshock magnitude (dots); mean values of number of foreshocks and aftershocks (squares) for each bin of magnitude (large 0.1); solid lines are regression fit of each cluster; we also report the slope values (s) and the intercept (i) of each line.

Figure 4.8 shows the temporal evolution of the foreshock and aftershock rates for both clusters. The foreshock rate of C_1^i (red line) is higher than the one of C_2^i (black line). This trend is confirmed by the survival function in Figure 4.9. For aftershocks, we do not find differences among the clusters.

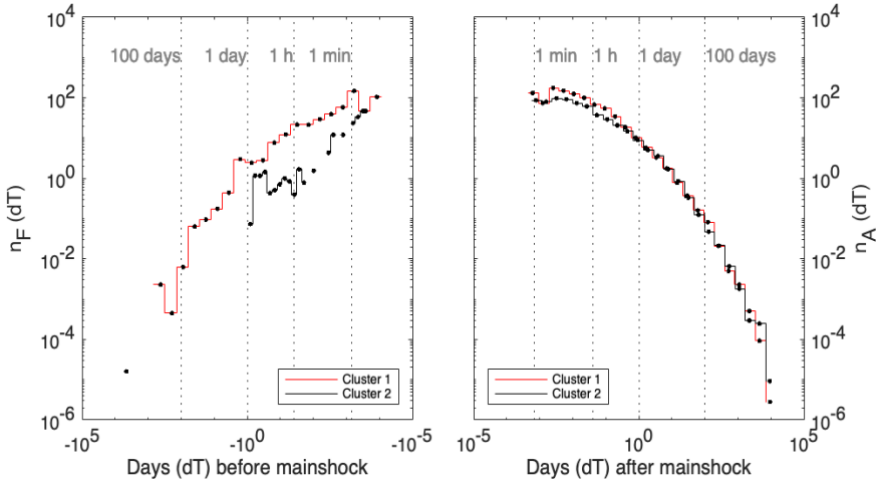


Figure 4-8 Rate for aftershock and foreshock sequences: x-axis is the day of events before or after the mainshocks; y-axis are the normalized time of foreshocks or aftershocks for specific time interval. Different function corresponds to different identified clusters.

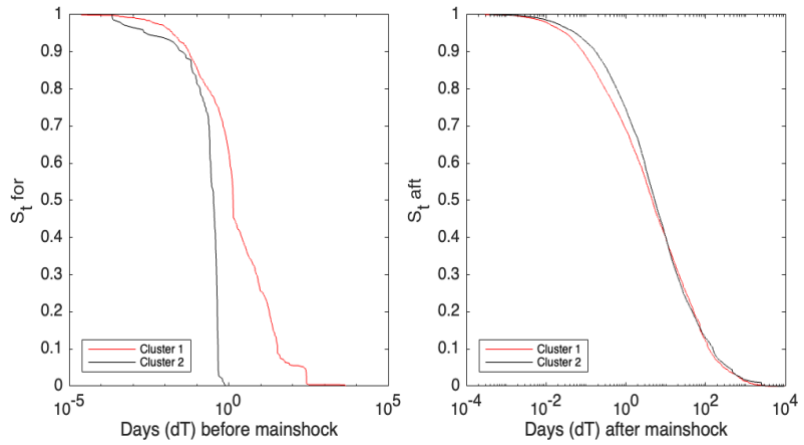


Figure 4-9 Survival function of rate of foreshocks (left) and aftershocks (right) for different identified clusters.

4.3.3 Results of cluster analysis using the most influential subset

In Table 4.1 we show the procedure used to select the most important subset of features to use for cluster analysis; it is easy to note that results do not change until specific features are removed from the dataset: *relative duration of foreshocks*, *area of foreshocks, foreshock productivity, and the magnitude of the largest events* (see columns 5 & 6 of Table 4.1). In these cases, there is a larger number of misclassified sequences respect to the initial clusters (larger than 10%). Following the logic behind the method, the features have a very important role in the sequence classification.

Finally, in Figure 4.10 we present the results of the cluster analysis using only the most important features; the optimal number of clusters is for *k-value* = 2 yet (Figure 4.11) we identify two clusters named C_1^s = 140 sequences and C_2^s = 93 sequences.

Table 4-1 Leave one out procedure to select the needful subset: for each selected subset (column 1), we report the removed feature from the previous subset (column 2), the eventually reinserted feature (column 3), the optimal number of clusters for the current subset (column 4), the number of misclassified sequences respect to cluster C_1^i (column 5) & C_2^i (column 6).

Row	Removed feature	Reinserted feature	optimal k-value	Miss-classified events C1	Miss-classified events C2
ALD – Depth of mainshock Distance fault – Relative Duration Foreshocks Relative Duration Aftershocks – Area Foreshocks Area Aftershocks – Magnitude diff (Main-Foreshocks) Magnitude diff (Main-Aftershocks) – Productivity (Foreshocks) Productivity (Aftershocks) – Magnitude of largest events	Relative Timing		2	7	2
Depth of mainshock – Distance fault Relative Duration Foreshocks – Relative Duration Aftershocks Area Foreshocks – Area Aftershocks Magnitude diff (Main-Foreshocks) – Magnitude diff (Main-Aftershocks) Productivity (Foreshocks) – Productivity (Aftershocks)	ALD		2	11	3

Magnitude of largest events -					
Distance fault – Relative Duration Foreshocks Relative Duration Aftershocks – Area Foreshocks Area Aftershocks – Magnitude diff (Main-Foreshocks) Magnitude diff (Main-Aftershocks) – Productivity (Foreshocks) Productivity (Aftershocks) – Magnitude of largest events	Depth of mainshock		2	11	3
Relative Duration Foreshocks – Relative Duration Aftershocks Area Foreshocks – Area Aftershocks Magnitude diff (Main-Foreshocks) – Magnitude diff (Main-Aftershocks) Productivity (Foreshocks) – Productivity (Aftershocks) Magnitude of largest events -	Distance fault		2	11	3
Relative Duration Aftershocks – Area Foreshocks Area Aftershocks – Magnitude diff (Main-Foreshocks)	Relative Duration Foreshocks	Relative Duration Foreshocks	3	55	20

Magnitude diff (Main-Aftershocks) – Productivity (Foreshocks) Productivity (Aftershocks) – Magnitude of largest events					
Relative Duration Foreshocks -Area Foreshocks – Area Aftershocks Magnitude diff (Main-Foreshocks) – Magnitude diff (Main-Aftershocks) Productivity (Foreshocks) – Productivity (Aftershocks) Magnitude of largest events -	Relative Duration Aftershocks		2	5	2
Relative Duration Foreshocks – Area Aftershocks – Magnitude diff (Main-Foreshocks) Magnitude diff (Main-Aftershocks) – Productivity (Foreshocks) Productivity (Aftershocks) – Magnitude of largest events	Area Foreshocks	Area Foreshocks	2	17	3
Relative Duration Foreshocks – Area Foreshocks Magnitude diff (Main-Foreshocks) – Magnitude diff (Main-Aftershocks) Productivity (Foreshocks) – Productivity (Aftershocks) Magnitude of largest events	Area Aftershocks		2	4	2

Relative Duration Foreshocks – Area Foreshocks Magnitude diff (Main-Aftershocks) – Productivity (Foreshocks) Productivity (Aftershocks) – Magnitude of largest events	Magnitude diff (Main-Foreshocks)		2	14	1
Relative Duration Foreshocks – Area Foreshocks Productivity (Foreshocks) – Productivity (Aftershocks) Magnitude of largest events -	Magnitude diff (Main-Aftershocks)		2	12	0
Relative Duration Foreshocks – Area Foreshocks Productivity (Aftershocks) – Magnitude of largest events	Productivity (Foreshocks)	Productivity (Foreshocks)	2	16	1
Relative Duration Foreshocks – Area Foreshocks Productivity (Foreshocks) Magnitude of largest events -	Productivity (Aftershocks)		2	8	0
Relative Duration Foreshocks – Area Foreshocks Productivity (Foreshocks)	Magnitude of largest events	Magnitude of largest events	3	42	3

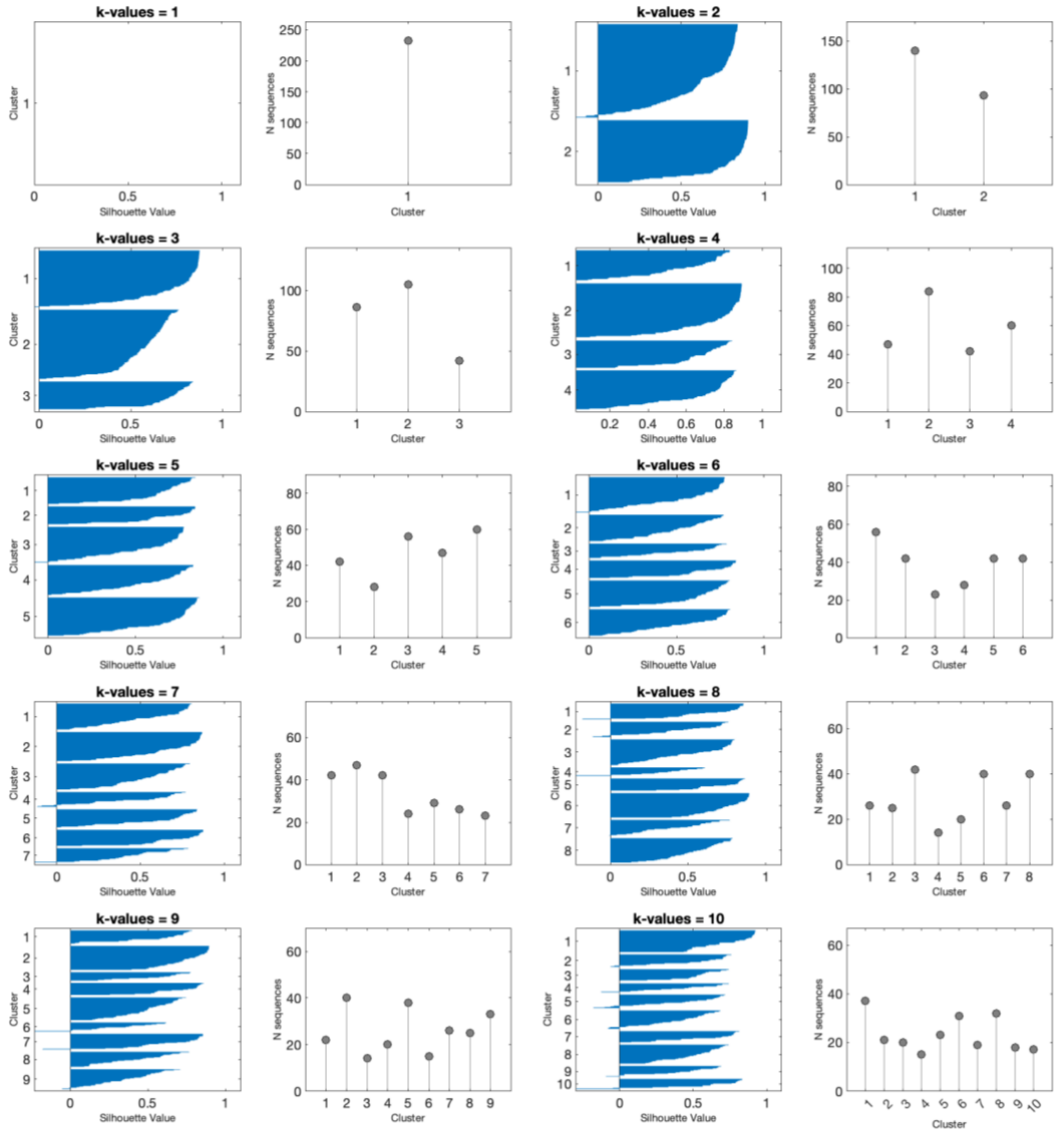


Figure 4-10 Same plots as Figure 4.3 but using the most important features for cluster analysis.

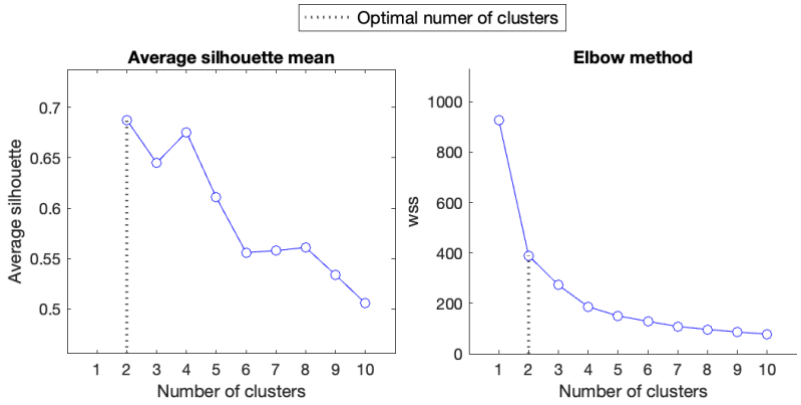


Figure 4-11 Same plots as Figure 4.4 but using the most important features for cluster analysis.

This prompt the question of why the selected features are the most important ones for the classification of seismic sequences. In this regard, we go into the details looking at the eCDFs of features for each identified cluster (Figure 4.12). The number of foreshocks seems to be the discriminating factor: C_2^s (black line) represents a group of sequences not having a huge number of foreshocks respect to C_1^s (red line), for this reason the relative duration, area and productivity values are lower than others. In addition, the magnitude of the largest events plays an important role: C_1^s is characterized by sequences with small-medium mainshock magnitude (no events with $M_M \geq 6$).

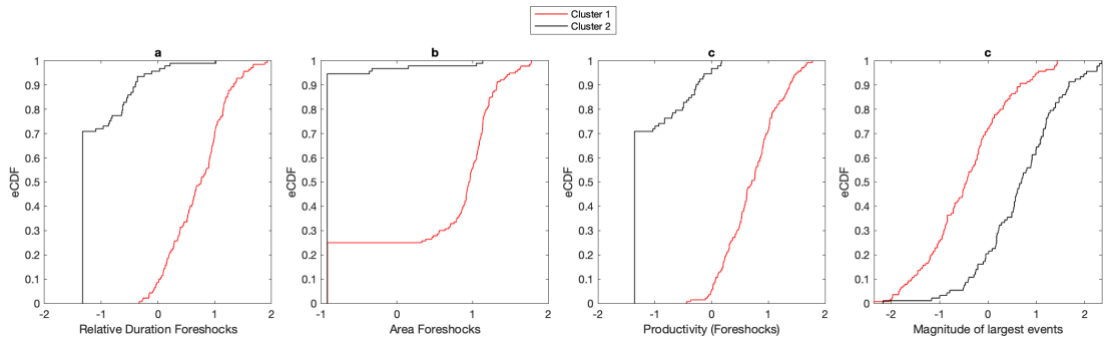


Figure 4-12 eCDFs of most important features: a) relative duration of foreshocks, b) area of foreshocks, c) foreshock productivity and d) the magnitude of the largest event for each identified cluster.

In the next section, we analyze in detail the sequences belonging to each single cluster identified with the most important features.

4.3.4 Physical interpretation of identified clusters using most important features

We saw that, using the most important features as dataset for cluster analysis, seismic sequences may be classified into two different clusters. From Figure 4.13 to Figure 4.17, we deepen physical reasons behind them as we show in Section 4.3.2 (using complete dataset).

C_1^s is located mostly in areas of high heat flow respect to C_2^s (Figure 4.13). This trend is confirmed by the p-values of the two-sample Kolmogorov-Smirnov and paired Wilcoxon tests (see annotations in Figure 4.13): being below 0.05, C_1^s come from different parent distributions with different medians respect to C_2^s. We obtain the same statistical results for magnitude frequency distribution MFDs (Figure 4.14); as we did in section 4.3.2, we perform Z-test to test the null hypothesis: $H_0: b_1 = b_2$. Also in this case H_0 is rejected as $Zscore \ll -3 \;\; \& \;\; p \ll 0.0013$ (Z-score table provided by Davis 2002). Furthermore, C_1^s has higher foreshock productivity values (Figure 4.15) and the slope of the straight line is lower than other as depends less on the mainshock magnitude (see annotation in Figure 4.15); in addition, the foreshock rate values (Figure 4.16) indicate a higher foreshock number per time unit (red line is above the other) as we see for the survival function (Figure 4.17, the red line is below the black one).

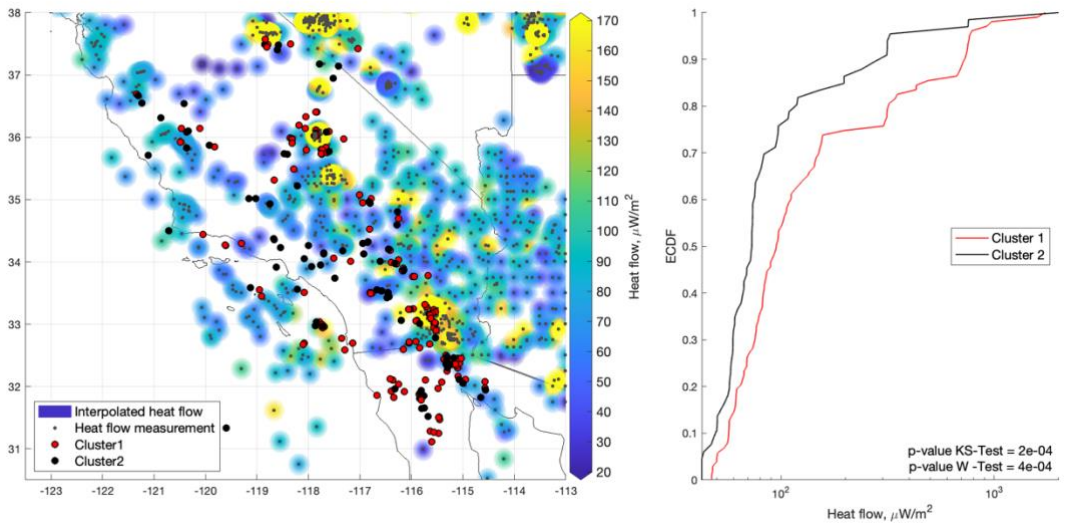


Figure 4-13 As Figure 4.5 but using the most important features for cluster analysis.

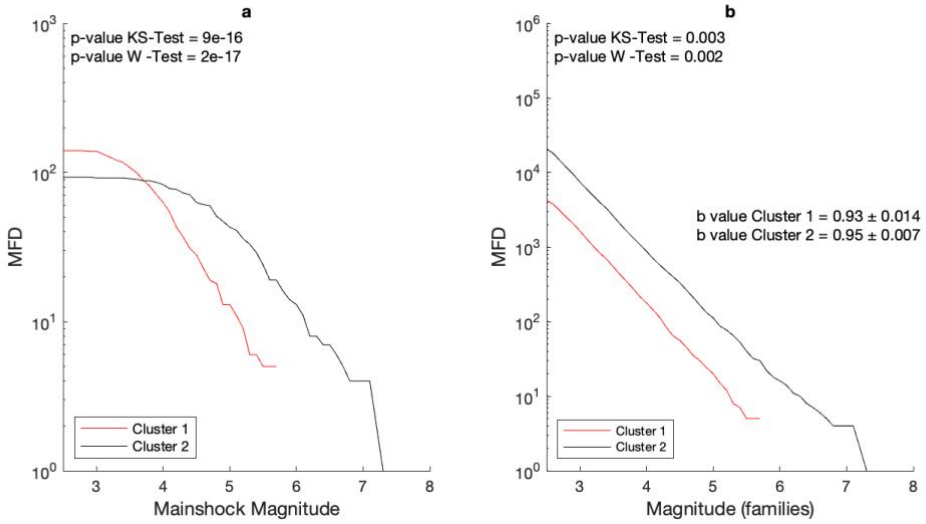


Figure 4-14 As Figure 4.6 but using the most important features for cluster analysis.

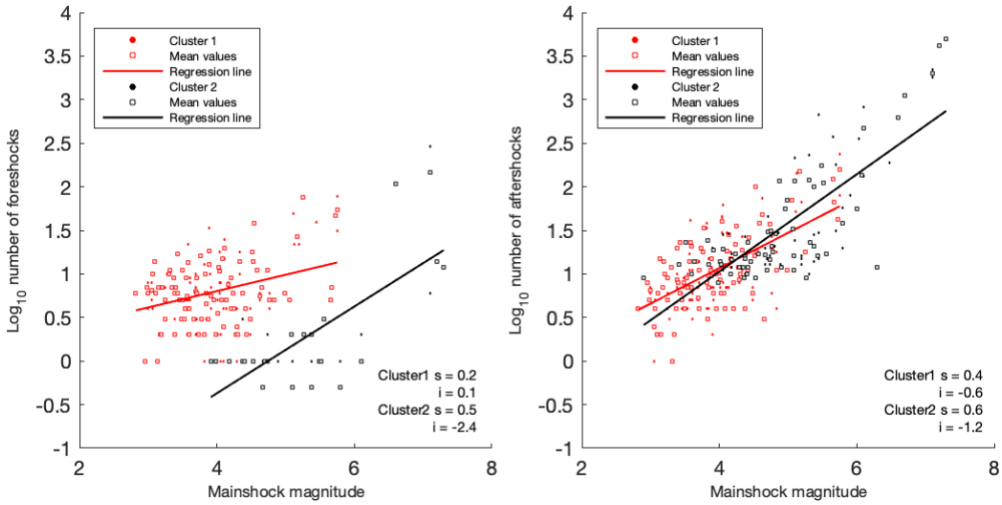


Figure 4-15 As Figure 4.7 but using the most important features for cluster analysis.

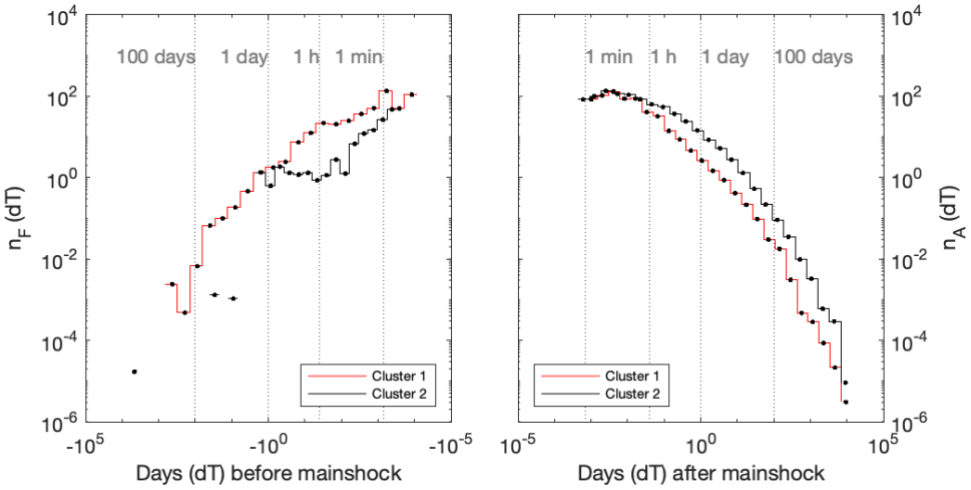


Figure 4-16 As Figure 4.8 but using the most important features for cluster analysis.

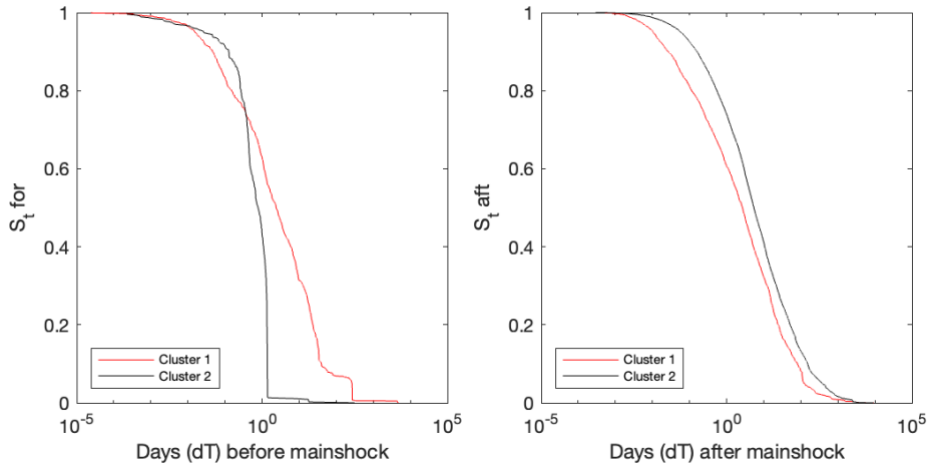


Figure 4-17 As Figure 4.9 but using the most important features for cluster analysis.

4.4 Discussion and Conclusion

The main goal of this study is to provide a quantitative, objective and statistic method to characterize seismic sequences and split them in different groups. The reason for doing that is simple: we want to understand if there are physical differences between the seismic sequences and therefore different mechanisms that originate them. In fact, due to a gap in the definition of sequence-types, it is hard to identify sequences in a certain category rather than another. Our study

may represent the solution for this type of problem: for any sequence it is possible to determine the cluster to which it belongs, due to the simplicity of the method and the easy availability of data it needs. The idea comes from previous works that have already highlighted important variation among sequences, in terms of their location (Manganiello et al., 2023; Zaliapin and Ben Zion, 2013) and number of foreshocks (Seif et al., 2019; Petrillo & Lippiello, 2021; Moutote et al., 2021). Here, we decided to quantify these differences considering space, time, and magnitude of a seismic dataset for southern California.

Initially, the results of the *k-means* cluster analysis allow us to identify two main groups; we wonder if all used dataset is necessary for the analysis, or if there are more important features useful for sequence classification. Through the leave one out method, we find that the most influential ones are: *relative duration of foreshocks, area of foreshocks, foreshock productivity and the magnitude of the largest event*. Performing the cluster analysis using this selected subset, we are able to identify two main clusters again, but with a slightly different number of sequences for each cluster.

The question arises spontaneously: why they are so influential on the cluster identification? Analyzing the eCDF of their values, we can see that the number of foreshocks is the strongest influential parameter: C_2^s has sequences with no or fewer foreshocks than those in C_1^s; C_1^s is also characterized by sequences with small-medium mainshock magnitude.

By comparing the heat flow values, the MFDs, the productivity and the rate of foreshocks and aftershocks of sequences, we began to outline the seismicity types: i) the sequences of C_2^s occur in areas of low flow and are characterized by low productivity of the foreshocks for different orders of magnitude; foreshocks rates are lower in the hours preceding the mainshocks, and seismicity is more focused after mainshocks; ii) sequences of C_1^s have higher

productivity of foreshocks, are located in areas of high heat flow and have small-medium mainshock magnitude sequences.

In doing so, we have outlined two types of sequences which differ from a physical and objective points of view; they can be more commonly named aftershock-type sequences (C_2^s) and swarm-type sequences (C_1^s). The distinction is much more pronounced using the most important features for cluster analysis: for example, using the whole dataset, also large seismic sequences (undoubtedly of aftershock-type, with mainshock magnitude M > 6) were part of swarm-type cluster C_1^i. These miss-classified sequences affected the productivity, MFD and rate values of swam-type cluster. This result highlights the importance of using only influential features to better appreciate differences among clusters.

Our results are in line with previous studies that identify atypical sequences compared to the widely known aftershock sequences: Zaliapin & Ben Zion (2013) identify swarm-type sequences in high-flow areas characterized by high values of average leaf depth (ALD > 5), i.e., with a topological structure that tends to extend in space and time. These findings are partly in agreement with our results, in particular for the spatial variability of the sequences with the heat flow values; however, ALD proved to be an unneeded feature within our cluster analysis. It should be noted that, in the current study, large seismic sequences (mainshock magnitude M > 7) are taken into consideration, which could influence the effect of the ALD values in the cluster analysis: in fact, Zaliapin and Ben Zion (2013) consider these sequences with a separate category, having intermediate characteristics between two main clusters identified by them (burst and swarm type). In addition, other studies (e.g., Enescu et al., 2009, Chen & Shearer, 2016; Ross et al., 2021) have shown that swarm-like activity is common in areas with high heat flow.

We also use another quantitative feature to split sequences in different types: the *relative timing* proposed by Chen & Shearer, (2016). Even if we cannot consider this characteristic as indispensable for the classification, the findings of Chen & Shearer, (2016) agree with our results: swarm-type sequences have larger relative timings respect to aftershock type sequences, and they are located in high heat flow region. The reason of such classification could be match with the high foreshocks productivity of swarm sequences discovered in this work.

In previous studies, the large number of foreshocks is also a widely debated feature and which seems to characterize swarm-type sequences (Manganiello et al., 2023; Petrillo & Lippiello, 2021; Moutote et al, 2021; Seif et al., 2019); in these works, it has also been shown that such sequences are not predicted by the ETAS models. This suggests that they could follow different laws than those used for the forecast, as the productivity, rate and MFD that we discover here.

As shown by previous study (Farrell et al., 2009), swarm sequences appear to be characterized by high b-values (up to 1.3 ± 0.1). Here, we perform statistical test on b-values and MFDs of identified clusters, and we can confirm that they are completely different due to the rejection of the null hypotheses.

Finally, the versatility of this method lies in the fact that it is possible to consider any characteristic, evaluate its importance and impact on the results. Our procedure may be useful to improve earthquake forecasting in future work, restricting features related only to foreshock information (i.e., before the largest event). We could characterize the on-going sequences, recognize their type (i.e., as aftershock or swarm sequence) and adjust the probability for a larger earthquake accordingly (e.g., lowering it upon indication of a swarm-type sequence).

5

Conclusion

Is it possible to predict large earthquakes? How far in advance? With what probability? These are some of the hottest topics plaguing the scientific community. The interest in being able to do so is twofold: saving human lives and enriching cultural background on the functioning of our Earth-system.

For these reasons, in this PhD thesis we are committed to providing skills that can be used to enhance earthquake forecasting models. To do that, we use different approaches that are aimed at the characterization of the seismic sequences. We believe that deeper knowledge on seismicity, existing typologies, physical parameters influencing them, etc., could be useful information to be implemented in the probabilistic models.

In this regard, in Chapter 2 we analyze foreshocks as possible *precursors*, and we answer the main question: do they have characteristics such as to be recognized a-*priori*? This could definitely improve the probability of large event occurrence. For that, we start from the null hypothesis that ETAS represents all the seismicity, and we compare the foreshocks observed in reality and those of synthetic catalogues. Any discrepancies could assume that foreshocks have some peculiar characteristics that differentiate them from normal seismicity and that would help to recognize them a-*priori*. However, we find none of this. One difference is noteworthy: there are particular sequences that ETAS models are unable to forecast. They occur mainly in areas of high heat flow, characterized by a high number of foreshocks and small-medium mainshock magnitude. In the literature, sequences with similar characteristics are defined as swarm-type,

although there is no objective method to identify them. Several physical mechanisms (e.g., fluids) could explain their occurrence or using not spatially uniform cascade model.

According to what has been identified in the previous chapter, we decide to further investigate what ETAS is unable to forecast. To do this, we use a scalar measure, the Average Leaf Depth, as possible discriminator between burst-like (with small ALD values) and swarm-like (with high ALD values) sequences and investigate if ETAS is able to predict both types. We find incongruous results exploring different ETAS models, with different parametrization and types of sources: the planar source model is characterized by very high ALD values, although they are only related to sequences of large mainshock magnitude (therefore not comparable to swarm-type sequences). Using ETAS models with point-source, the trends in ALD values remain different between those observed and those predicted. Based on that, we assert that the ETAS model used to reproduce synthetic sequences plays a very important role in the calculation of ALD. For this reason, the ability of ALD to differentiate the types of sequences is not objective at all and can be influenced by several parameters.

The finding obtained in Chapter 2 and Chapter 3 arose the need to find (quasi-) real-time methods to discriminate swarm-like from ETAS-like sequences. For this reason, we decided to focus on an analysis that could improve earthquake forecasting: being able to identify an on-going swarm-like sequence could reduce the forecast probability for a future large earthquake.

In Chapter 4 we provide a new quantitative, objective, and statistical method to characterize seismic sequences and split them in different groups. We identify two main seismicity groups which differ from a physical and objective points of view: one having high productivity of foreshocks, located in areas of high heat flow, with an important concentration of events before the mainshocks (i.e., swarm-type cluster). The other occur in areas of low heat flow, with low

productivity and rate of foreshocks for different orders of mainshock magnitude (i.e., aftershock type cluster). Noteworthy, the characteristics that allows the best discretization of seismicity are the *relative duration of foreshocks, area of foreshocks, foreshock productivity and magnitude of largest event.*

These results are not surprising but confirm the results of Chapter 2 and 3: sequences not reproduced by ETAS are those with a high number of foreshocks and small-medium mainshock magnitude; consequently, we would expect that such sequences delineate a larger area, duration, and productivity of foreshocks.

5.1 Outlook

In this thesis, I hope to have contributed useful insights that could help to improve earthquake forecasting models. However, various issues still remain unresolved or need to be investigated. Here, I list some suggestions that should be considered in future research:

- In Chapters 2 and 3 we have better outlined the discrepancies between the ETAS model and the real seismicity; many elements led us to think that such discrepancy is due to swarm seismicity. We have also shown that the heat flow values could strongly influence the occurrence of swarm-type sequences. For this reason, we believe that it is necessary to deepen the physical reasons that cause these anomalous sequences.
- In Chapter 4 we perform a cluster analysis collecting a dataset of characteristics that can represent earthquake sequences in different domains (space, time, and magnitude). The method is versatile as it can consider any feature, evaluate its importance in the discriminative power. Therefore, our method can be used in future research to explore the impact that any other feature has on the characterization of seismicity. This could help to understand the mechanisms behind the origin of the different types of seismicity.

- Finally, we remind to stay always interested about the things that surround our life, understand their mechanisms, and derive benefits from them (*science is power*, and knowledge help to face the dangers of life, from the point of view of a *sapient*).

SI.1

Supporting Information for Chapter 2

This is published as Supporting Information of:

Manganiello, E., Herrmann, M., & Marzocchi, W. (2023). New physical implications from revisiting foreshock activity in southern California. *Geophysical Research Letters*, *50*, e2022GL098737. https://doi.org/10.1029/2022GL098737

Introduction

The supporting information contains additional information about the ETAS models used for the analyses and their verification. It also reports the results using alternative methods to infer and select anomalous foreshock sequences (e.g., the spatiotemporal windows (STW) method to identify mainshocks and their foreshocks, using a significance level of 95%, using an alternative analysis of TEST1, and using an independent dataset of anomalous foreshock sequences).

SI.1.1 Used ETAS models

We used the stochastic ETAS aftershock simulator program developed by K. Felzer (see Felzer et al., 2002 and Table S1), which is calibrated to southern California using spatially invariant triggering parameter values of Hardebeck et al., 2008, see Table S2. The program simulates background events and triggered

earthquakes in time and space. It makes use of Monte Carlo methods and empirical aftershock relationships following the ETAS model of Ogata (1988).

We use another ETAS model proposed by Seif et al. (2017) with spatially invariant triggering parameters calibrated for southern California. The model simulates background events and triggered earthquakes in time and space using the relationships of Ogata (1988). The supplement of Seif et al. (2017) explains each parameter of the model as well as the maximum likelihood estimation (MLE) procedure. We report their estimated parameters in Table S3.

We prefer to use reliable (see Text S2) and independent (made by someone else) ETAS models for a few important reasons that are worth being mentioning. First, we lower the degrees of freedom of our analyses, avoiding that we, unconsciously, tune our results. Second, we make our results more comparable with previous analyses that used such classical ETAS models. Third, by using two different ETAS models with different parameterization, we demonstrate the stability of our results. Fourth, although more complex ETAS models should fit past data better, this does not mean that they are more reliable; results of the first prospective experiments carried out by the Collaboratory for the Study of Earthquake Predictability (CSEP) show that simple models may outperform more complex models (e.g., Nanjo et al., 2012).

SI.1.2 Verifying the reliability of the ETAS models

For the purpose of this work, we verify the reliability of both ETAS models to produce realistic earthquake rates for different magnitude ranges. Specifically, we adopt a Turing-style test (Page & van der Elst, 2018) by comparing the number of events in the real catalog with the distribution of simulated earthquakes in the synthetic catalogs (Figure S1 and S2). We also apply the same kind of analysis to different earthquake magnitude classes $\boldsymbol{C_M}$ = {2.5 ≤

$M < 3.0$, $3.0 \leq M < 4.0$, $4.0 \leq M < 5.0$, $5.0 \leq M < 6.0$, $M \geq 6.0$} (Figure S3 and S4). In all cases, the real observation (solid vertical line) is within the 95% confidence interval (vertical dashed lines), indicating that the ETAS models produce synthetic catalogs with earthquake rates that are consistent with the observations for different magnitude ranges.

SI.1.3 Alternative analysis of TEST1

For TEST1, we argue that the approach of labeling anomalous and normal foreshock sequences is conservative, because comparing a single sequence against the average behavior of foreshock sequences may lead to wrongly label more actual normal foreshock sequences as 'anomalous' (i.e., false positives) than wrongly labeling anomalous foreshock sequences as 'normal' (i.e., false negatives). To investigate this aspect, we perform an alternative analysis by building two eCDFs of N_F (i.e., without normalizing by N_M): one for the real catalog ($eCDF^{real}$) and one for all synthetic catalogs combined ($eCDF^{ETAS}$). If the 99th percentile of $eCDF^{real}$ is larger than the corresponding percentile of $eCDF^{ETAS}$ in a certain class, we label each foreshock sequence as 'anomalous' whose N_F is above the 99th percentile of $eCDF^{ETAS}$. We apply this procedure using K. Felzer's ETAS model and the NN method to identify mainshocks and their foreshocks; results are shown in Figure S6.

SI.1.4 Results for a significance level of 95%

Figures S12 shows the results of TEST1 using K. Felzer's ETAS model and a significance level of 95%. Of a total of 152 foreshock sequences, we found 65 (43%) anomalous foreshock sequences using the NN method. Figure S13 shows the results for TEST2: we found 52 of 152 (34%) foreshock sequences to be anomalous using the NN method.

Table S1 Parameters used in K. Felzer's ETAS simulator.

Start date of simulation	1-1-1981
Start date of synthetic catalogs	1-1-1983
End date of synthetic catalogs	31-12-2019
Lower magnitude limit of active earthquakes, M_0	2.5
Lower magnitude limit in synthetic catalog	2.5
Lower magnitude limit for modelling planar sources	6.5
Upper magnitude limit	7.9
Minimum aftershock distance from parent event	0.001 (km)
Maximum aftershock distance from parent event	500 (km)

Table S2 Used ETAS parameters as given by Hardebeck et al., 2008 for $M_0 = 2.5$. μ is the background rate; K, c, and p are parameters of Omori's law; n is the aftershock decay with distance (exponent in r^{-n}); α is the productivity law exponent (productivity scaling with magnitude); and $b\beta$ is the Gutenberg–Richter b-value multiplied by ln(10) and describes the magnitude distribution.

μ	***K***	***c (days)***	***p***	***n***	***α, β***
spatially variable	**0.008**	**0.095**	**1.34**	**1.37**	**ln(10) ≈ 2.30**

Table S3 Used ETAS parameters as given by Seif et al., 2017 for M_0 = 2.5 and fixed α. For a description of K, α, β, p, and c see Table S2. q, d, and γ describe the spatial distribution of earthquakes.

K	***α, β***	***p***	***c***	***q***	***d***	**γ**
0.0804	**ln(10) ≈ 2.30**	**1.11**	**0.014**	**1.9**	**0.005 km^2**	**1.49**

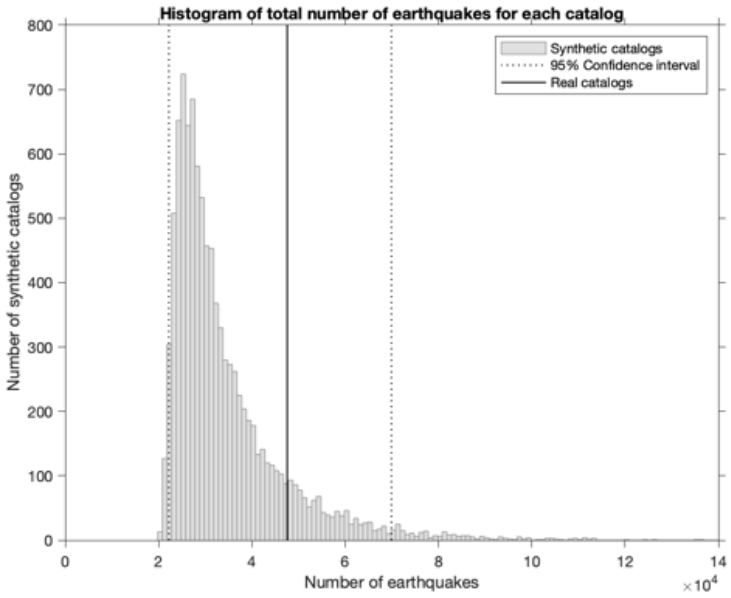
Table S4 p-values for comparing eCDFs of heat flow values at locations of normal versus anomalous foreshock sequences for each analysis.

ETAS model by K. Felzer		**NN-99**	**STW-99**	**NN-95**
TEST1	KS test	0.0004	0.035	0.0003
	W. rank test	0.000006	0.009	0.000004
TEST2	KS test	0.019	0.016	0.003
	W. rank test	0.018	0.047	0.0004

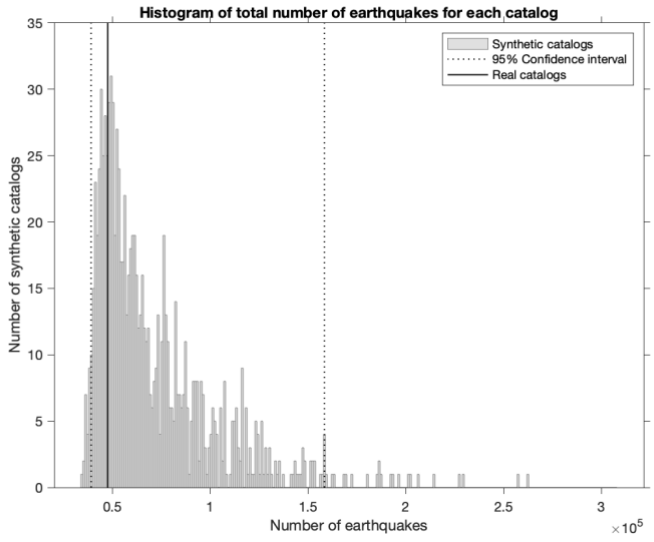
ETAS model by S. Seif		**NN-99**	**STW-99**
TEST1	KS test	0.003	0.0008
	W. rank test	0.0001	0.0002

TEST2	KS test	0.021	0.041
	W. rank test	0.004	0.038

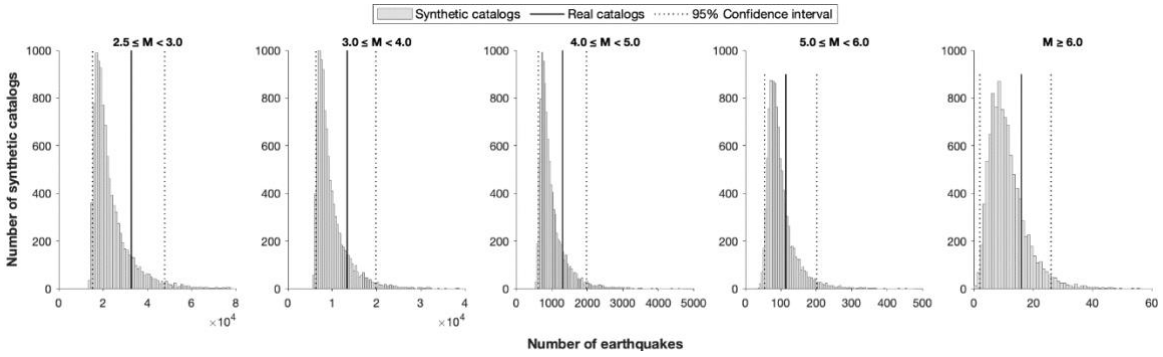
	Alternative TEST1 (ETAS model of K. Felzer, NN-99)	**Independent dataset (G. Petrillo)**
KS test	0.013	0.002
W. rank test	0.004	0.001



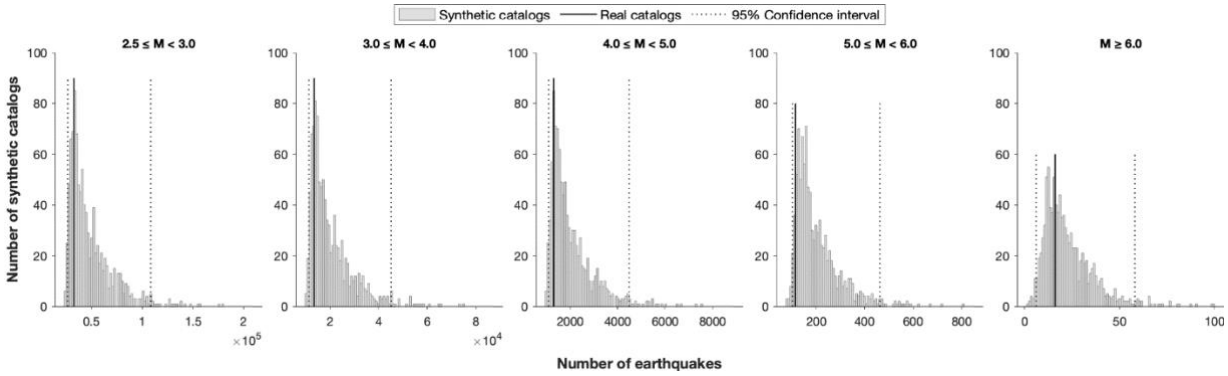
S1 Total number of events in the synthetic catalogs (distribution) and the real catalog (solid vertical line), using events above M2.5. The dashed vertical lines refer to the 95% confidence interval (i.e., the 2.5th – 97.5th percentile range of the distribution). For more information, see Text S2.



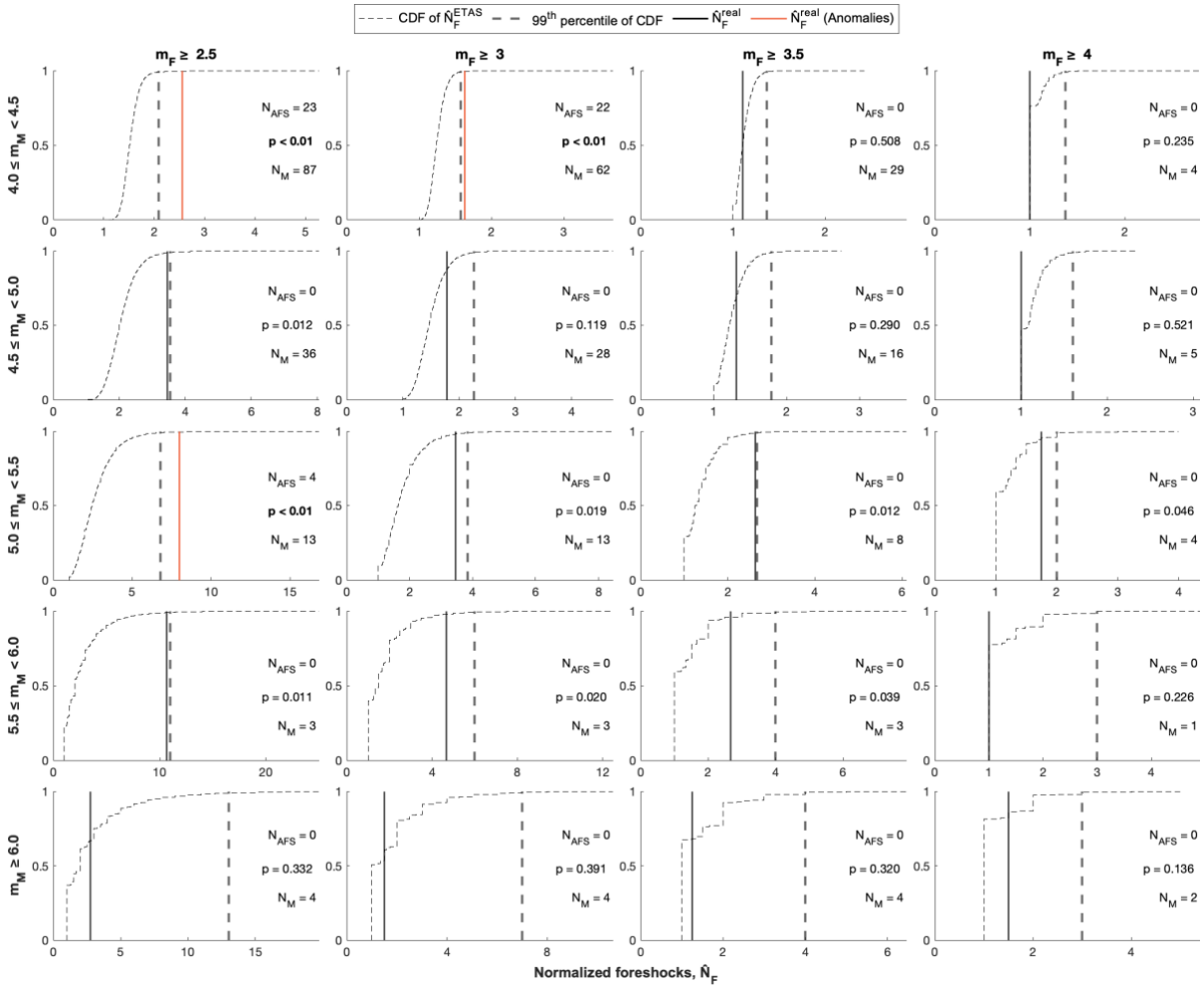
S2 Like Figure S1 but for the second ETAS model (by S. Seif).



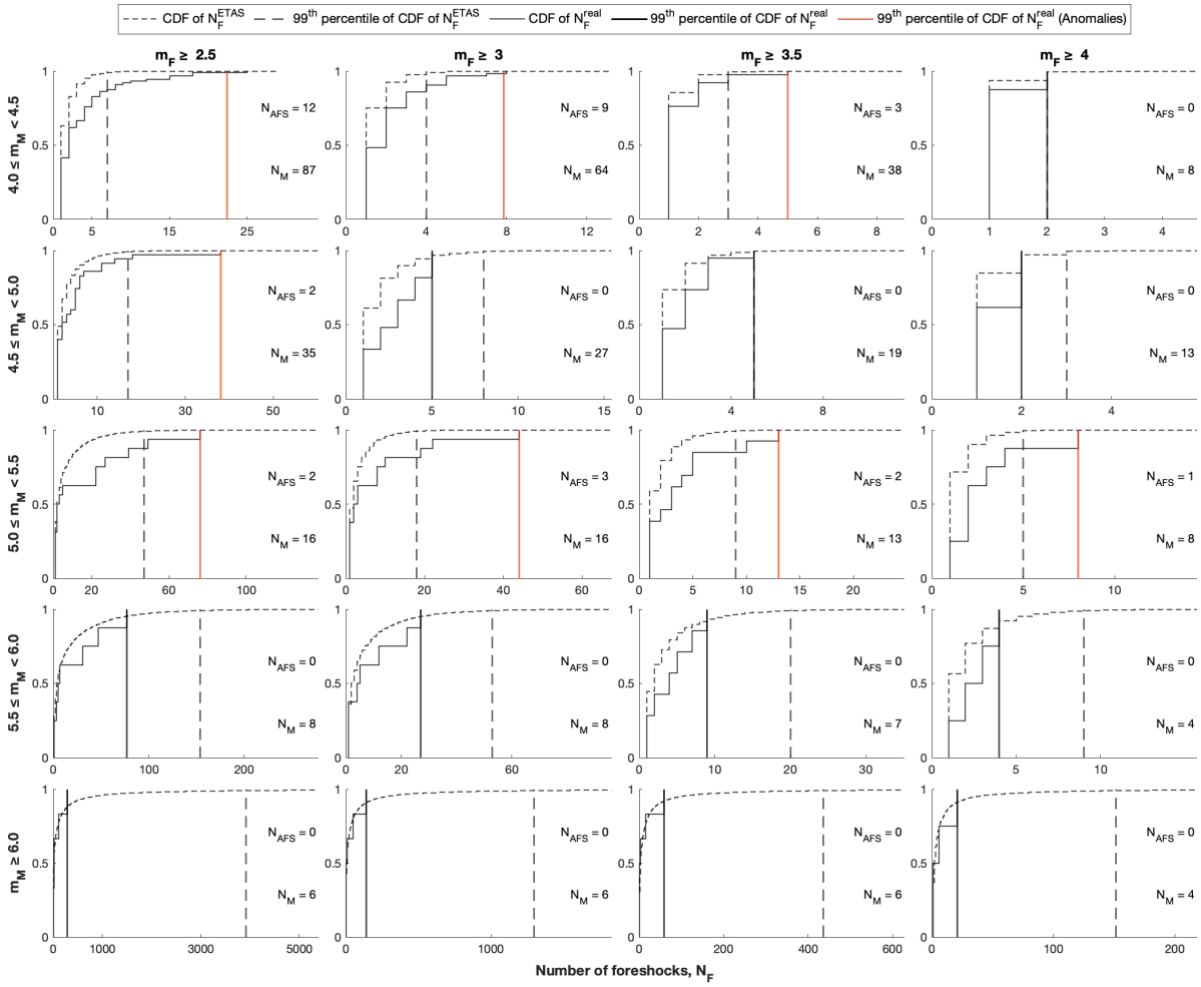
S3 Like Figure S1 but for different magnitude classes.



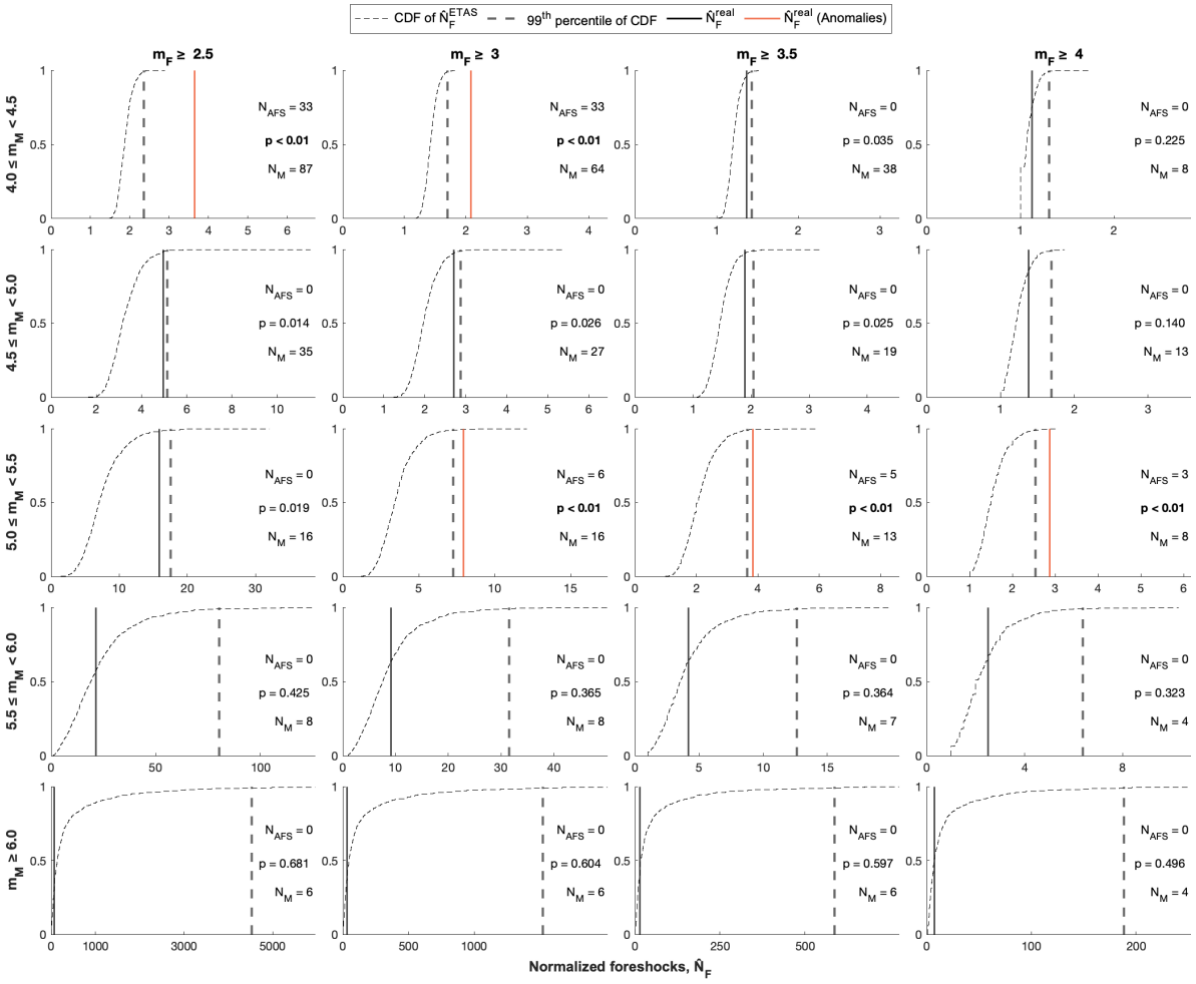
S4 Like Figure S3 but for the second ETAS model (by S. Seif).



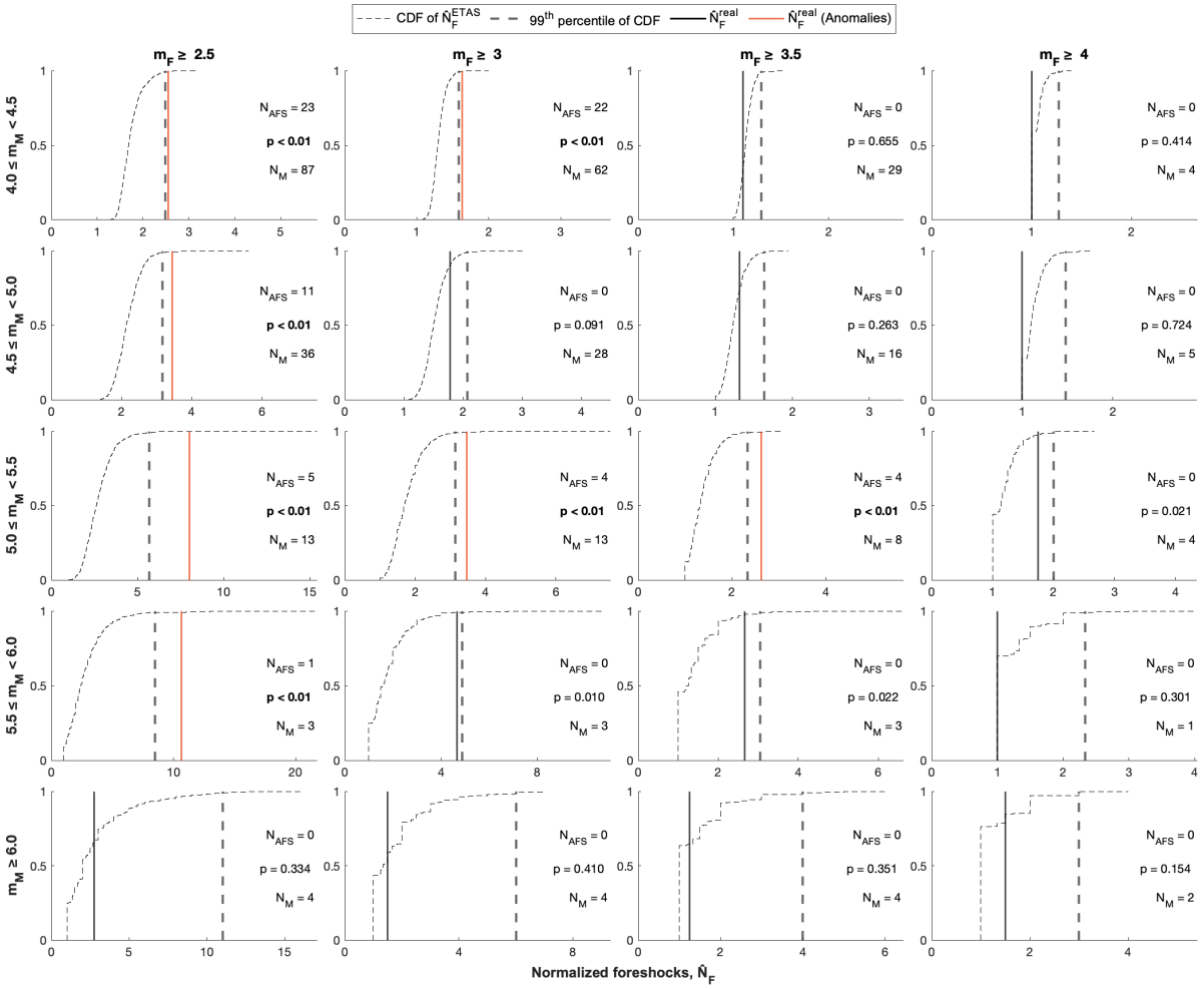
S5 Like Figure 2.1 in Chapter 2 (TEST1) but using the STW method.



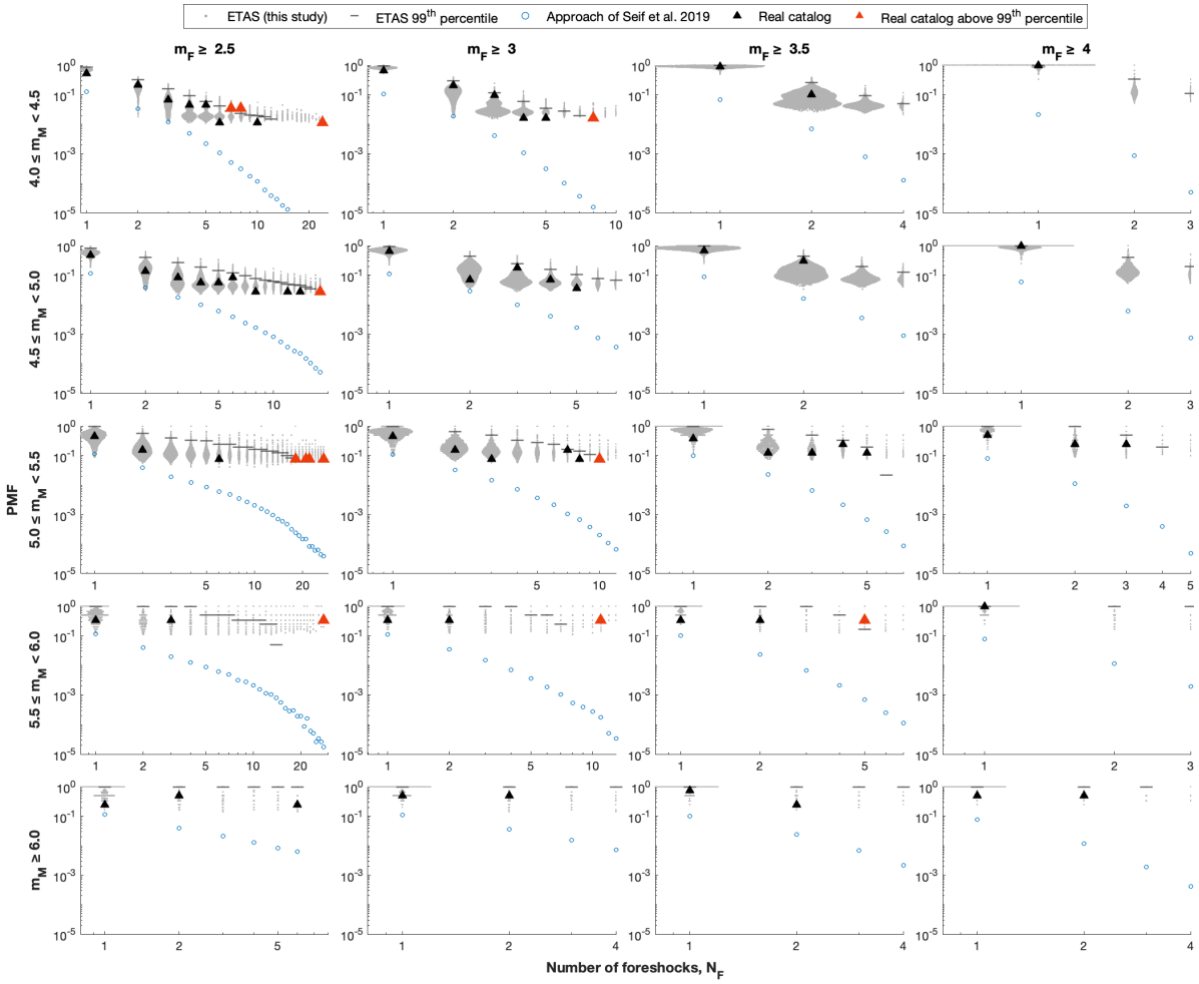
S6 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using the individual number of foreshocks, N_F (alternative analysis of TEST1). In this way, the empirical Cumulative Distribution Function (eCDF) of N_F can be constructed for both the real catalog (solid curve) and all 10'000 synthetic catalogs combined (dashed curve); vertical lines show their corresponding 99th percentile (real catalog: solid; synthetic catalogs: dashed). If the former is above the latter, the solid vertical line becomes red, indicating more anomalous foreshock sequences than expected. Each subplot also reports the number of anomalous foreshock sequences, N_{AFS}, and the number of mainshocks, N_M.



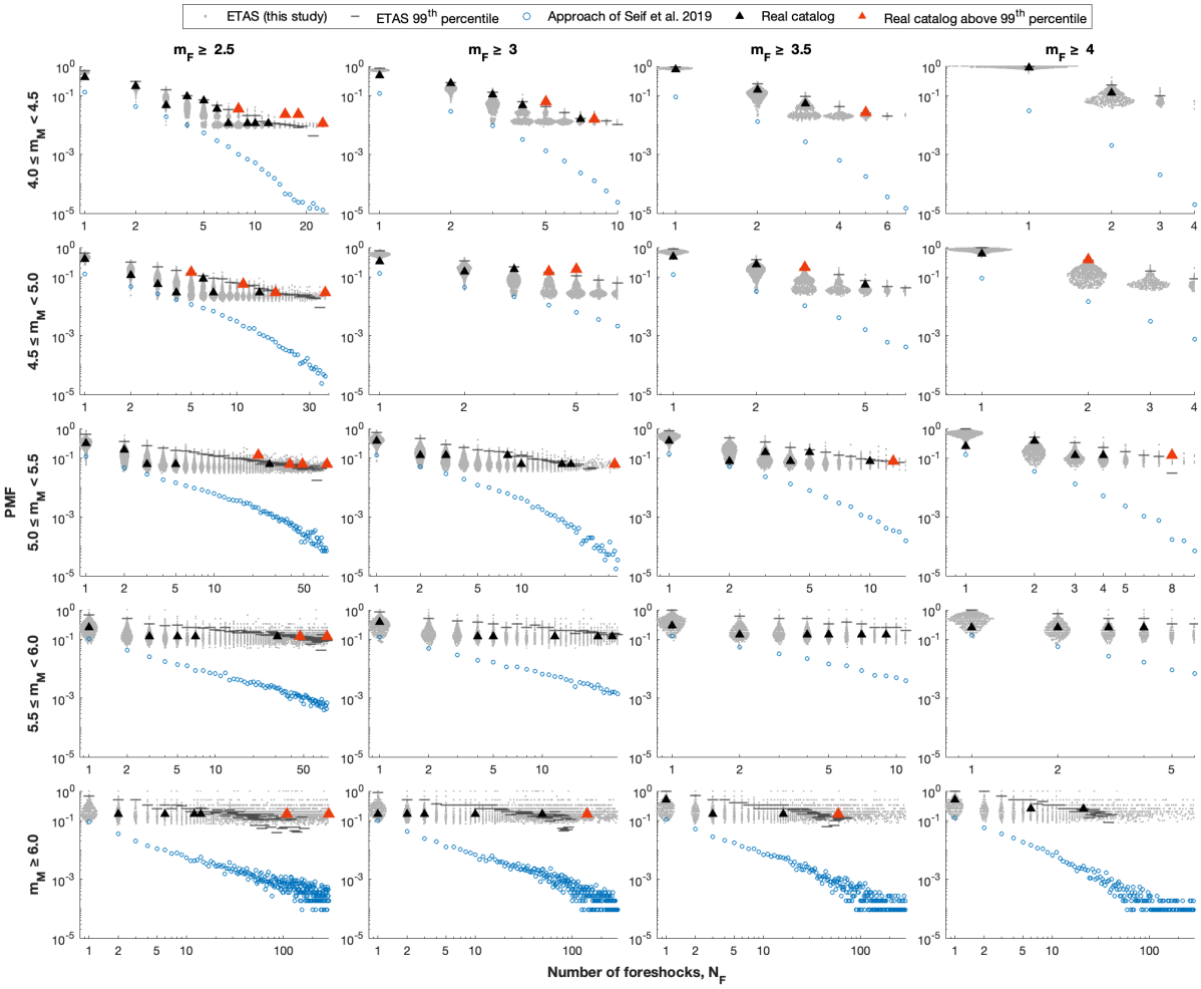
S7 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using the second ETAS model. Results are based on 1000 synthetic catalogs.



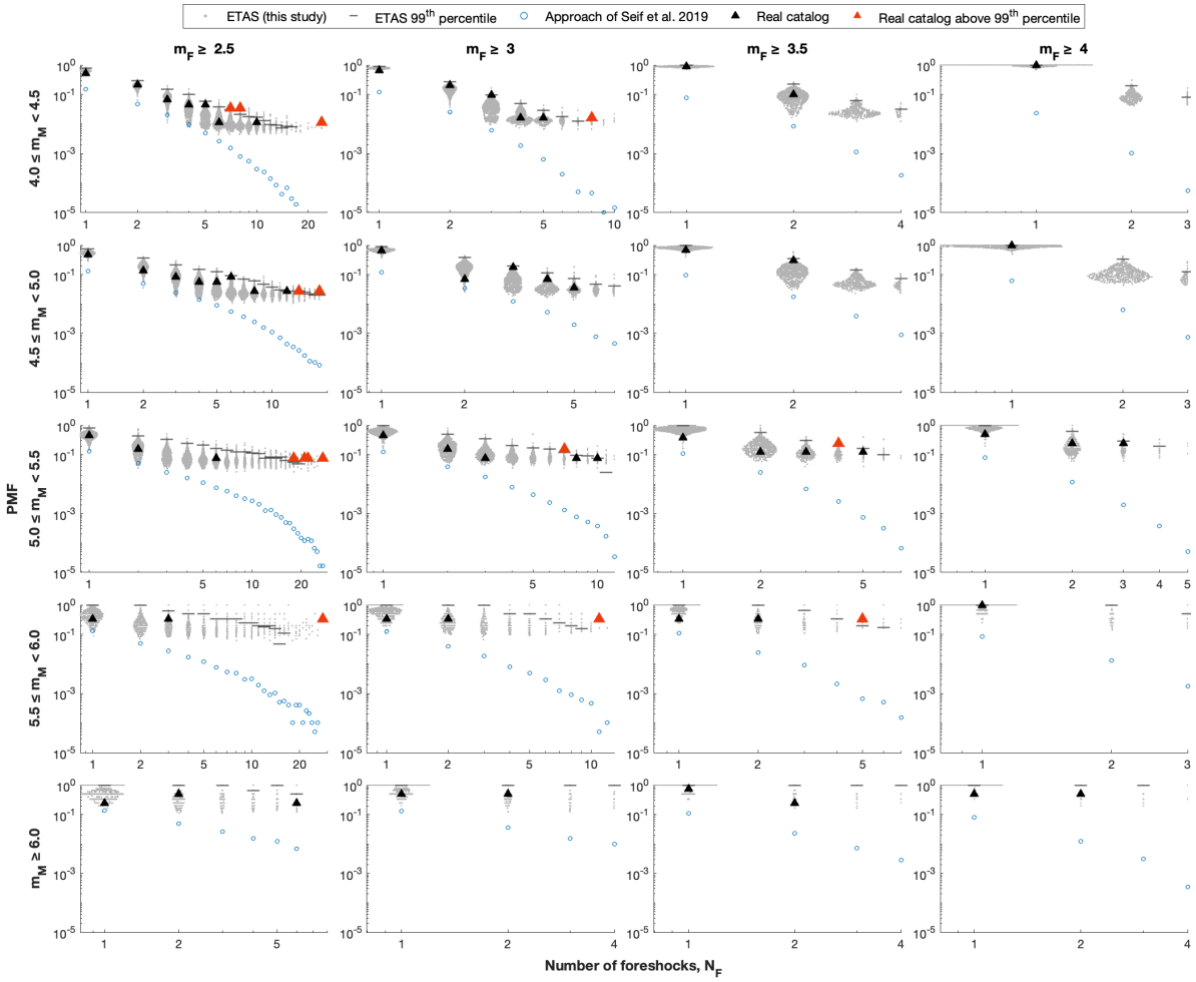
S8 Like Figure S5 (TEST1, STW method) but using the second ETAS model. Results are based on 1000 synthetic catalogs.



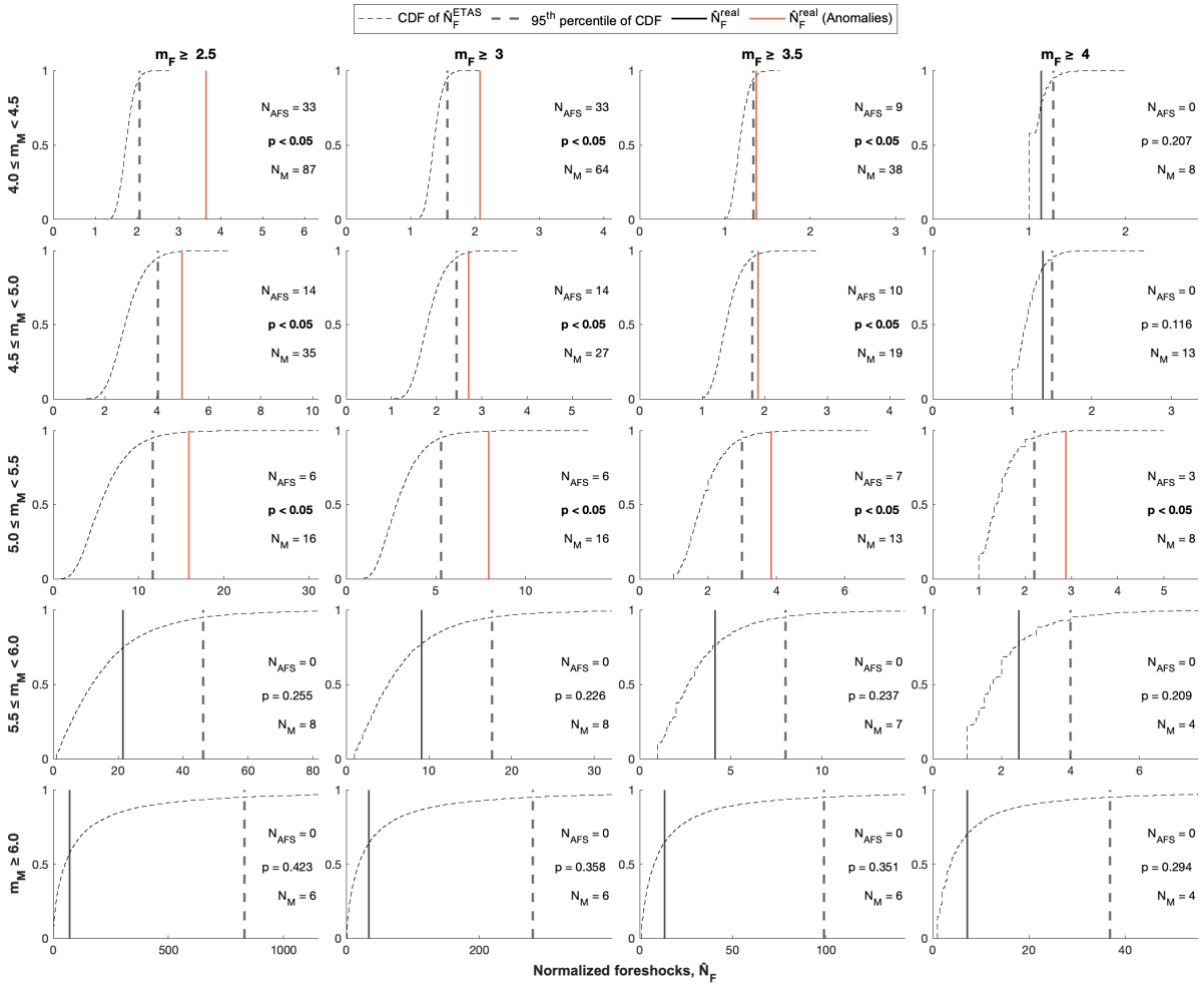
S9 Like Figure 2.2 in Chapter 2 (TEST2) but using the STW method.



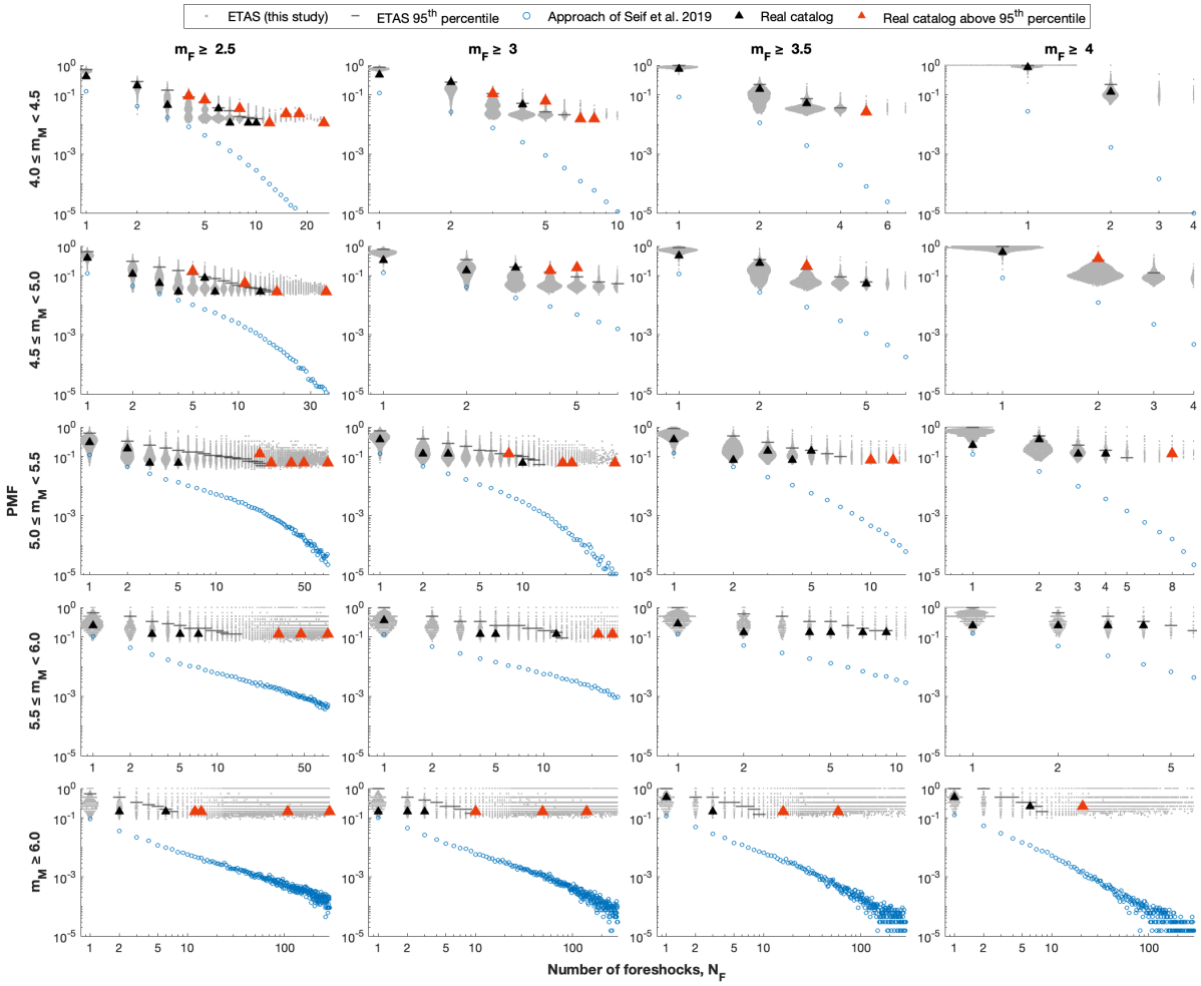
S10 Like Figure 2.2 in Chapter 2 (TEST2, NN method) but using the second ETAS model. Results are based on 1000 synthetic catalogs.



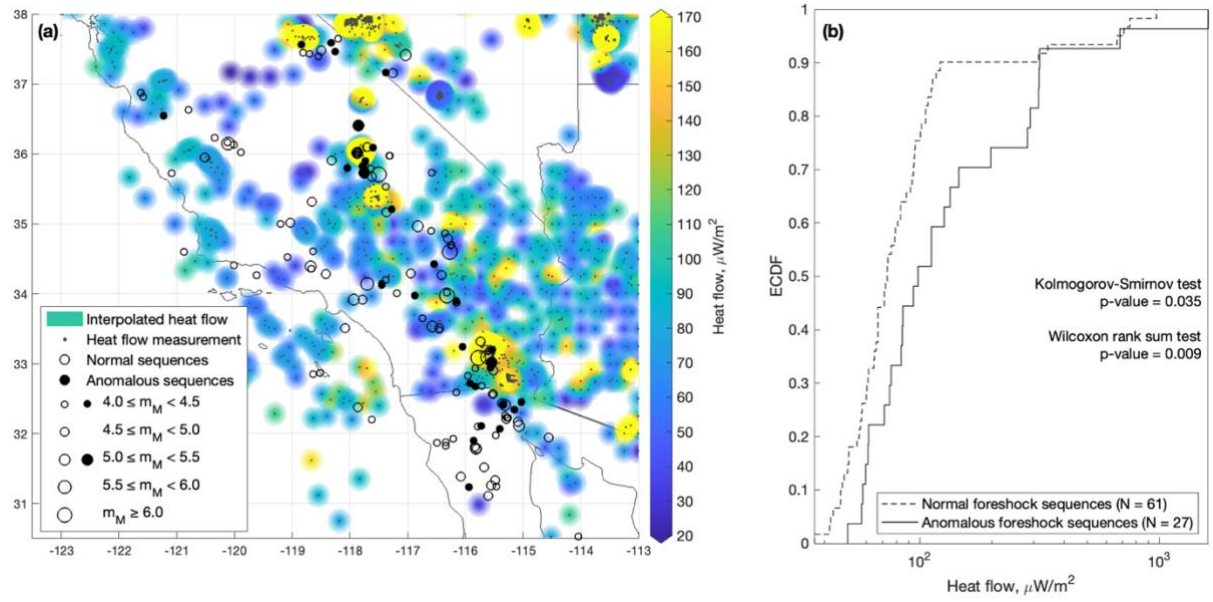
S11 Like Figure S9 (TEST2, STW method) but using the second ETAS model. Results are based on 1000 synthetic catalogs.



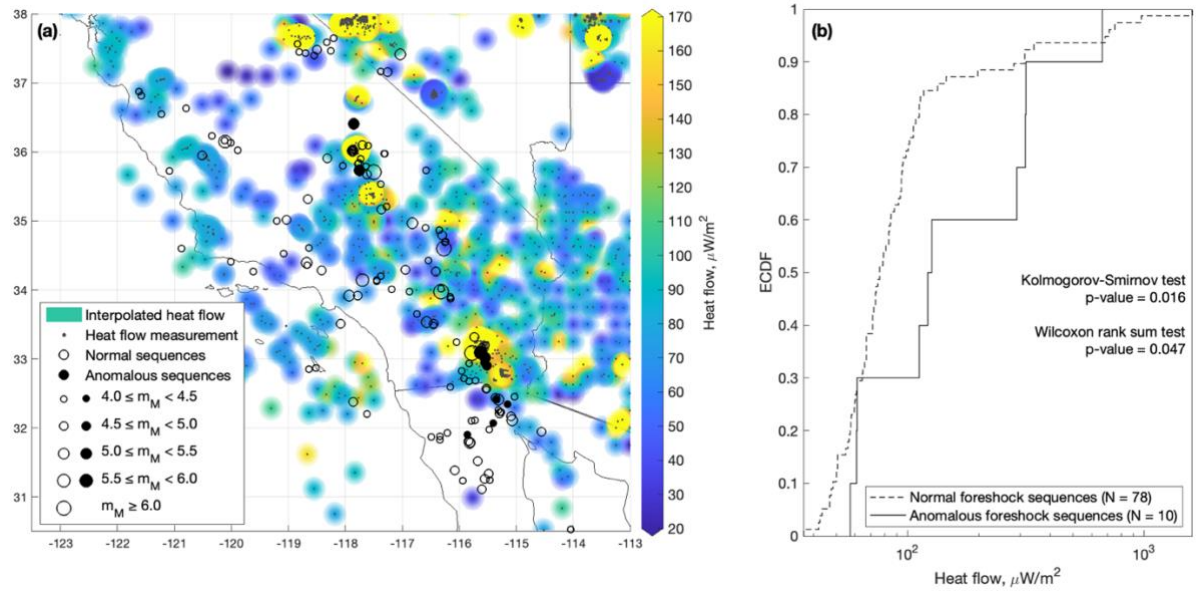
S12 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using a significance level of 95%.



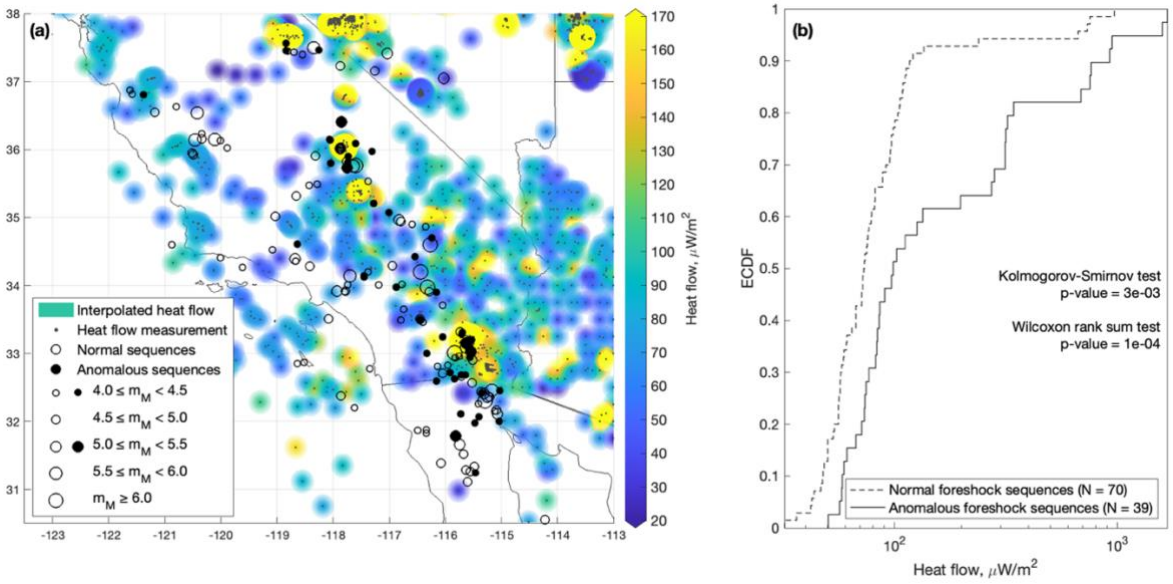
S13 Like Figure 2.2 in Chapter 2 (TEST2, NN method) but using a significance level of 95%.



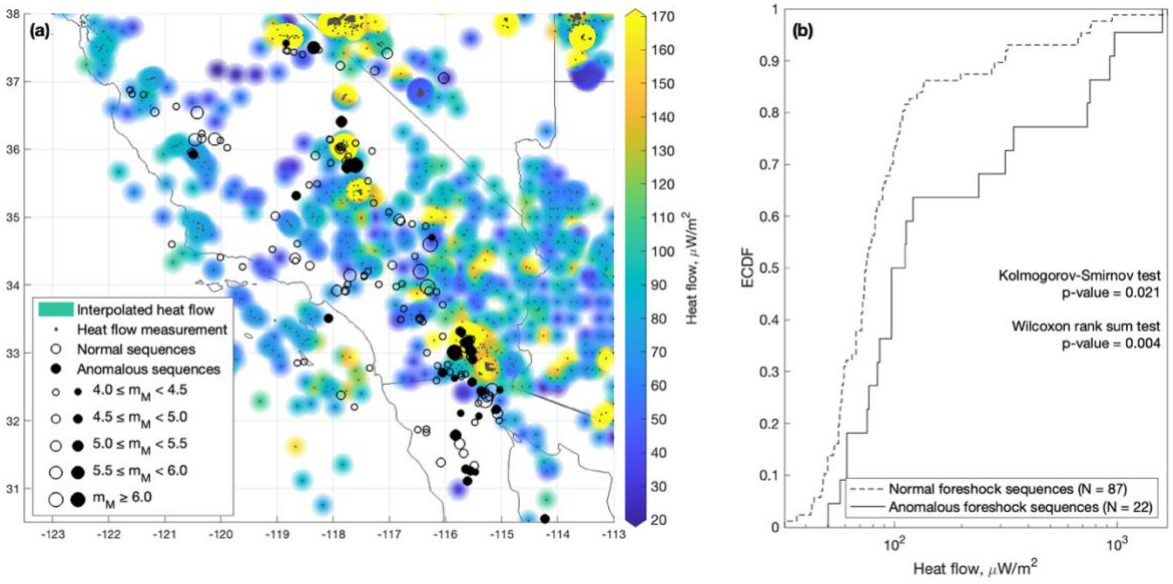
S14 Like Figure 2.3 in Chapter 2 (TEST1) but using the STW method.



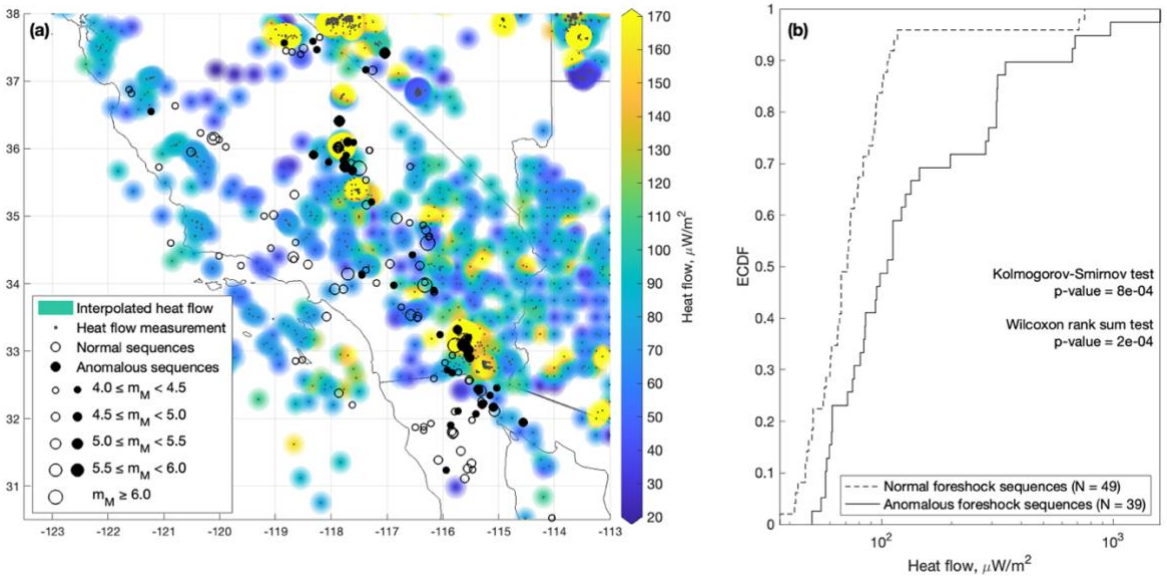
S15 Like Figure 2.4 in Chapter 2 (TEST2) but using the STW method.



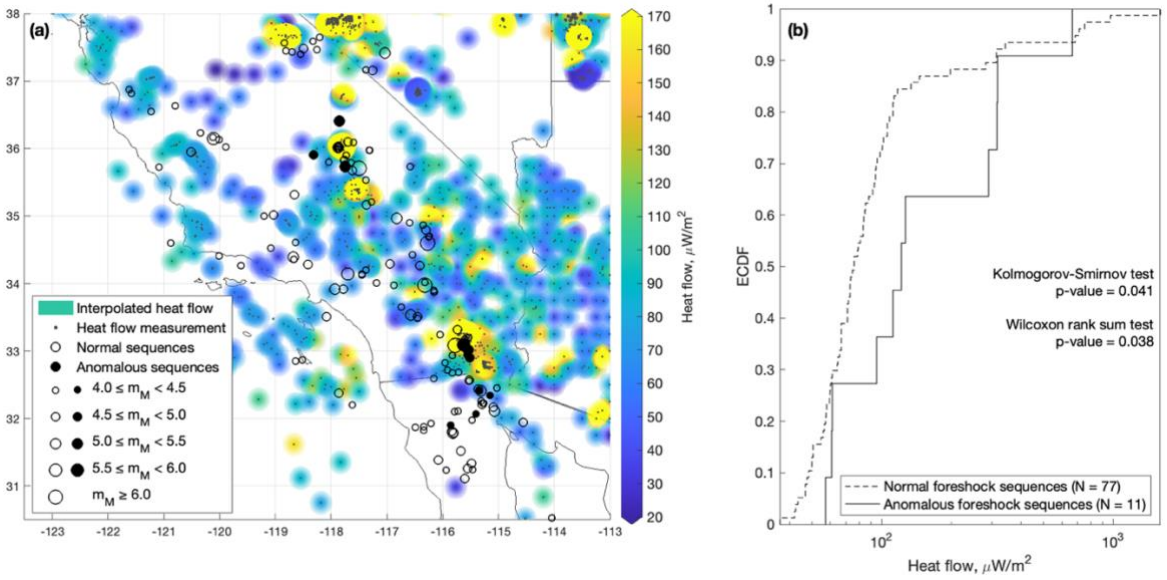
S16 Like Figure 2.3 in Chapter 2 (TEST1, NN method) but using the second ETAS model.



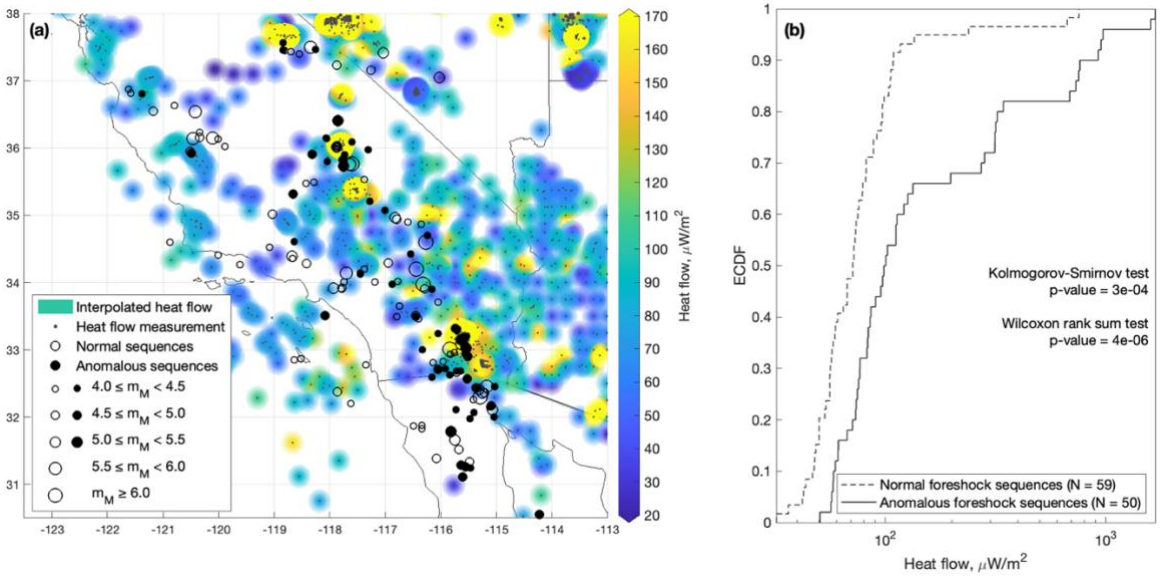
S17 Like Figure 2.4 in Chapter 2 (TEST2, NN method) but using the second ETAS model.



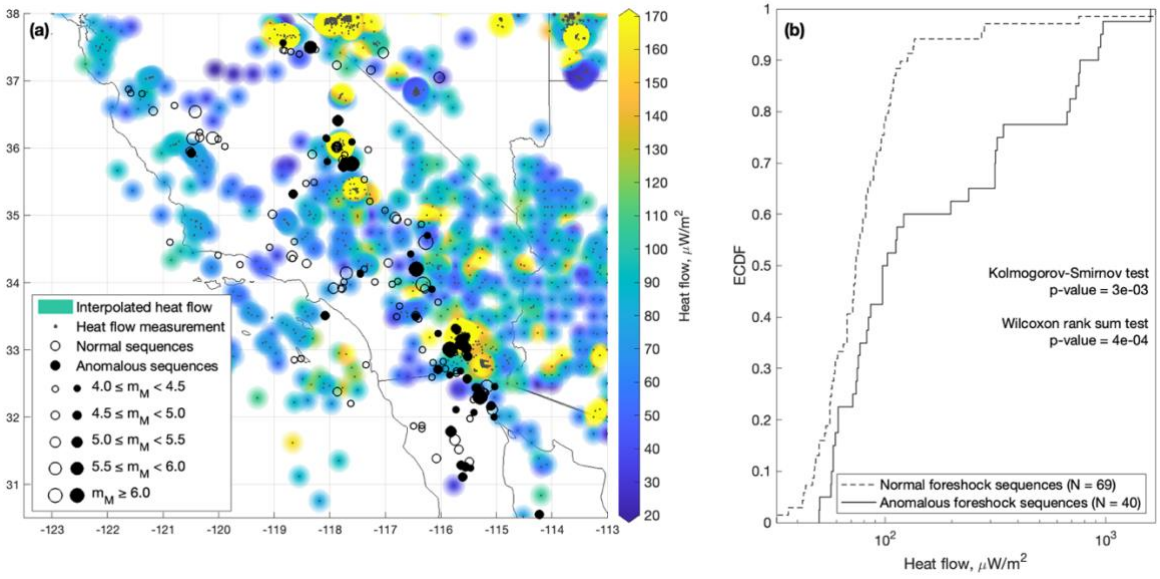
S18 Like Figure S14 (TEST1, STW method) but using the second ETAS model.



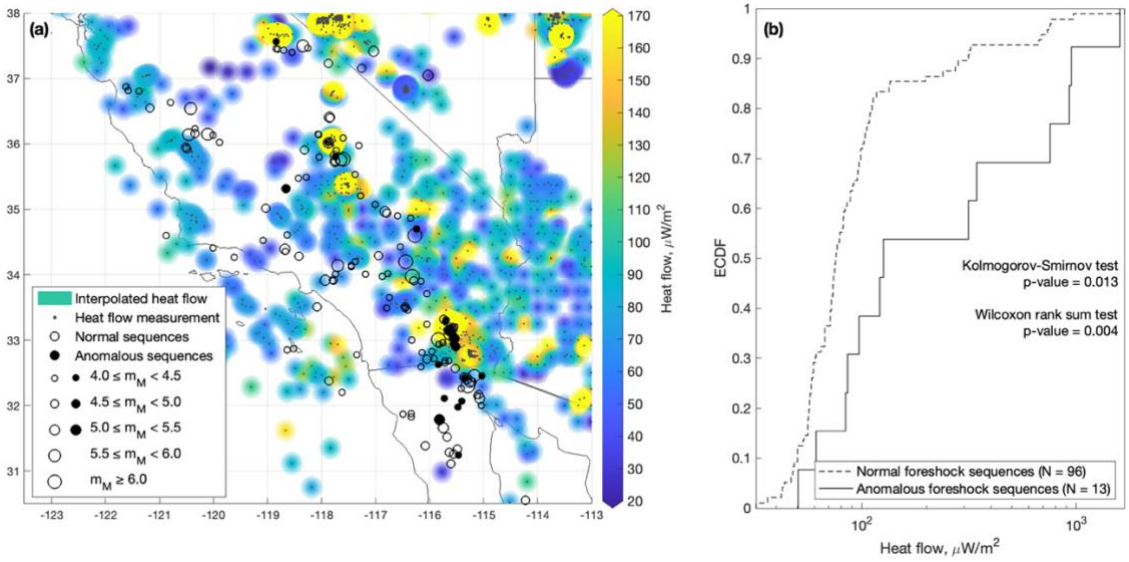
S19 Like Figure S15 (TEST2, STW method) but using the second ETAS model.



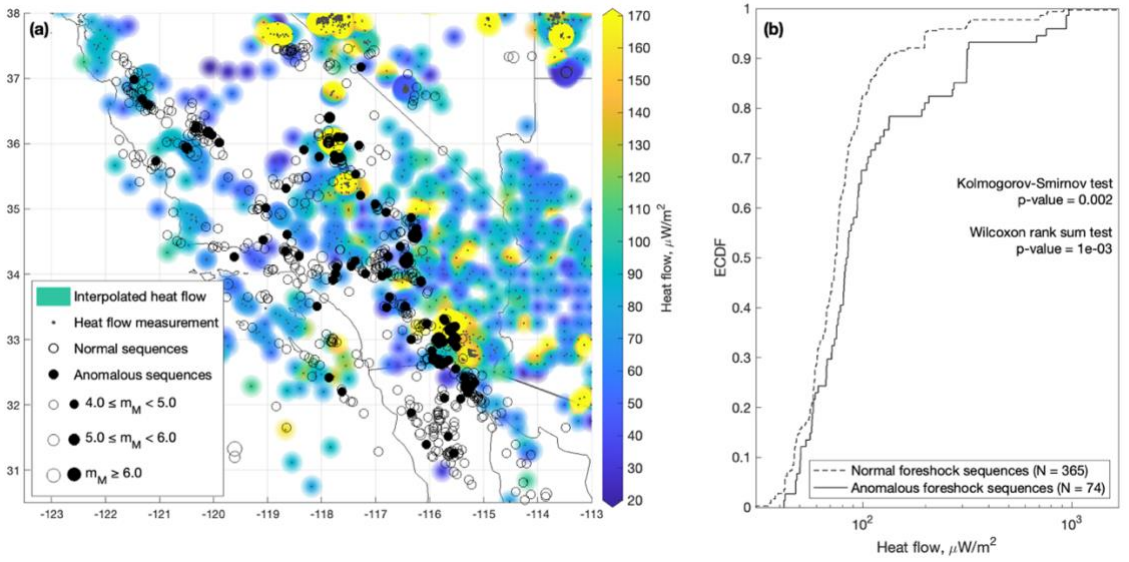
S20 Like Figure 2.3 in Chapter 2 (TEST1, NN method) but using a significance level of 95%.



S21 Like Figure 2.4 in Chapter 2 (TEST2, NN method) but using a significance level of 95%.



S22 Like Figure 2.3 in Chapter 2 (TEST1, NN method, 99th percentile) but with anomalous sequences identified using an alternative analysis that uses the distributions of the individual number of foreshocks (not the average, see Figure S6).



S23 Like Figure 2.3 in Chapter 2 but with the locations of normal and anomalous foreshock sequences identified by Petrillo & Lippiello (2021).

SI.2

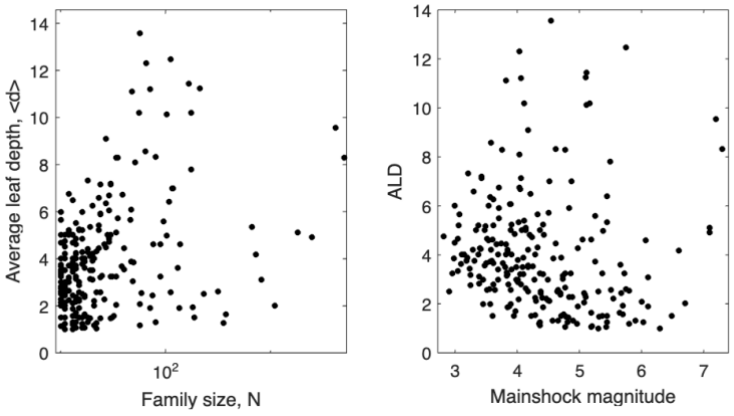
Supporting Information for Chapter 4

Average Leaf Depth

For each seismic sequence, the average leaf depth (ALD) is the sum of the generation level of each leaf (i.e., last event of each branch), $\sum d_g$, divided by the total number of leaves n_g (for more details see Chapter 3, Method section):

$$ALD = \frac{\sum d_g}{n_g}$$

In Figure S2.1, we plot the computed ALD of each sequence vs family size (i.e., number of events of each family) and the ALD vs mainshock magnitude (M_m) of the sequences. As in Zaliapin and Ben Zion (2013), we can recognize: i) sequences with high ALD, low family size and low mainshock magnitude ($M_m < 6$) reflecting the swarm-like behavior, ii) sequences with low ALD, higher mainshock magnitude ($M_m \geq 5$ & $M_m < 7$) and higher family size corresponding to aftershock sequences, and iii) sequences with a large mainshock size ($M_m \geq 7$) that consist likely of combined of previous types.



S2.1 (On the left) ALD vs number of events for each family; (on the right) ALD vs mainshock magnitude of each family

Relative timing of the largest event

The relative timing is the definition of Chen and Shearer (2016) used to characterize cluster types:

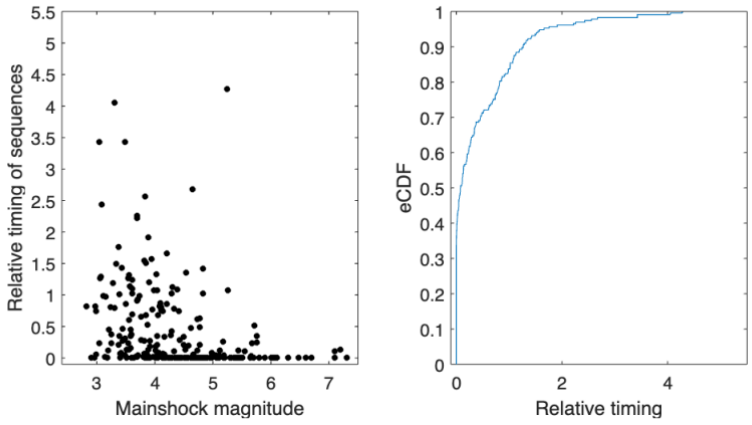
$$t_{rel} = \frac{(T_m - T_0)}{mean(T_i - T_0, i = 1 \dots N)}$$

T_m is the time of the mainshock (largest event), T_0 is the time of the first event of the sequence and T_i is the time of other events in the sequence ($i = 1 \dots N$).

Noteworthy, the relative timing t_{rel} is only computed for foreshock sequences. If there are no foreshocks for some sequences, relative timing is equal to the minimum value (larger than 0) rescaled of one unit.

We plot the relative timing t_{rel} vs magnitude of the largest event (Figure S2.2, left) and the eCDF of t_{rel} (Figure S2.2, right). Chen & Shearer (2016) define a sequence as swarm-type when $t_{rel} \geq 0.2$ (see Figure S2.5, from Chen &

Shearer, 2016), considered more prone to spatial migration. Only sequences whose largest magnitude is rather low ($M < 5.5$) have higher t_{rel}.



S2.2 (On the left) Relative timing vs mainshock magnitude of each sequence; (On the right) eCDF of relative timing values

Relative duration of foreshocks and aftershocks

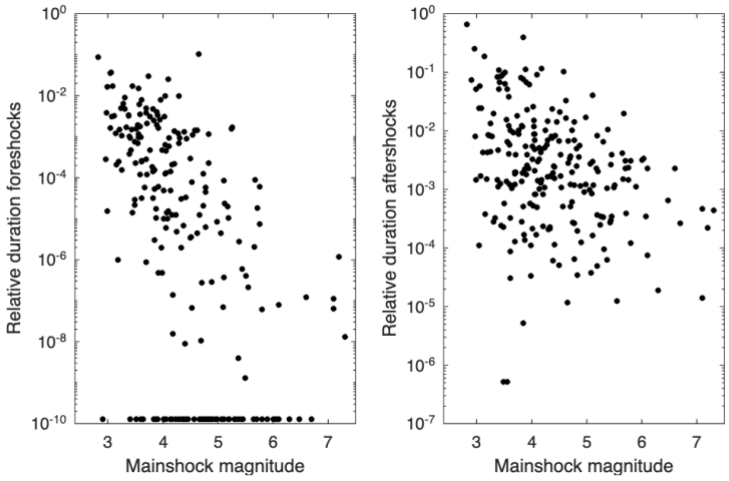
The relative duration of foreshocks (RD_{fors}) and aftershocks (RD_{aft}) respect to the time occurrence of the mainshock is a feature that can be used to make a strict division between clusters: traditional sequences tend to have low relative duration values (both for foreshocks and aftershocks) respect to swarm-type sequences (Zalipain et al., SCEC3 report):

$$RD_{fors} = \frac{(t_m - t_{first})}{10^{M_m}}$$

$$RD_{aft} = \frac{(t_{last} - t_m)}{10^{M_m}}$$

in which t_m is the time of the mainshock, t_{first} is the time of the first event, t_{last} is the time of the last event of the sequence. If there are no foreshocks for some sequences, relative duration is equal to the minimum value (larger than 0) rescaled of one unit.

Figure S2.3 show the relative duration of sequences vs their mainshock magnitude. It seems that both are inversely correlated, albeit with a large scatter. Especially moderate-sized sequences scatter the most.



S2.3 On the left: relative duration of foreshock sequences vs their mainshock magnitude; on the right: same plot but for aftershock sequences.

Area of foreshocks and aftershocks

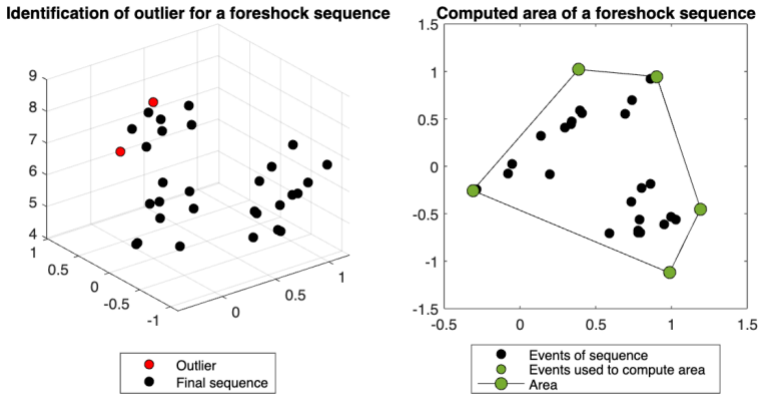
To characterized sequences in the space-domain, we compute the area of foreshocks and aftershocks for each sequence. To do that:

1) we convert the coordinate (Latitude and Longitude) of each event in *km* respect to the first event of the sequence;

2) we compute the 3D component Mahalanobis distance, and we use an inversion chi-square function to remove outlier (above 0.01 significance values) (Hadi,1992) for sequences with at least 4 foreshocks or aftershocks (example of a foreshock sequence in Figure S2.4 (left));

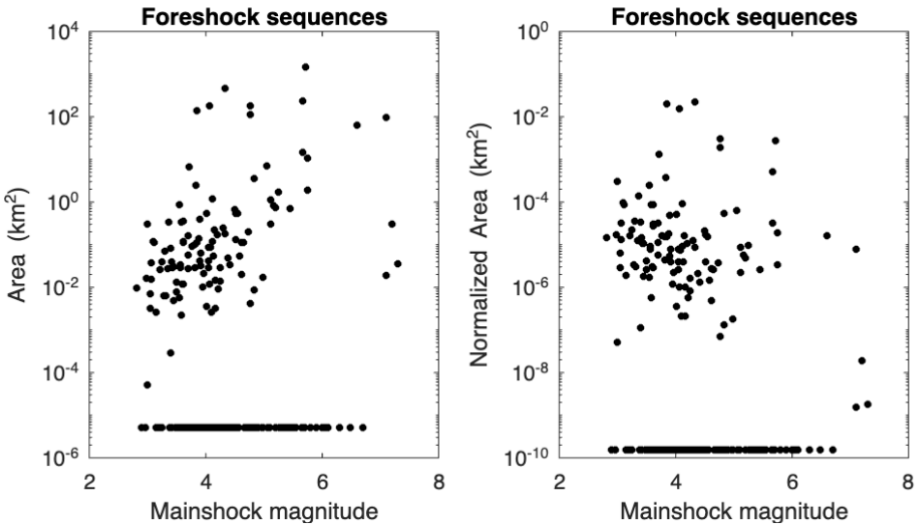
3) we compute the area of each sequence (considering x and y location of no-outlier events) using the convex hull function (example of a foreshock sequence in Figure S2.4 (right));

4) we plot the results in Figure S2.5 & 6 for foreshock and aftershocks sequences respectively, for no-normalized and normalized values respect to the magnitude of the mainshock sequences;

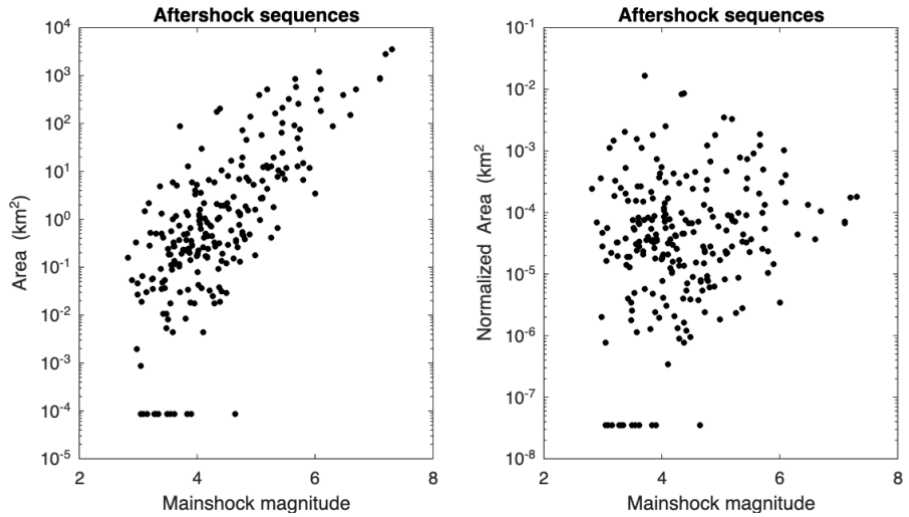


S2.4 (On the right) 3D location of foreshock sequence; red points are identified outlier events; (On the left) 2D location of foreshock sequence after removed outlier; green points are identified to compute the area of the foreshock sequence.

Normalized values are used for cluster analysis. Noteworthy, small mainshock magnitude sequences ($M_m \geq 4$ & $M_m < 7$) have foreshock area larger than higher mainshock magnitude sequences ($M_m > 7$) (Figure S2.5). In the case of aftershock sequences, we see an increasing trend (Figure S2.6).



S2.5 On the left: area of foreshocks sequences vs their mainshock magnitude; on the right: same plot but values are normalized by the mainshock magnitude.

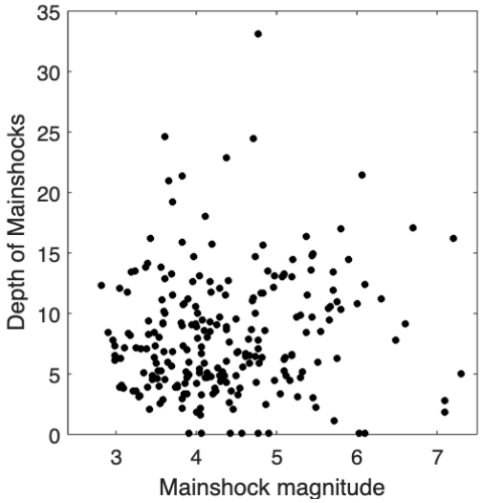


S2.6 On the left: area of aftershocks sequences vs their mainshock magnitude; on the right: same plot but values are normalized by the mainshock magnitude.

If there are no foreshocks or aftershocks for some sequences (or a number of events lees than 4), area values are equal to the minimum value (larger than 0) rescaled of one unit.

Depth of each sequence

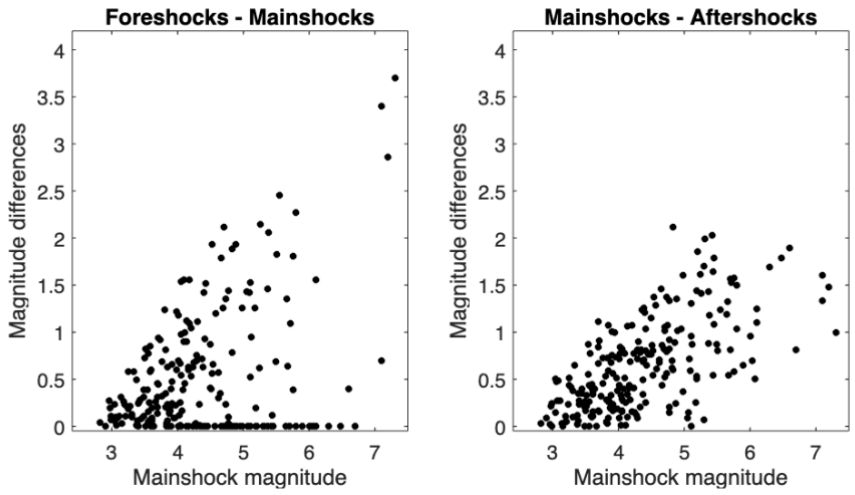
To fully investigate sequences in space-domain, we decide also to look at the depth of the mainshock of sequences. We compare them with the mainshock magnitude in Figure S2.7. All sequences, except one, are above 25 km, and we do not see a correlation. All events with a negative depth are replaced with the minimum positive depth rescaled of one unite.



S2.7 Depth of mainshock vs mainshock magnitude.

Magnitude differences (foreshocks-mainshocks & mainshocks-aftershocks)

Looking at the magnitude domain, we compute the difference between the magnitude of the mainshock and the largest foreshock and aftershock of each sequence. Then, we plot these values vs mainshock magnitude. From Figure S2.8 we can see an increasing trend: larger mainshocks have larger magnitude difference respect their foreshocks and aftershocks. If there are no foreshocks or aftershocks for some sequences, the magnitude difference is equal to the minimum value (larger than 0) rescaled of one unite.



S2.8 On the left: magnitude differences between the mainshock and the largest foreshock of each sequence vs mainshock magnitude; on the right: same plot but considering the largest aftershock of each sequence.

Normalized Productivity

The productivity of foreshocks (P_{fors}) and aftershocks (P_{aft}) is also a feature that can be used to make a strict division between clusters: we expect that traditional sequences should have large P_{aft} respect to P_{fors}, and we cannot assert that for swarm-type sequences (Zalipain et al., SCEC3 report).

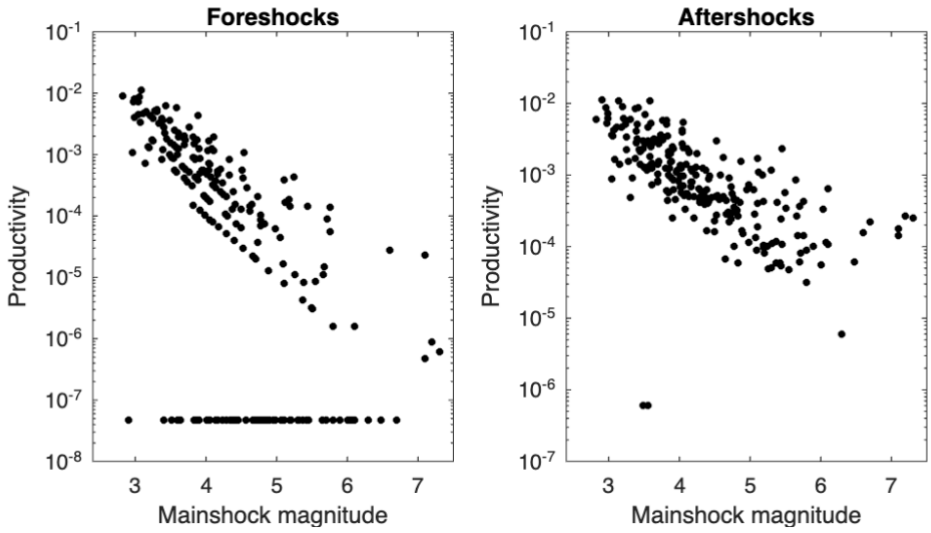
Here we compute the productivity using the following formulation:

$$P_{fors} = \frac{n_{fors}}{10^{Mm}};$$

$$P_{aft} = \frac{n_{aft}}{10^{Mm}};$$

where n_{fors} is the number of events before the mainshock, n_{aft} is the number of events after the mainshock, and M_m is the mainshock magnitude. We look at the computed results in Figure S2.9.

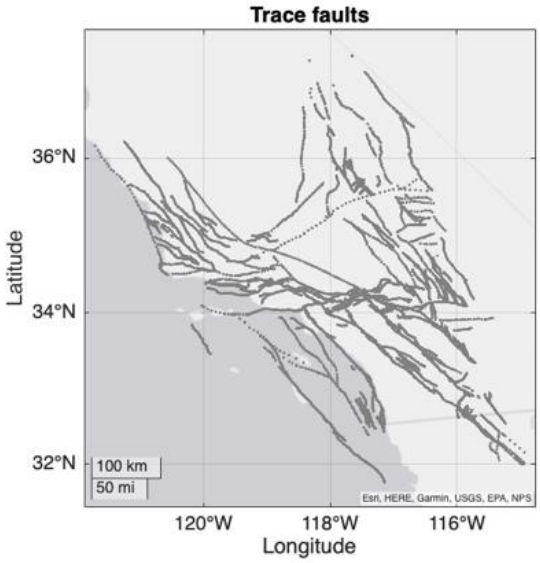
If there are no foreshocks or aftershocks, the productivity is equal to the minimum value (larger than 0) rescaled of one unite.



S2.9 On the left: productivity of foreshocks vs mainshock magnitude; on the right: same plot but considering aftershocks.

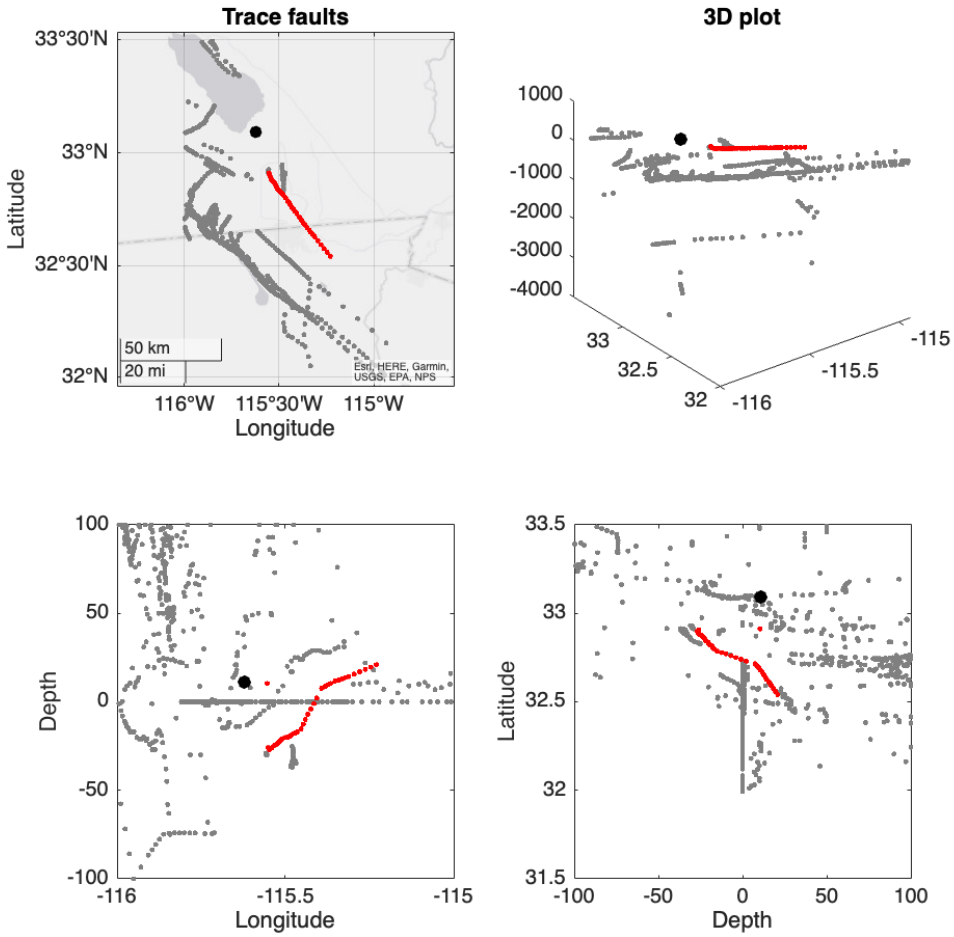
Nearest fault of each sequence

Associating sequences with their faults could be an interesting feature to characterize sequences in southern California. Different clusters of sequences could be characterized by specific fault types. To perform this kind of association, we use the dataset provided by SCEC Community Fault Model (CFM, dataset provided here: https://www.scec.org/research/cfm), which is a representation of active faults in southern California; Figure S2.10 shows the location of used fault traces.



S2.10 Fault locations from used SCEC datasets.

Figure S2.11 illustrates a method for associating the mainshock with his source fault in the Southern California (Plesch et al., 2020): we compute the 3D distance between each point of each fault and the mainshock location, considering the shortest one for the association. Figure S2.11 shows also a mainshock with its associated fault.



S2.11 Location of a mainshock (black point) and its associated fault (red points); gray points are the location of all other faults. Different subplots correspond to different visualizations of the same example.

List of Figures

Figure 1-1 Simple representation of self-exciting point processes 1-7

Figure 2-1 Results of TEST1 for various classes of the mainshock magnitude mM (rows) and thresholds for the foreshock magnitude mF (columns). Each subplot displays the number of normalized foreshocks NF for the real catalog (vertical line; red if anomalous, black otherwise) and the empirical Cumulative Distribution Function (eCDF, dashed curve) with its 99th percentile (dashed vertical line) for 10'000 synthetic catalogs. Each subplot also reports the number of anomalous foreshock sequences, $NAFS$, the p-value for TEST1, and the number of mainshocks, NM. The results are based on K. Felzer's ETAS model and the NN method; Figure S5 shows results using the STW method, and Figure S7 and S8 using the second ETAS model. Note that each subplot uses a different NF-axis range to account for the varying data range. 2-24

Figure 2-2 Results of TEST2 showing probability mass functions (PMFs) of the number of foreshocks NF for various classes of mM (rows) and mF (columns). The PMFs are shown for (i) the real catalog (triangles), (ii) all synthetic catalogs (small gray dots as swarm distributions) with their 99th percentile (gray horizontal bars), and (iii) when considering all synthetic catalogs as a single compound catalog (blue open circles, using the approach of Seif et al., 2019). Triangles become red when they are located above the 99th percentile of (ii). The results are based on K. Felzer's ETAS model and the NN method; Figure S9 shows results using the STW method, and Figures S10 and S11 using the second ETAS model. Note that each subplot uses a different NF-axis range.. 2-25

Figure 2-3 Correlating foreshock sequences with the heat flow. (a) Locations of normal (empty circles) and anomalous foreshock sequences (filled circles)

identified with TEST1 overlayed on a heat flow map. The circles sizes scales with m_M (see legend). The interpolated heat flow map is based on sampled heat flow measurements (small gray dots, see Data and Methods section); (b) eCDFs of heat flow values at locations of normal (dashed curve) and anomalous foreshock sequences (solid curve); both eCDFs are compared using two statistical tests (see annotation with corresponding p-values). The results are based on K. Felzer's ETAS model and the NN method; Figure S14 shows results using the STW method, and Figures S16 and S18 using the second ETAS model. 2-27

Figure 2-4 Like Figure 2.3 but with foreshock sequences labeled as 'anomalous' or 'normal' using TEST2. Figure S9 shows results using the STW method, and Figures S17 and S19 using the second ETAS model. 2-27

Figure 3-1 Average leaf depth, ALD, of two different types of simplistic sequences; dg is the sum of events over each leaf, and ng is the number of leaves; ALD is the ratio of dg and ng (figure inspired by the work of Zaliapin & Ben Zion, 2013)..... 3-35

Figure 3-2 Average leaf depth, ALD, vs family size, N, for the real catalog (a) and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). Dots are colored according to the density of the data. Higher density areas have more overlapped data. 3-36

Figure 3-3 Average leaf depth, ALD, vs mainshock magnitude, M_M, for the real catalog (a) and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). Dots are colored according to the density of the data. Vertical line highlights the threshold $MM = 6.5$ used for Felzer planar source model. Higher density areas have more overlapped data. 3-37

Figure 3-4 Distribution of ALD in terms of histograms for (a) the real catalog and 1000 synthetic catalogs reproduced using Felzer model with point source (b), Felzer model with planar source (c), and Seif model with point source (d). The vertical dotted lines represent $ALD = 5$, the maximum ALD found by Zaliapin and Ben Zion (2013) using one synthetic ETAS catalog; curly brackets represent the fraction of families with $ALD > 5$;..... 3-38

Figure 4-1 Simple explanation of the proposed method: 1) Collect information on seismic sequences as explained in section 4.2.1; 2) perform a k-means cluster analysis on collected dataset and identify the optimal number of clusters to split sequences, as explained in section 4.2.2; 3) look at statistical properties of each identified cluster, as explained in section 4.2.3; 4) perform the leave one out method to identify the most important features from the collected dataset, as explained in section 4.2.4. 4-51

Figure 4-2 Histogram of normalized feature values. 4-53

Figure 4-3(On the left) Silhouette plots for each k-value = [1:10]: x-axis corresponds to the silhouette value of each element; y-axis corresponds to the number of identified clusters. (On the right) Corresponding stem plot for each silhouette plot: x-axis represents the number of clusters and y-axis represents the number of sequences inside each cluster. 4-54

Figure 4-4 Different methods used to identify the optimal number of clusters highlight by the vertical dotted lines. 4-55

Figure 4-5 On the left, interpolated heat flow map with sampled points and mainshock location of sequences of each cluster; on the right eCDF of interpolated heat flow values of each mainshock of each cluster; eCDF are compared using two statistical tests (see annotation with corresponding p-values). 4-56

Figure 4-6 a) Magnitude Frequency Distribution of mainshocks for each identified cluster; b) Magnitude Frequency Distribution of all events for each

identified cluster; b-value is also computed and reported for each cluster with the uncertainty. MFDs are compared using two statistical tests (see annotation with corresponding p-values). 4-57

Figure 4-7 Number of foreshock and aftershock for each mainshock magnitude (dots); mean values of number of foreshocks and aftershocks (squares) for each bin of magnitude (large 0.1); solid lines are regression fit of each cluster; we also report the slope values (s) and the intercept (i) of each line. 4-58

Figure 4-8 Rate for aftershock and foreshock sequences: x-axis is the day of events before or after the mainshocks; y-axis are the normalized time of foreshocks or aftershocks for specific time interval. Different function corresponds to different identified clusters 4-58

Figure 4-9 Survival function of rate of foreshocks (left) and aftershocks (right) for different identified clusters. 4-59

Figure 4-10 Same plots as Figure 4.3 but using the most important features for cluster analysis 4-64

Figure 4-11 Same plots as Figure 4.4 but using the most important features for cluster analysis. 4-64

Figure 4-12 eCDFs of most important features: a) relative duration of foreshocks, b) area of foreshocks, c) foreshock productivity and d) the magnitude of the largest event for each identified cluster 4-65

Figure 4-13 As Figure 4.5 but using the most important features for cluster analysis. 4-66

Figure 4-14 As Figure 4.6 but using the most important features for cluster analysis. 4-67

Figure 4-15 As Figure 4.7 but using the most important features for cluster analysis. 4-67

Figure 4-16 As Figure 4.8 but using the most important features for cluster analysis. 4-68

Figure 4-17 As Figure 4.9 but using the most important features for cluster analysis. 4-68

S1 Total number of events in the synthetic catalogs (distribution) and the real catalog (solid vertical line), using events above M2.5. The dashed vertical lines refer to the 95% confidence interval (i.e., the 2.5^{th} – 97.5^{th} percentile range of the distribution). For more information, see Text S2. 5-82

S2 Like Figure S1 but for the second ETAS model (by S. Seif). 5-82

S3 Like Figure S1 but for different magnitude classes. 5-83

S4 Like Figure S3 but for the second ETAS model (by S. Seif). 5-83

S5 Like Figure 2.1 in Chapter 2 (TEST1) but using the STW method. 5-84

S6 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using the individual number of foreshocks, N_F (alternative analysis of TEST1). In this way, the empirical Cumulative Distribution Function (eCDF) of N_F can be constructed for both the real catalog (solid curve) and all 10'000 synthetic catalogs combined (dashed curve); vertical lines show their corresponding 99^{th} percentile (real catalog: solid; synthetic catalogs: dashed). If the former is above the latter, the solid vertical line becomes red, indicating more anomalous foreshock sequences than expected. Each subplot also reports the number of anomalous foreshock sequences, N_{AFS}, and the number of mainshocks, N_M. 5-85

S7 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using the second ETAS model. Results are based on 1000 synthetic catalogs. 5-86

S8 Like Figure S5 (TEST1, STW method) but using the second ETAS model. Results are based on 1000 synthetic catalogs. 5-87

S9 Like Figure 2.2 in Chapter 2 (TEST2) but using the STW method. 5-88

S10 Like Figure 2.2 in Chapter 2 (TEST2, NN method) but using the second ETAS model. Results are based on 1000 synthetic catalogs. 5-89

S11 Like Figure S9 (TEST2, STW method) but using the second ETAS model. Results are based on 1000 synthetic catalogs. 5-90

S12 Like Figure 2.1 in Chapter 2 (TEST1, NN method) but using a significance level of 95%. 5-91

S13 Like Figure 2.2 in Chapter 2 (TEST2, NN method) but using a significance level of 95%. 5-92

S14 Like Figure 2.3 in Chapter 2 (TEST1) but using the STW method. 5-93

S15 Like Figure 2.4 in Chapter 2 (TEST2) but using the STW method. 5-93

S16 Like Figure 2.3 in Chapter 2 (TEST1, NN method) but using the second ETAS model. 5-94

S17 Like Figure 2.4 in Chapter 2 (TEST2, NN method) but using the second ETAS model. 5-94

S18 Like Figure S14 (TEST1, STW method) but using the second ETAS model. 5-95

S19 Like Figure S15 (TEST2, STW method) but using the second ETAS model. 5-95

S20 Like Figure 2.3 in Chapter 2 (TEST1, NN method) but using a significance level of 95%. 5-96

S21 Like Figure 2.4 in Chapter 2 (TEST2, NN method) but using a significance level of 95%. 5-96

S22 Like Figure 2.3 in Chapter 2 (TEST1, NN method, 99th percentile) but with anomalous sequences identified using an alternative analysis that uses the distributions of the individual number of foreshocks (not the average, see Figure S6). 5-97

S23 Like Figure 2.3 in Chapter 2 but with the locations of normal and anomalous foreshock sequences identified by Petrillo & Lippiello (2021). 5-97

S2.1 (On the left) ALD vs number of events for each family; (on the right) ALD vs mainshock magnitude of each family 5-99

S2.2 (On the left) Relative timing vs mainshock magnitude of each sequence; (On the right) eCDF of relative timing values 5-100

S2.3 On the left: relative duration of foreshock sequences vs their mainshock magnitude; on the right: same plot but for aftershock sequences. 5-101

S2.4 (On the right) 3D location of foreshock sequence; red points are identified outlier events; (On the left) 2D location of foreshock sequence after removed outlier; green points are identified to compute the area of the foreshock sequence. 5-102

S2.5 On the left: area of foreshocks sequences vs their mainshock magnitude; on the right: same plot but values are normalized by the mainshock magnitude. 5-103

S2.6 On the left: area of aftershocks sequences vs their mainshock magnitude; on the right: same plot but values are normalized by the mainshock magnitude. 5-103

S2.7 Depth of mainshock vs mainshock magnitude..... 5-104

S2.8 On the left: magnitude differences between the mainshock and the largest foreshock of each sequence vs mainshock magnitude; on the right: same plot but considering the largest aftershock of each sequence..... 5-105

S2.9 On the left: productivity of foreshocks vs mainshock magnitude; on the right: same plot but considering aftershocks. 5-106

S2.10 Fault locations from used SCEC datasets. 5-107

S2.11 Location of a mainshock (black point) and its associated fault (red points); gray points are the location of all other faults. Different subplots correspond to different visualizations of the same example. 5-108

List of Tables

Table 4-1 Leave one out procedure to select the needful subset: for each selected subset (column 1), we report the removed feature from the previous subset (column 2), the eventually reinserted feature (column 3), the optimal number of clusters for the current subset (column 4), the number of misclassified sequences respect to cluster $C1i$ (column 5) & $C2i$ (column 6). 4-60

Table S1 Parameters used in K. Felzer's ETAS simulator..... 5-79

Table S2 Used ETAS parameters as given by Hardebeck et al., 2008 for M_0 = 2.5. μ is the background rate; K, c, and p are parameters of Omori's law; n is the aftershock decay with distance (exponent in r^{-n}); α is the productivity law exponent (productivity scaling with magnitude); and bβ is the Gutenberg–Richter b-value multiplied by ln(10) and describes the magnitude distribution. 5-79

Table S3 Used ETAS parameters as given by Seif et al., 2017 for M_0 = 2.5 and fixed α. For a description of K, α, β, p, and c see Table S2. q, d, and γ describe the spatial distribution of earthquakes. 5-80

Table S4 p-values for comparing eCDFs of heat flow values at locations of normal versus anomalous foreshock sequences for each analysis. 5-80

Bibliography

Agnew, D. C., & Jones, L. M. (1991). Prediction probabilities from foreshocks. *Journal of Geophysical Research, 96*(B7), 11959. https://doi.org/10.1029/91JB00191

Baiesi, M., & Paczuski M. (2004). Scale-free networks of earthquakes and aftershocks. *Physical Review E*, *69*, 066106. https://doi.org/10.1103/PhysRevE.69.066106.

Bayliss, K., Naylor, M., & Main, I. G. (2019). Probabilistic identification of earthquake clusters using rescaled nearest neighbour distance networks. *Geophysical Journal International*, 217, 487–503. https://doi.org/10.1093/gji/ggz034

Bird, P., Jackson, D. D., Kagan, Y. Y., Kreemer, C., Stein, R. S. (2015). GEAR1: A Global Earthquake Activity Rate Model Constructed from Geodetic Strain Rates and Smoothed Seismicity. *Bulletin of the Seismological Society of America*, 105 (5): 2538–2554. https://doi.org/10.1785/0120150058

Blackwell, D. D., & Richards, M. (2004). Geothermal Map of North America, AAPG Map, scale 1:6,500,000, Product Code 423. url: https://www.smu.edu/-/media/Site/Dedman/Academics/Programs/Geothermal-Lab/Graphics/Geothermal_MapNA_7x10in.gif

Bouchon, M., Karabulut, H., Aktar, M., Özalaybey, S., Schmittbuhl, J., & Bouin, M. P. (2011). Extended Nucleation of the 1999 Mw 7.6 Izmit Earthquake. *Science*, *331*(6019), 877–880. http://doi.org/10.1126/science.1197341

Cattania, C., & Segall, P. (2021). Precursory slow slip and foreshocks on rough faults. *Journal of Geophysical Research: Solid Earth*, 126(4), e2020JB020430. https://doi.org/10.1029/2020JB020430

Chen, X. & Shearer, P. M. (2016). Analysis of Foreshock Sequences in California and Implications for Earthquake Triggering. *Pure Applied Geophysical*, *173*(1), 133–152. https://doi.org/10.1007/s00024-015-1103-0

Chen, X., Shearer, P. M., Abercrombie R.E. (2012). Spatial migration of earthquakes within seismic clusters in Southern California: Evidence for fluid diffusion. *Journal of Geophysical Research,* Vol. 117, B04301. https://doi.org/10.1029/2011JB008973

Chiaraluce, L., Chiarabba, C., De Gori, P., Di Stefano, R., Improta, L., Piccinini, D., Schlagenhauf, A., Traversa, P., Valoroso, L., & Voisin, C. (2011). The 2009 L'Aquila (central Italy) seismic sequence. *Bollettino di Geofisica Teorica ed Applicata*, *52*(3), 367-387. https://doi.org/10.4430/bgta0019

Davis, J., C. (2002). Statistics and Data Analysis in Geology*, Wiley*.

Ellsworth, W. L., & Beroza, G. C. (1995). Seismic Evidence for an Earthquake Nucleation Phase. *Science*, *268*, 851-855. http://doi.org/10.1126/science.268.5212.851

Ellsworth, W. L., & Bulut, F. (2018). Nucleation of the 1999 Izmit earthquake by a triggered cascade of foreshocks. *Nature Geoscience*, *11*(7), 531–535. https://doi.org/10.1038/s41561-018-0145-1

Enescu, B., Hainzl, S., and Ben-Zion, Y. (2009). Correlations of Seismicity Patterns in Southern California with Surface Heat Flow Data. *Bulletin of the Seismological Society of America*, *99* (6), 3114-3123. https://doi.org/10.1785/0120080038

Farrell, J., Husen, S., Smith, R.B. (2009). Earthquake swarm and b-value characterization of the Yellowstone volcano-tectonic system, *Journal of Volcanology and Geothermal Research*,188, 1–3,260-276. https://doi.org/10.1016/j.jvolgeores.2009.08.008

Farrelly, C. (2021). "Positive Biology" and Well-Ordered Science. Measuring well-being: Interdisciplinary perspectives from the social sciences and the humanities. *New York: Oxford*.

Felzer, K. R., Becker, T. W., Abercrombie, R. E., Ekström, G., & Rice, J. R. (2002). Triggering of the 1999 M_w 7.1 Hector Mine earthquake by aftershocks of the 1992 M_w 7.3 Landers earthquake. *Journal of Geophysical Research*, *107*(B9), 2190. https://doi.org/10.1029/2001JB000911

Gomberg, J. (2018). Unsettled earthquake nucleation. *Nature Geoscience*, 11(7), 463–464. https://doi.org/10.1038/s41561-018-0149-x

Greco, J. (2010). Achieving knowledge: A virtue-theoretic account of epistemic normativity. *Cambridge University Press*.

Gulia, L., & Wiemer, S. (2019). Real-time discrimination of earthquake foreshocks and aftershocks. *Nature*, *574*(7777), 193-199. https://doi.org/10.1038/s41586-019-1606-4

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(3), 761-771. https://doi.org/10.1111/j.2517-6161.1992.tb01449.x.

Hainzl, S., & Ogata, Y. (2005). Detecting fluid signals in seismicity data through statistical earthquake modeling. *Journal of Geophysical Research: Solid Earth*, 110(B5). https://doi.org/10.1029/2004JB003247

Hardebeck, J. L., Felzer, K. R., & Michael, A. J. (2008). Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant.

Journal of Geophysical Research, *113*, 3B08310. https://doi.org/10.1029/2007JB005410

Harte, D. S. (2015). Model parameter estimation bias induced by earthquake magnitude cut-off. *Geophysical Journal International*, 204(2), 1266–1287. https://doi.org/10.1093/gji/ggv524

Hauksson, E., Yang, W., & Shearer, P. M. (2012). Waveform relocated earthquake catalog for Southern California (1981 to June 2019). *Bulletin of the Seismological Society of America*, *102*(5), 2239–2244. https://doi.org/10.1785/0120120010

Jesseph, D. (2010). Scientia in Hobbes. In: Sorell, T., Rogers, G., Kraye, J. (eds) Scientia in Early Modern Philosophy. *Studies in History and Philosophy of Science*, 24. https://doi.org/10.1007/978-90-481-3077-1_8

Jordan, T. H., Chen, Y. T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., et al. (2011). Operational earthquake forecasting: State of knowledge and guidelines for utilization. *Annals of Geophysics*, 54(4), 319–391. https://doi.org/10.4401/ag-5350

Kato, A. & Ben-Zion, Y. (2021). The generation of large earthquakes. *Nature Reviews Earth & Environment*, 2(1), pp.26-39. https://doi.org/10.1038/s43017-020-00108-w

Kato, A., Iidaka, T., Ikuta, R., Yoshida, Y., Katsumata, K., Iwasaki, T., Sakai, S., Thurber, C., Tsumura, N., Yamaoka, K., Watanabe, T (Page M. T.)., Kunitomo, T., Yamazaki, F., Okubo, M., Suzuki, S., & Hirata, N. (2010). Variations of fluid pressure within the subducting oceanic crust and slow earthquakes. *Geophysical Research Letters*, *37* (14). https://doi.org/10.1029/2010GL043723

Kato, A., Obara, K., Igarashi, T., Tsuruoka, H., Nakagawa, S., & Hirata, N. (2012). Propagation of Slow Slip Leading Up to the 2011 Mw 9.0 Tohoku-Oki Earthquake. *Science*, *335*(6069), 705-708. https://doi.org/10.1126/science.1215141

Kaufman, L., and P. J. Rouseeuw. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *Hoboken, NJ: John Wiley & Sons, Inc.* https://doi.org/10.1002/9780470316801

Kumazawa, T., Ogata, Y., Kimura, K., Maeda, K., & Kobayashi, A. (2016). Background rates of swarm earthquakes that are synchronized with volumetric strain changes. *Earth and Planetary Science Letters,* 442, 51-60. http://dx.doi.org/10.1016/j.epsl.2016.02.049

Li, Y., and Wu H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, 1104-1109, 1875-3892. https://doi.org/10.1016/j.phpro.2012.03.206

Lippiello, E., Giacco, F., Marzocchi, W., Godano, G., & Arcangelis, L. D. (2017). Statistical features of foreshocks in instrumental and ETAS catalogs. Pure and Applied Geophysics, 174, 1679-1697. https://doi.org/10.1007/s00024-017-1502-5

Llenos, A. L., McGuire, J. J., & Ogata, Y. (2009). Modeling seismic swarms triggered by aseismic transients. *Earth and Planetary Science Letters,* 281(1-2), 59-69. https://doi.org/10.1016/j.epsl.2009.02.011

Lohman, R. B., and J. J. McGuire (2007), Earthquake swarms driven by aseismic creep in the Salton Trough, California. *Journal of Geophysical Research*, 112, B04405, https://doi.org/10.1029/2006JB004596

Lombardi, A.M., Marzocchi W., Selva J. (2006). Exploring the evolution of a volcanic seismic swarm: the case of the 2000 Izu Islands swarm. *Geophysical Research Letters*, 33, L07310. http://doi.org/10.1029/2005GL025157

Mancini, S., Segou, M., Werner, M. J., & Cattania, C. (2019). Improving physics-based aftershock forecasts during the 2016–2017 Central Italy Earthquake Cascade. *Journal of Geophysical Research: Solid Earth*, *124*(8), 8626-8643. https://doi.org/10.1029/2019JB017874

Manganiello, E., Herrmann, M., & Marzocchi, W. (2023). New physical implications from revisiting foreshock activity in southern California. *Geophysical Research Letters*, 50, e2022GL098737. https://doi.org/10.1029/2022GL098737

Marzocchi, W., & Zhuang, J. (2011). Statistics between mainshocks and foreshocks in Italy and Southern California. *Geophysical Research Letters*, *38*(9), 2011GL047165. https://doi.org/10.1029/2011GL047165

Marzocchi, W., Spassiani, I., Stallone, A., & Taroni, M. (2020). How to be fooled searching for significant variations of the b-value. *Geophysical Journal International* 220.3 (2020): 1845-1856. https://doi.org/10.1093/gji/ggz541

McLaskey, G.C. (2019). Earthquake initiation from laboratory observations and implications for foreshocks. *Journal of Geophysical Research*: Solid Earth, 124(12), pp.12882-12904 https://doi.org/10.1029/2019JB018363

Meng, H., & Fan, W. (2021). Immediate foreshocks indicating cascading rupture developments for 527 M 0.9 to 5.4 Ridgecrest earthquakes. *Geophysical Research Letters*, *48*, e2021GL095704. https://doi.org/10.1029/2021GL095704

Mogi, K. (1963). Some discussions on aftershocks, foreshocks and earthquake swarms the fracture of a semi-infinite body caused by an inner stress origin and

its relation to the earthquake phenomena, 3. *Bull. Earthq. Res. Inst. Univ. Tokyo*. 41, 615–658. https://doi.org/10.15083/0000033716

Molchan, G. M., & Dmitrieva, O. E. (1992). Aftershock identification: methods and new approaches. *Geophysical Journal International*, *109*(3), 501–516. https://doi.org/10.1111/j.1365-246X.1992.tb00113.x

Moutote, L., Marsan, D., Lengliné, O., & Duputel, Z. (2021). Rare occurrences of non-cascading foreshock activity in Southern California. *Geophysical Research Letters*, *48*, e2020GL091757. https://doi.org/10.1029/2020GL091757

Nandan, S., Ouillon, G., Wiemer, S., and Sornette, D. (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California, *Geophysical Research Letters Solid Earth*, *122*, 5118–5143. https://doi.org/10.1002/2016JB013266

Nanjo, K. Z., Tsuruoka, H., Yokoi, S., Ogata Y., Falcone, G., Hirata, N., Ishigaki, Y., et al. (2012). Predictability study on the aftershock sequence following the 2011 Tohoku-Oki, Japan, earthquake: first results, *Geophysical Journal International*, Vol 191, Issue 2, 653–658. https://doi.org/10.1111/j.1365-246X.2012.05626.x

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83 (401), 9-27. https://doi.org/10.1080/01621459.1988.10478560

Page, M. T. and van der Elst, N. J. (2022). Aftershocks Preferentially Occur in Previously Active Areas, *The Seismic Record*. 2(2), 100–106, https://doi.org/10.1785/0320220005.

Page, M. T., & van der Elst, N. J. (2018). Turing-Style Tests for UCERF3 Synthetic Catalogs, *Bulletin of the Seismological Society of America*, *108*(2), 729–741. https://doi.org/10.1785/0120170223

Petrillo, G., & Lippiello, E. (2021). Testing of the foreshock hypothesis within an epidemic like description of seismicity. *Geophysical Journal International*, 225, 1236–1257. https://doi.org/10.1093/gji/ggaa611

Petrillo, G., & Zhuang, J. (2022). The debate on the earthquake magnitude correlations: a meta-analysis. Sci Rep 12, 20683. https://doi.org/10.1038/s41598-022-25276-1

Plesch, A., Shaw, J. H., Ross, Z. E., & Hauksson, E. (2020). Detailed 3D fault representations for the 2019 Ridgecrest, California, earthquake sequence. *Bulletin of the Seismological Society of America*, *110*(4), 1818-1831. https://doi.org/10.1785/0120200053

Reasenberg, P. A., & Jones, L. M. (1989). Earthquake hazard after a mainshock in California. *Science*, *243*(4895), 1173-1176. https://doi.org/10.1126/science.243.4895.1173

Ross, G. J. (2021). Bayesian Estimation of the ETAS Model for Earthquake Occurrences. *Bulletin of the Seismological Society of America*, 111 (3): 1473–1480. https://doi.org/10.1785/0120200198

Ross, Z. E, Cochran, E. S., Trugman, D. T., & Smith, J. D. (2021). 3D fault architecture controls the dynamism of earthquake swarms. *Science*, *368*(6497), 1357-1361. https://doi.org/10.1126/science.abb0779

Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., ... & Zhuang, J. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, *89*(4), 1305-1313. https://doi.org/10.1785/0220180053

Seif, S., A. Mignan, J. D. Zechar, M. J. Werner and S. Wiemer. (2017) Estimating ETAS: the effects of truncation, missing data, and model

assumptions. *Journal of Geophysical Research: Solid Earth*, 122:449-469. https://doi.org/10.1002/2016JB012809

Seif, S., Zechar, J. D., Mignan, A., Nandan, S., & Wiemer, S. (2019). Foreshocks and Their Potential Deviation from General Seismicity. *Bulletin of the Seismological Society of America*, *109* (1), 1–18. https://doi.org/10.1785/0120170188

Shelly, D. R., W. L. Ellsworth, & D. P. Hill (2016), Fluid-faulting evolution in high definition: Connecting fault structure and frequency-magnitude variations during the 2014 Long Valley Caldera, California, earthquake swarm, J. *Geophysical Research Solid Earth*, 121, 1776–1795, https://doi.org/10.1002/2015JB012719.

Stallone A., and Marzocchi W. (2019). Empirical evaluation of the magnitude-independence assumption. *Geophysical Journal International*, 216, 820–839. https://doi.org/10.1093/gji/ggy459

Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M. J., Wiemer, S., Zechar, J. D., Heiniger, L., & Euchner, F. (2018). Prospective CSEP Evaluation of 1-Day, 3-Month, and 5-Yr Earthquake Forecasts for Italy. *Seismological Research Letters*, *89*, 1251-1261. https://doi.org/10.1785/0220180031

Tinti, S., & Mulargia, F. (1987). Confidence intervals of b values for grouped magnitudes. *Bulletin of the Seismological Society of America,* 77 (6): 2125–2134. https://doi.org/10.1785/BSSA0770062125

Van Stiphout, T., J. Zhuang, and D. Marsan (2012), Seismicity declustering. *Community Online Resource for Statistical Seismicity Analysis,* doi:10.5078/corssa52382934. Available at http://www.corssa.org.

Vidale, J. E., & Shearer, P. M. (2006). A survey of 71 earthquake bursts across southern California: Exploring the role of pore fluid pressure fluctuations and

aseismic slip as drivers. *Journal of Geophysical Research: Solid Earth*, 111(B5), https://doi.org/10.1029/2005JB004034.

Zaliapin, I., & Ben-Zion, I. (2013). Earthquake clusters in southern California II: Classification and relation to physical properties of the crust. *Journal of Geophysical Research: Solid Earth*, *118*, 2865-2877. http://doi.org/10.1002/jgrb.50178

Zaliapin, I., & Ben-Zion, I. Correlation between seismic clustering properties and regional physical conditions**.** *Report for SCEC3 funded project*. https://files.scec.org/s3fs-public/reports/2011/11026_report.pdf

Zaliapin, I., Gabrielov, A., Keilis-Borok, V., & Wong, H. (2008). Clustering Analysis of Seismicity and Aftershock Identification. *Physical Review Letters*, *101*, 018501. https://doi.org/10.1103/PhysRevLett.101.018501

Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J., & Jordan, T. H. (2010). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience*, *22*(12), 1836-1847. https://doi.org/10.1002/cpe.1519

Zhuang j., Ogata Y., Vere-Jones D. (2002). Stochastic Declustering of Space-Time Earthquake Occurrences, *Journal of the American Statistical Association*, 97:458, 369-380, https://doi.org/10.1198/016214502760046925

Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, 63(3):207–216. https://doi.org/10.5047/eps.2010.12.010