ICOSSAR 2021
The 13th International Conference on Structural Safety and Reliability (ICOSSAR 2021), June 21-25, 2021, Shanghai, P.R. China
J. Li, Pol D. Spanos, J.B. Chen & Y.B. Peng (Eds)

Validating the medium-term performance of an Epidemic-Type Aftershock Sequence (ETAS) model for application to seismic hazard assessments

Salvatore Iacoletti[a], Gemma Cremen[b], Carmine Galasso[c]

[a]PhD Student, Department of Civil, Environmental and Geomatic Engineering, University College London, UK, salvatore.iacoletti.19@ucl.ac.uk.
[b]Lecturer, Department of Civil, Environmental and Geomatic Engineering, University College London, UK, g.cremen@ucl.ac.uk.
[c]Professor, Department of Civil, Environmental and Geomatic Engineering, University College of London, UK, and Scuola Universitaria Superiore (IUSS) Pavia, Italy, c.galasso@ucl.ac.uk.

ABSTRACT: Moderate-to-large magnitude earthquakes induce considerable short-to-medium-term increases in seismic hazard, due to the subsequent occurrence of aftershocks. This study validates the medium-term (i.e. months to years) performance of an advanced formulation of the Epidemic-Type Aftershock Sequence (ETAS) model for implementation in simulation-based Probabilistic Seismic Hazard Analysis (PSHA). The aim of the work is to determine whether the ETAS models calibrated with two different calibration procedures can forecast adequate numbers of aftershocks, and consistent spatial and magnitude-frequency distributions, for three years after a given large-magnitude mainshock. The validation procedure employs only out-of-sample testing and its evaluation metrics are the catalog-based scores recently proposed for the Collaboratory for the Study of Earthquake Predictability (CSEP) framework. It is demonstrated using the New Zealand catalog of the period 1990-2020, which includes seven sequences. The findings of this study can be used to guide future implementations of the considered ETAS formulation, possibly in combination with a simulation-based mainshock PSHA.

KEYWORDS: Aftershock modeling; Epidemic-Type Aftershock Sequence (ETAS); Model validation; Model calibration; Simulation-based Probabilistic Seismic Hazard Analysis.

1 INTRODUCTION

While conventional probabilistic seismic hazard analysis (PSHA) generally only considers time-independent long-term mainshock occurrences, the importance of also accounting for the short-to-medium-term space-time clustering of earthquakes has been recognized in recent years (e.g. Papadopoulos et al. 2020). The significance of non-mainshock events is underscored by recent earthquake sequences across the world (e.g. 2016 Central Italy sequence), which have demonstrated that aftershocks can be large and damaging (e.g. Papadopoulos et al. 2020).

Various aftershock forecasting (occurrence) models have been proposed in the literature for simulating the possible future seismicity of earthquake sequences. This study focuses on the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata 1998). ETAS is widely used in short-term Operational Earthquake Forecasting (OEF, i.e. the real- or near-real-time simulation of on-going sequences), for which it can provide reasonably good forecasts from a few days to a few months after a specific mainshock (e.g. Werner et al. 2011). However, few studies have examined its ability to bridge the methodological gap between the long-term (i.e. decades) focus of traditional PSHA and the medium-term (i.e. few months to few years) analyses required to capture the hazard increases observed after large mainshocks.

This study examines the challenges of calibrating the ETAS model for simulating aftershock sequences on a time scale that is compatible with simulation-based PSHA. The particular objective of this work is to investigate two different calibration methodologies and assess whether the resulting ETAS models are able to describe common characteristics of the considered sequences (i.e. productivity, magnitude and spatiotemporal distribution of events) in the

medium term (i.e. few months to few years). The first methodology, hereinafter referred to as region-wide calibration, calibrates ETAS model parameters based on the entire historic catalog of a large area (i.e. entire countries or other sizeable spatial extents) and is the conventional approach for ETAS calibration (e.g. Papadopoulos et al. 2020, Zhuang et al. 2011). The second calibration methodology, hereinafter referred to as sequence-averaged calibration, first determines sequence-specific ETAS model parameters (from particular space-time seismicity increases due to a specific mainshock occurrence) and then adopts the average of the resulting values across all sequences. The validation methodology is demonstrated on the New Zealand catalog between 1990 and 2020.

2 BACKGROUND ON ETAS

2.1 *Adopted ETAS model*

The underlying mathematical modeling framework of the ETAS model belongs to the class of branching processes (Zhuang et al. 2011), i.e. each parent (i.e. mainshock or subsequent aftershock) earthquake produces a random number of offspring (or child) aftershocks, according to a fixed probability distribution. Different variations (or formulations) of the model have been proposed by various researchers (see Zhuang et al. 2011). The following ETAS formulation is used in this study (Ogata 1998, Zhuang et al. 2011):

$$\lambda(t,x,y) = \mu(x,y) + \sum_{i:t_i<t} \xi(t,x,y|t_i,x_i,y_i) \quad (1)$$

where the ETAS conditional intensity $\lambda(t,x,y)$ (i.e. the rate of earthquake occurrence given that no event occurred in the previous period, unit days^{-1}·degrees^{-2}) at time t and location (x,y) is the superposition (i.e. sum) of a time-independent background intensity function $\mu(x,y)$ and the epidemic aftershock contribution $\xi(t,x,y|t_i,x_i,y_i)$ of all the i^{th} events that occurred before time t. The time-independent component is usually considered to be proportional to the total seismicity: $\mu(x,y) = \nu \cdot u(x,y)$, where $\nu > 0$ is the relaxation factor and $u(x,y)$ is the total (spatial) seismicity function. The epidemic aftershock contribution can be written as:

$$\xi(t,x,y|t_i,x_i,y_i) = \kappa(m_i)g(t-t_i)f(x-x_i,y-y_i,m_i) \quad (2)$$

where $\kappa(m_i)$ is the mean of a Poisson distribution that describes the number of child events associated with the i^{th} parent earthquake of magnitude m_i. In this study, the following $\kappa(m_i)$ (Zhuang et al. 2011) is used for region-wide ETAS calibrations:

$$\kappa(m_i) = A\, e^{\alpha(m_i - m_{min})} \quad (3)$$

where $A > 0$ and $\alpha > 0$ are unknown parameters and m_{min} is the minimum magnitude below which earthquakes do not trigger other earthquakes (Sornette & Werner 2005). A time-dependent version of Equation 3 is used for sequence-specific ETAS calibrations:

$$\kappa(m_i,t_i) = A\, e^{\alpha(m_i - m_c(t_i))} \quad (4)$$

where $m_c(t_i)$ is the time-dependent completeness magnitude (e.g. Page et al. 2016), which is discussed in the next subsection. $g(t-t_i)$ is the temporal PDF describing the occurrence time of a child event relative to that of the parent event (t_i) and is based on the modified Omori's law (Utsu et al. 1995). It is assumed that this PDF is a function of the time lag $(t-t_i) > 0$ and is independent of t_i. The truncated version of $g(t-t_i)$ used in this study can be written as (Cattania et al. 2018):

$$g(t-t_i) = \begin{cases} \frac{1-p}{(c+t_a)^{1-p}-c^{1-p}}(c+t-t_i)^{-p} \\ \text{if } p \neq 1 \\ \frac{1}{\log(c+t_a)-\log c}(c+t-t_i)^{-p} \\ \text{if } p = 1 \end{cases} \quad (5)$$

where $c > 0$ and p are the unknown parameters of Omori's law, and t_a is the truncation parameter (in days). Hainzl et al. (2016) suggested that the temporal ETAS parameters of the truncated temporal PDF can be calibrated with limited bias for a truncation of 1600 days (around 4.4 years). A 5-year truncation is therefore used in this work (i.e. $t_a = 1825$ days). $f(x-x_i,y-y_i,m_i)$ is the PDF of the location of an aftershock (i.e. offspring event) for a parent event with magnitude m_i and location (x_i,y_i). Several versions of $f(x-x_i,y-y_i,m_i)$ have been proposed in the literature (e.g. Zhuang et al. 2011, Ogata 1998, Guo et al. 2015). The most popular isotropic power-law spatial PDF is (Ogata & Zhuang 2006, Zhuang et al. 2011, Werner et al. 2011):

$$f(x-x_i,y-y_i,m_i) = \frac{q-1}{\pi \cdot D\exp[\gamma(m_i-m_{min})]}\left[1+\frac{(x-x_i)^2+(y-y_i)^2}{D\exp[\gamma(m_i-m_{min})]}\right]^{-q} \quad (6)$$

where q, D, and γ are unknown parameters. A number of studies have attempted to find a more realistic description of the (anisotropic) spatial distribution of aftershocks. Guo et al. (2015) incorporated finite fault models (where available) by dividing the available rupture surface for the i^{th}

earthquake into n_i finite patches and considered each patch an individual triggering source, according to:

$$f(x-x_i, y-y_i, m_i) = \frac{1}{n_i} \frac{q-1}{\pi \cdot D \exp[\gamma (m_p - m_{min})]} \sum_{j=1}^{n_i} \left[1 + \frac{(x-x_{ij})^2 + (y-y_{ij})^2}{D \exp[\gamma (m_p - m_{min})]}\right]^{-q} \quad (7)$$

where the j^{th} patch on the i^{th} surface source has location (x_{ij}, y_{ij}) and is associated with the same magnitude $m_p \geq m_{min}$. m_p can be calculated from the empirical model given by Wells & Coppersmith (1994). In this paper, Equation 7 is used where a finite fault model is available for both calibration and (out-of-sample) validation of ETAS, and Equation 6 is used otherwise. Due to instabilities of the calibration process, the q parameter of the spatial PDF is often constrained to 1.5 (e.g. Lombardi & Marzocchi 2010).

2.2 *Short-term magnitude incompleteness during sequences*

The completeness magnitude often significantly increases after medium-to-large earthquakes due to seismic-network limitations (Helmstetter et al. 2006). Helmstetter et al. (2006) investigated the issue of undetected events after three medium-to-large magnitude earthquakes in the Californian catalog from the Advanced National Seismic System (ANSS), and suggested the following model to describe the short-term variation of the completeness magnitude at time t:

$$m_{c,st}(t, t_k) = m_k - 4.5 - 0.75 \log(t - t_k) \quad (8)$$

where $t - t_k$ is the time lag from the k^{th} event with corresponding magnitude m_k. While Equation 8 is not universal, it has been used for New Zealand studies in the literature (e.g. Cattania et al. 2018). The k^{th} event in Equation 8 is usually the mainshock (e.g. Page et al. 2016). However, the short-term variation of the completeness magnitude can also be observed for large aftershocks. Hence, in this study Equation 8 is used for all events with $m_k \geq 6$, such that the $m_c(t_i)$ in Equation 4 is calculated as (Savran et al. 2020):

$$m_c(t_i) = \max\left(\max_{k:t_k<t_i}\left(m_{c,st}(t, t_k), m_{min}\right)\right) \quad (9)$$

where the internal max operator only considers events before time t_i with $m_k \geq 6$.

2.3 *Simulating synthetic catalogs with ETAS*

The ETAS model can be used to produce multiple realizations of earthquake sequences progressing over time. There are different approaches to building an ETAS simulator (e.g. Harte 2013, Zhuang & Touati 2015). Algorithm #16 in Zhuang & Touati (2015) is used in this study for models that are calibrated using the region-wide approach, to simulate events in a chosen time window and spatial region, starting from a given (non-declustered) catalog of events $\{(t_i, x_i, y_i, m_i): i = 1,2,\ldots,K\}$ up to the starting time of the simulation. For models calibrated using the sequence-averaged approach, the algorithm above is modified by computing $m_c(t_i)$ with Equation 9 at every step and calculating the productivity with Equation 4. The magnitude of the aftershocks is generated from a Gutenberg-Richter (GR) distribution with a minimum magnitude equal to m_{min} for region-wide calibrations, or $m_{c,st}$ from Equation 8 otherwise.

3 CALIBRATION METHODOLOGY

3.1 *Calibrating the ETAS model*

The parameters of the ETAS model that require calibration for a specific analysis are $\boldsymbol{\theta} = (\nu, \alpha, A, c, p, D, q, \gamma)$. Maximum Likelihood Estimation (MLE) is typically used to determine $\boldsymbol{\theta}$ for a given input catalog (H_T), which involves maximizing the log-likelihood $LL(\boldsymbol{\theta}|H_T)$ for "target events" inside a target space region S and a target time window $T = [t_{start}, t_{end}]$ (Ogata 1998):

$$LL(\boldsymbol{\theta}|H_T) = \sum_{i=1}^{N} \delta_i \log(\lambda(t_i, x_i, y_i)) - \int_{t_{start}}^{t_{end}} \iint_S \lambda(t, x, y) dx dy dt \quad (10)$$

where $\delta_i = 1$ if the event is a target event and $\delta_i = 0$, otherwise. Numerical approximations of the spatial integral in Equation 10 are required because it does not have a closed-form solution. Here, the Ogata (1998) radial partitioning method of S is used to approximate the integral term. Maximizing the likelihood is achieved by minimizing the negative log-likelihood, which can be performed with any available numerical method. Here, the trust-region algorithm for large-scale bound-constrained nonlinear optimization problems is used (Lalee et al. 1998). Note that the input catalog extends to an auxiliary area larger than the aforementioned window S to avoid boundary effects, i.e. biases caused by

neglecting events located outside the employed window that trigger, or are triggered by, events within it (Harte 2013, Seif et al. 2017). The time window T is also enlarged to appropriately account for events that occurred before t_{start} (Harte 2013, Seif et al. 2017).

3.2 *Region-wide calibration*

The region-wide calibration procedure involves (e.g. Papadopoulos et al. 2020): (1) developing an historic earthquake catalog for the (large) area of interest (in this study, all magnitudes are converted to moment magnitudes); (2) assessing the completeness magnitude of the whole catalog (e.g. Seif et al. 2017); (3) choosing a magnitude threshold m_{min} and the corresponding catalog length. The resulting catalog should be complete above m_{min} in the target space-time window and is used as the input catalog for the MLE procedure (see Section 3.1); (4) calibrating the b-value of the GR magnitude-frequency distribution (MFD) with the maximum likelihood method proposed by Aki (1965); (5) choosing the target space-time window and the auxiliary space-time window. The auxiliary window captures two years before the start of the target (t_{start}) time window and 100km (~1deg) around the target space window (S) (e.g. Papadopoulos et al. 2020, Harte 2013, Seif et al. 2017); (6) calibrating the background seismicity model with the stochastic declustering procedure (Zhuang et al. 2002); and (7) choosing an ETAS formulation (Eq. 1 in this study) and performing the calibration procedure (see Section 3.1).

3.3 *Sequence-averaged calibration*

Each ETAS model determined using sequence-averaged calibration (herein referred to as sequence-averaged ETAS models) represents the average characteristics (i.e. productivity, magnitude and spatiotemporal distribution of aftershocks) of the included sequences. Calibration of the ETAS model for each sequence is similar to the region-wide procedure (e.g. Guo et al. 2015, Harte 2013), with the following nuances: (1) the sequence-specific ETAS calibration is carried out on each detected sequence (see Section 3.4); (2) each sequence has its own target space-time window, which encompasses seismicity increases due to the mainshock (see Section 3.4); (3) an a-priori uniform background seismicity is adopted to speed up the calibration process (e.g. Papadopoulos et al. 2020); (4) the issue of undetected early aftershocks (see Section 2.2) has a larger effect on the sequence-averaged calibration (e.g. Helmstetter et al. 2006) than the region-wide calibration (see Section 2.2). To avoid potential underestimation due to variable short-term magnitude incompleteness, the b-values of the GR distribution are calculated excluding events that occur within 10 days of mainshocks.

3.4 *Sequence detection and definition*

Single sequences and the corresponding target/auxiliary space-time windows are detected from a given earthquake catalog with the following algorithm: (1) apply Gardner & Knopoff (1974)'s (GK74) declustering technique to the catalog; (2) select sequences with at least 100 aftershocks that occur within the larger space-time window described in Section 3.2; (3) for each selected sequence: (a) apply the Bayesian change point methodology proposed by Gupta & Baker (2017) to identify the points in space where a change in seismicity rate is detected; (b) define the target space region S as the ellipsoid that encompasses all of the detected points in space; (c) define the auxiliary space region as the target space region scaled by a factor of 1.5; (d) select all events from the catalog that occurred within the auxiliary space region and within [-t_a, 985] days of the mainshock (i.e. the largest magnitude event in the considered sequence). 985 days is equivalent to the maximum empirical time window provided by GK74; (e) define the target time window (T) as the time of occurrence of the mainshock (t_{start}) until one day after the last considered event (t_{end}).

4 VALIDATION METHODOLOGY

ETAS models calibrated using the region-wide procedure (herein referred to as region-wide ETAS models) are validated with an adapted version of the walk-forward cross-validation approach. Cross-validation estimates how accurately a predictive model performs on an independent dataset not used for calibration (i.e. retrospective test). A number of region-wide ETAS models are calibrated on portions of the target window, called training periods, which are roughly 50%, 60%, 70%, or 80% of the entire target time window. The first sequence of the testing period (or remaining portion of the target time window) – henceforth referred to as the "testing

sequence" – is used to validate each calibrated model. Note that the same 2-year auxiliary time window (see Section 3.2) is used for all validation models. A similar cross-validation approach is used for sequence-averaged ETAS models. A number of sequence-averaged ETAS models are built by averaging the ETAS parameters of all the sequences in the same training periods used for the region-wide ETAS models. The testing sequence is then used for validation.

Simulations of the testing sequences are carried out with the specific objective of validating only the aftershock part of Equation 1. The ETAS simulator (see Section 2.3) is fed with five years of data before the starting date of the simulations. The starting date of the simulations corresponds to the mainshock occurrence time of the sequence of interest. The simulation region is 500km (~5deg) larger than the region-wide auxiliary space window (Section 3.2), to enable the sequences to extend in space as necessary. The maximum magnitude of the offspring GR distribution is set equal to the magnitude of the testing sequence mainshock. This limit on the maximum magnitude is justified because the ETAS simulator in this study only describes the declustered aftershock seismicity (i.e. aftershocks cannot be larger than the mainshocks of the testing sequences). The maximum magnitude is further capped to the maximum magnitude for the area as specified in a case-specific seismic hazard model. The results of the stochastic declustering process (see step 6 in Section 3.2) are used to simulate the background seismicity for region-wide models (details in Zhuang & Touati 2015). For sequence-averaged models, the corresponding sequence-specific uniform background rate (i.e. sequence-specific ν) is used for this purpose.

The final result of the procedure is an ensemble of 10,000 synthetic catalogs of the considered sequence. The length of the synthetic catalogs is equal to three years (similar in duration to the Canterbury sequence, for example). Model forecasts can be considered acceptable if their predictions are reasonably similar to observations in the testing sequence. In this paper, this similarity is quantified using the new catalog-based metrics developed within the Collaboratory for the Study of Earthquake Predictability (CSEP) (Savran et al. 2020, Zechar 2010), which are now described.

The number (N-) test, with the test scores δ_1 and δ_2, compares the number of forecasted events against the number of observed events. If $\delta_1 < 0.5$, the forecasts generally underpredict the number of events and if $\delta_2 < 0.5$, the forecasts generally overpredict the number of events. The number of forecasted events is deemed to be inconsistent if $\delta_1 < 0.025$ or $\delta_2 < 0.025$ (Savran et al. 2020).

The magnitude (M-) test, with the test score γ_m, evaluates whether the forecasted MFD is inconsistent with the observed one. Larger γ_m values (close to one) indicate greater discrepancies between the forecasted and the observed magnitudes (Savran et al. 2020). The forecast is deemed to be inconsistent if $\gamma_m > 0.95$ (Savran et al. 2020).

The spatial (S-) test, with the test score γ_S, evaluates whether the forecasted spatial distributions are inconsistent with the observed locations. Smaller γ_S values (close to zero) indicate greater inconsistency between the forecasted and observed spatial distribution of events (Zechar 2010). The forecast is deemed to be inconsistent if $\gamma_S < 0.05$ (Savran et al. 2020).

The pseudolikelihood (PL-) test, with the γ_{PL}test score, simultaneously captures discrepancies in the spatial and rate components of the forecasts (i.e. potential discrepancies in both the S-test and the N-test should be reflected in the PL-test). Smaller γ_{PL} values (close to zero) indicate greater inconsistencies between forecasts and observations (Zechar 2010). The forecast is deemed to be inconsistent if $\gamma_{PL} < 0.05$ (Savran et al. 2020).

The testing spatial region (i.e. where the CSEP tests are evaluated) for both region-wide and sequence-averaged model validations is a square region at least 100km (~1deg) larger than the target space region defined for each sequence of interest (see Section 3.4). A square grid of 0.1deg (centred on $1/10^{th}$ degree coordinates of latitude and longitude) and a magnitude grid from 4 to 8 in 0.1 increments are used in this study for these tests (Savran et al. 2020).

5 CASE STUDY

5.1 *New Zealand catalog*

The GeoNet catalog of New Zealand earthquakes (https://quakesearch.geonet.org.nz/) is used as a case study in this paper. The GeoNet catalog measures event magnitudes across several different scales. The base magnitudes reported on the GeoNet website are converted to M_W, as follows: (1) approximately 2,660 magnitudes are set to the corresponding M_W

estimated from the regional moment tensor analysis, available since 2003 (Ristau 2009); (2) the remaining M_L for events before September 2012 (401,932 events) are converted using the equations proposed by Ristau (2013); (3) the remaining M_L for events after September 2012 (1,951 events) are converted using the equations of Ristau (2016); and (4) GeoNet preferred magnitudes (187,013 events with symbol M in the original GeoNet catalog) are assumed to be equivalent to M_W.

Consistent with the CSEP experiment for New Zealand (Gerstenberger & Rhoades 2010), only shallow events with a maximum depth of 40km are included in the following analyses. The events are also filtered in space using the boundaries given by Gerstenberger & Rhoades (2010) (CSEP region), but extended by about 1deg in all directions. The resulting catalog appears to be complete from 1990 for a M_W above 3.5 (according to the method of Stepp 1972) and contains 11,022 events. The calculated b-value of the GR distribution is 0.95. Figure 1 shows the location of events with depth $\leq$ 40 km that are included in the GeoNet catalog between 1990 to 2020, along with the CSEP region.

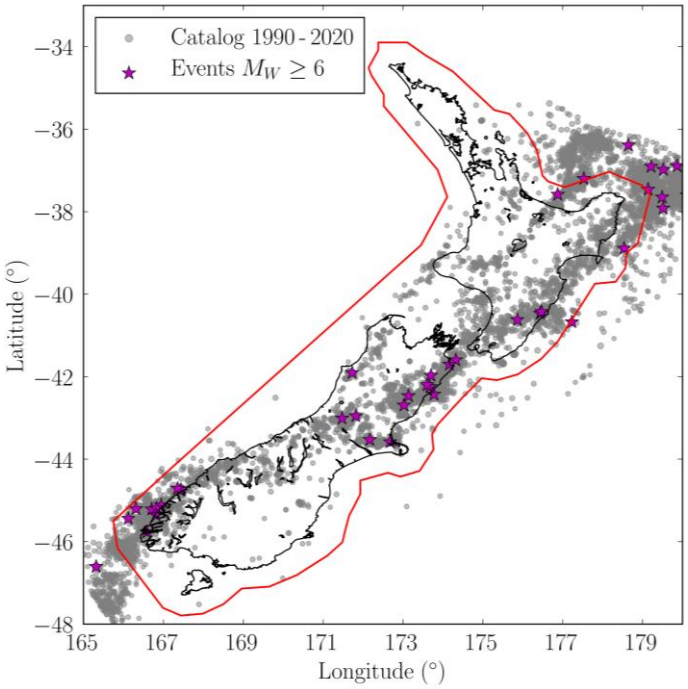

Figure 1. Location of depth $\leq$ 40 km in the 1990-2020 GeoNet catalog. The red polygon represents the CSEP region (Gerstenberger & Rhoades, 2010).

5.2 *New Zealand sequence detection*

Nine sequences are detected in the New Zealand GeoNet catalog between 1990 and 2020. These are numbered from #1 to #9 according to the magnitude of the mainshock (in descending order). Sequences #1, #3 and #8 are the 2016 Kaikōura, the 2010-2012 Canterbury and 2013 Cook Strait sequences, respectively. Figure 2 shows sequence #1 (2016 Kaikōura) with the corresponding finite fault model from Hamling et al. (2017).

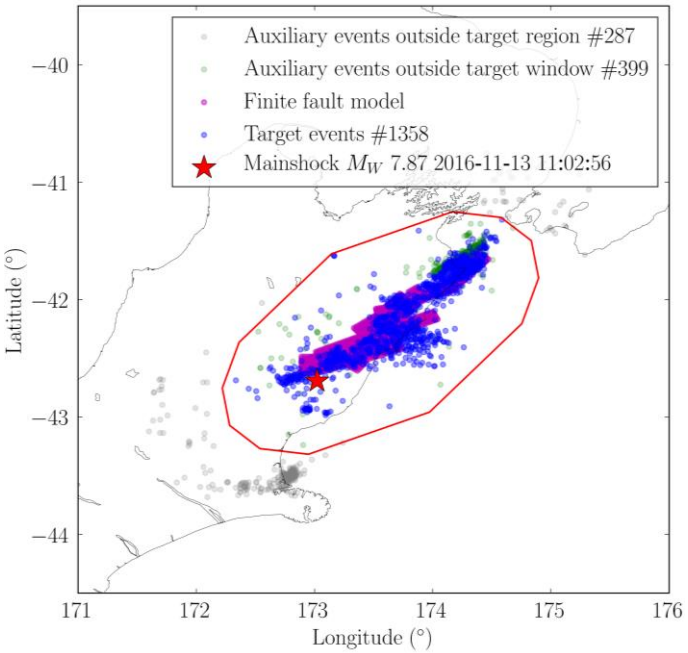

Figure 2. Sequence #1 (2016 Kaikōura), including the finite fault model from Hamling et al. (2017). The red ellipsoid is the target space region generated with the algorithm in Section 3.4.

5.3 *Results of the region-wide calibration*

Table 1 reports the training data and corresponding testing sequences used for the region-wide ETAS models. The auxiliary time window adopted is 1990-1992. The target space region is the CSEP region (Gerstenberger & Rhoades 2010; see Section 5.1 and Fig. 1). The auxiliary area extends about 1deg in all directions beyond the edge of the target region (Harte 2013). The settings of the stochastic declustering procedure (see step 6 in Section 3.2) are taken from Harte (2013) and Zhuang et al. (2011). The calibrated values of $\boldsymbol{\theta}$ are shown in Table 2.

Table 1. Training periods used for calibrating each region-wide ETAS model. Also shown are the corresponding testing sequences in the testing period.

Model	Start	End	Target events	Auxiliary events	Testing sequence
NZ-07	1/1/1992	1/1/2007	2438	2896	#2
NZ-10	1/1/1992	1/1/2010	3066	3060	#3
NZ-13	1/1/1992	1/1/2013	4063	3375	#8
NZ-16	1/1/1992	1/1/2016	4927	3541	#1

Table 2. Region-wide ETAS model parameters.

Model	ν	A	α	c	p	D	q	γ	b
NZ-07	0.90	0.19	1.62	2.9E-3	1.09	4.8E-4	1.96	0.84	1.00
NZ-10	0.91	0.25	1.36	2.8E-3	1.08	5.3E-4	1.94	0.78	0.94
NZ-13	0.93	0.25	1.43	2.4E-3	1.07	3.4E-4	1.86	0.78	0.94
NZ-16	0.94	0.24	1.46	2.8E-3	1.07	2.7E-4	1.84	0.79	0.96

5.4 *Results of the sequence-averaged calibration*

The $q = 1.5$ constraint is used to calibrate the sequence-specific ETAS models (see Section 2.1). Sequences #2 and #5 are not stable (see Harte 2013). This means that the associated sequence-specific ETAS models (#2 and #5) are not capable of producing a finite number of events in an infinite amount of time and so the model parameters are not plausible. Sequences #2 and #5 are thus not used for the sequence-averaged models in the following.

The sequence-averaged ETAS models are built by averaging all calibrated parameters of the sequences in the training periods of Table 1 (i.e. S_{AV}-07 is the sequence-averaged equivalent of NZ-07); see Table 3.

Table 3. Sequence-averaged ETAS model parameters.

Model	A	α	C	p	D	q	γ	b
S_{AV}-07	0.34	1.56	1.3E-2	1.22	2.5E-4	1.5	0.98	1.10
S_{AV}-10	0.34	1.56	1.3E-2	1.22	2.5E-4	1.5	0.98	1.10
S_{AV}-13	0.33	1.60	1.1E-2	1.19	2.1E-4	1.5	0.90	1.10
S_{AV}-16	0.33	1.56	1.1E-2	1.20	1.9E-4	1.5	0.79	1.09

5.5 *Validation of the region-wide models*

Each region-wide ETAS model is validated using the corresponding (out-of-sample) testing sequence in Table 1 (i.e. retrospective testing is used). Synthetic catalogs for these sequences are generated with the corresponding ETAS parameters in Table 2. For this case study, the cap to the maximum magnitude of the GR distribution is $M_W 7.2$, which is equal to the maximum magnitude chosen for most seismogenic areas in the National Seismic Hazard Model for New Zealand (Stirling et al. 2012). The b-value of the GR distribution is reported in Table 2. Table 4 provides the CSEP test scores of the validation. As an example, Figure 3 displays the 95% range (2.5^{th} - 97.5^{th} percentiles) forecast of the 10,000 simulations produced by NZ-16 for sequence #1.

According to Table 4, all region-wide ETAS models fail the N-test (i.e. $\delta_1 < 0.025$), significantly underpredicting the number of events in the testing sequences. All models also fail the M-test; NZ-07 and NZ-10 fail the S-test and the PL-test. This implies that the calibrated region-wide ETAS models are not able to capture the characteristics of the corresponding testing sequence, at least for the case study examined here.

Table 4. Cross-validation results for each region-wide ETAS model. (F) stands for failed.

Model	δ_1	δ_2	γ_m	γ_S	γ_{PL}
NZ-07	0.017 (F)	0.983	0.982 (F)	0.007 (F)	0.003
NZ-10	0.000 (F)	1.000	1.000 (F)	0.007 (F)	0.000
NZ-13	0.000 (F)	1.000	0.999 (F)	0.974	1.000
NZ-16	0.000 (F)	1.000	0.999 (F)	0.245	0.718

5.6 *Validation of the sequence-averaged models*

Simulations of the testing sequences for the sequence-averaged ETAS models are carried out using the maximum magnitude discussed in Section 5.5 and the b-values in Table 3. Validation results for the sequence-averaged ETAS models are reported in Table 5.

All CSEP N-test scores (δ_1 and δ_2) (except in the case of S_{AV}-07) are greater than 0.025, which indicates that the number of forecasted events is consistent with the observed number in each corresponding sequence. This inference is reflected in Figure 3, which compares observed sequence #1 with the 95% range (2.5^{th}-97.5^{th} percentiles) forecast of the 10,000 simulations produced by S_{AV}-16. Except in the case of sequence #8, M-test (γ_m) scores are also satisfactory, meaning that the forecasted MFDs are consistent with those observed. All sequences pass the S-test (γ_S) score, implying that the forecasted spatial distributions of aftershocks are consistent with observations. The generally satisfactory N-test and S-test performances are reflected in the PL-test scores (i.e. $\gamma_{PL} \geq 0.05$ in all cases). Based on three-year forecasts of the same testing sequences, the majority of CSEP scores imply that the sequence-averaged ETAS models have notably better predictive power than the region-wide models.

Table 5. Cross-validation results for each sequence-averaged ETAS model. (F) stands for failed.

ETAS	δ_1	δ_2	γ_m	γ_S	γ_{PL}
S_{AV}-07	1.000	0.000 (F)	0.002	0.220	0.053
S_{AV}-10	0.307	0.695	0.761	0.393	0.591
S_{AV}-13	0.292	0.713	0.689	1.000	1.000
S_{AV}-16	0.473	0.530	0.001	0.661	0.727

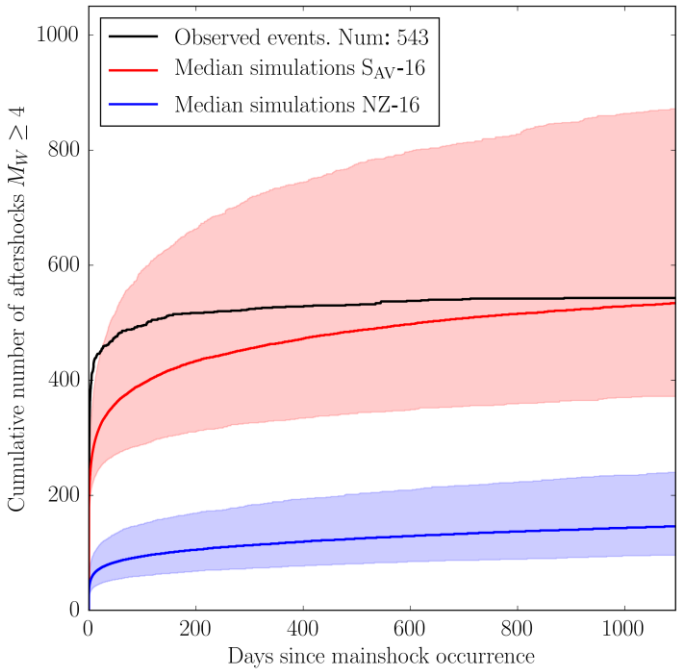

Figure 3. Observed sequence #1 (2016 Kaikōura), compared with the corresponding median and the 95% range (2.5^{th} and 97.5^{th} percentiles) forecast of the 10,000 simulations produced by NZ-16 and S_{AV}-16.

6 DISCUSSION

The (retrospective) validations of the two considered calibration methodologies indicate that the sequence-averaged ETAS models (Section 5.4) perform better than conventional region-wide ETAS models (Section 5.3). The number of aftershocks generated by the sequence-averaged ETAS models in the three-year long synthetic catalogs is consistent with the observed (testing) sequences, whereas the corresponding region-wide models fail to reproduce a reasonable number. The M-test, the S-test and the PL-test also generally produce better results in the case of the sequence-averaged ETAS models. This means that averaging the ETAS parameters across detected sequences leads to more accurate forecasts of common aftershocks characteristics than the conventional region-wide ETAS calibration procedure (e.g. Zhuang et al. 2011), at least in the New Zealand case study considered here.

The poor N-test performance of the region-wide ETAS models is caused by the underestimation of the productivity parameters (α in particular), which is not new in the literature (e.g. Harte 2013, Papadopoulos et al. 2020). It has been demonstrated that the region-wide calibration procedure can lead to significantly underestimated values of α due to: (1) the use of the isotropic spatial PDF (Eq. 6); (2) the excessive smoothing of the background distribution (e.g. Harte 2013); and (3) the aftershock incompleteness (Hainzl et al. 2013). In this study, an anisotropic spatial PDF (Eq. 7) is used to mitigate problem (1), where a finite fault model is available. The smoothing bandwidth suggested by Harte (2013) for New Zealand is used to avoid (2) in this work. Problem (3) is difficult to address for region-wide calibrations, as the available time-dependent magnitude incompleteness formulation (i.e. Eq. 8) is designed for sequences and does not contain the required space component. This issue is addressed in the sequence-specific calibrations, by including the time-dependent magnitude incompleteness model directly in the ETAS formulation. This is consistent with the approach proposed by Page et al. (2016).

Another reason why region-wide calibrations can be inadequate is the fact that the common space-time ETAS formulation has constant parameters while seismicity patterns vary spatially, showing various clustering features due to geological and tectonic processes. The sequence-specific calibrations take this heterogeneity into account and allow the analyst to focus on the declustered seismicity associated with large sequences (which is a key concern when coupling a mainshock PSHA with an aftershock simulator). The sequence-averaged models then represent the average characteristics of the considered large sequences.

One limitation of this study is that the time-dependent m_c violates some assumptions of the ETAS model. Commonly used ETAS formulations (e.g. Papadopoulos et al. 2020, Zhuang et al. 2011), consider the MFD to be independent of the rate in Equation 1 and the input catalog (for calibrating the model) to be complete. Among others, these theoretical assumptions are used in Equation 10 and the MLE process (Zhuang et al. 2011). However, the empirical earthquake data available soon after a medium-to-large magnitude earthquake do not support these assumptions (see Section 2.2). The time-dependency of m_c is used in this paper to improve the performance of the ETAS model, as it is

well-established in the literature and a key component for improving aftershock forecasts (Page et al. 2016). Introducing region-specific magnitude incompleteness models might further improve the forecasting performance. In fact, there is some sequence-to-sequence variability of the parameters in Equation 8 (e.g. Helmstetter et al. 2006), such that they should be evaluated on an individual sequence basis. Other means to improve the performance of the sequence-averaged ETAS models (not accounted for in this study) could be: including tectonic region information, exploring the intersequence variability in test performance and constraining $\alpha = \beta$, where β is the logarithmic version of the GR b-value (e.g. Page et al. 2016, Helmstetter et al. 2006, Seif et al. 2017, Papadopoulos et al. 2020).

Another limitation of this study is that the tests are, in a strict sense, only partially retrospective. This is because the simulations account for the finite fault models (where available), which may not be known in a fully retrospective test (e.g. Cattania et al. 2018). This limitation does not invalidate the conclusions of the study, however. In fact, approximate finite fault models for ETAS can be derived from estimated fault geometries, based on a minimal amount of summary data (i.e. on seismogenic depths, rake, dip and strike angles), or more complex fault-based source models (e.g. Iacoletti et al. 2021) that are available from the mainshock simulations.

7 CONCLUSIONS

Damaging earthquake sequences and short-to-medium-term (i.e. on the order of a few years) increases in seismic hazard that follow moderate-to-large magnitude earthquakes demonstrate the need to account for aftershock occurrence in long-term PSHAs. However, literature on calibrating and validating aftershock forecast models for use in simulation-based PSHAs is somewhat scarce. This study applied a cross-validation methodology to evaluate two different aftershock occurrence (i.e. ETAS) calibration procedures and assess whether the resulting models are able to capture common characteristics (i.e. productivity, magnitude and spatiotemporal distribution of events) of the considered sequences. The first investigated procedure is the commonly used region-wide calibration approach, where the entire historic earthquake catalog for the (wide) region at hand is used to fit the ETAS model. The second (sequence-averaged) calibration methodology first fits ETAS models on a sequence-by-sequence basis and then averages the resulting parameters.

The employed cross-validation methodology involves only out-of-sample (i.e. retrospective) testing of the calibrated ETAS models, using the catalog-based scores recently proposed for the CSEP framework (Savran et al. 2020) as evaluation metrics for the forecasting performance. The methodology was demonstrated for the New Zealand catalog and seven sequences in the period 1990-2020. Resulting test scores imply (at least for the case study of interest) that the conventional region-wide calibration procedure does not produce ETAS models that appropriately capture aftershock characteristics. This suggests that the conventional region-wide calibration procedure may not be suitable for building an ETAS simulator to generate aftershocks in combination with a mainshock PSHA model. On the other hand, sequence-averaged ETAS models predict more accurate synthetic catalogs in terms of numbers of aftershocks, as well as spatial and magnitude-frequency distributions.

This study provided useful insights for calibrating an ETAS model in the context of simulation-based PSHA. In particular, based on the findings of this work, it is recommended that modelling general aftershock behaviour in a region of interest with ETAS begins by calibrating model parameters on a sequence-by-sequence basis (instead of considering the whole historic catalog) and then averaging the resulting values.

REFERENCES

Aki, K. 1965. Maximum likelihood estimate of b in the formula logN=a-bM and its confidence limits. *Bulletin of the Earthquake Research Institute Tokyo University* 43: 237–239.

Cattania, C., Werner, M.J., Marzocchi, W., Hainzl, S., Rhoades, D., Gerstenberger, M., Liukis, M., Savran, W., Christophersen, A., Helmstetter, A., Jimenez, A., Steacy, S., Jordan, T.H. 2018. The Forecasting Skill of Physics-Based Seismicity Models during the 2010–2012 Canterbury, New Zealand, *Earthquake Sequence. Seismological Research Letters* 89, 1238–1250.

Gardner, J.K., Knopoff, L. 1974. Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America* 64, 1363–1367.

Gerstenberger, M.C., Rhoades, D.A. 2010. New Zealand Earthquake Forecast Testing Centre. *Pure Appl. Geophys.*

167, 877–892.

Guo, Y., Zhuang, J., Zhou, S. 2015. An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *J. Geophys. Res. Solid Earth* 120, 3309–3323.

Gupta, A., Baker, J.W. 2017. Estimating spatially varying event rates with a change point using Bayesian statistics: Application to induced seismicity. *Structural Safety* 65, 1–11.

Gutenberg, B. & Richter, C.F. 1944. Frequency of earthquakes in California. *Bulletin of the Seismological Society of America* 34: 185–188.

Hainzl, S., Christophersen, A., Rhoades, D. & Harte, D. 2016. Statistical estimation of the duration of aftershock sequences. *Geophysical Journal International* 205(2): 1180–1189.

Hainzl, S., Zakharova, O. & Marsan, D. 2013. Impact of aseismic transients on the estimation of aftershock productivity parameters. *Bulletin of the Seismological Society of America* 103(3): 1723–1732.

Hamling, I.J., Hreinsdóttir, S., Clark, K., Elliott, J., Liang, C., Fielding, E., Litchfield, N., Villamor, P., Wallace, L., Wright, T.J., D'Anastasio, E., Bannister, S., Burbidge, D., Denys, P., Gentle, P., Howarth, J., Mueller, C., Palmer, N., Pearson, C., Power, W., Barnes, P., Barrell, D.J.A., Van Dissen, R., Langridge, R., Little, T., Nicol, A., Pettinga, J., Rowland, J., Stirling, M., 2017. Complex multifault rupture during the 2016 M w 7.8 Kaikōura earthquake, New Zealand. *Science* 356, eaam7194.

Harte, D.S. 2013. Bias in fitting the ETAS model: A case study based on New Zealand seismicity. *Geophysical Journal International* 192(1): 390–412.

Iacoletti, S., Cremen, G., Galasso, C., 2021. Advancements in multi-rupture time-dependent seismic hazard modeling, including fault interaction. *Earth-Science Reviews* 220, 103650.

Lalee, M., Nocedal, J., & Plantega, T. 1998. On the implementation of an algorithm for large-scale equality constrained optimization. *SIAM Journal on Optimization* 8.3: 682-706.

Lombardi, A. M., Marzocchi, W. 2010. The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment, *Ann. Geophys* 53, no. 3, 155–164.

Ogata, Y. 1998. Space-time point-process models for earthquake occurrences, *Annals of the Institute of Statistical Mathematics*, 50, 379-402.

Ogata, Y., & Zhuang, J. 2006. Space-time ETAS models and an improved extension, *Tectonophysics*, 413(1-2), 13-23.

Page, M.T., van der Elst, N., Hardebeck, J., Felzer, K., Michael, A.J. 2016. Three Ingredients for Improved Global Aftershock Forecasts: Tectonic Region, Time-Dependent Catalog Incompleteness, and Intersequence Variability. *Bulletin of the Seismological Society of America* 106, 2290–2301.

Papadopoulos, A.N., Bazzurro, P., Marzocchi, W. 2020. Exploring probabilistic seismic risk assessment accounting for seismicity clustering and damage accumulation: Part I. Hazard analysis: Earthquake Spectra.

Ristau, J. 2009. Comparison of Magnitude Estimates for New Zealand Earthquakes: Moment Magnitude, Local Magnitude, and Teleseismic Body-Wave Magnitude. *Bulletin of the Seismological Society of America* 99, 1841–1852.

Ristau, J. 2013. Update of Regional Moment Tensor Analysis for Earthquakes in New Zealand and Adjacent Offshore Regions. *Bulletin of the Seismological Society of America* 103, 2520–2533.

Ristau, J., Harte, D., Salichon, J., 2016. A Revised Local Magnitude (ML) Scale for New Zealand Earthquakes. *Bulletin of the Seismological Society of America* 106, 398–407.

Savran, W.H., Werner, M.J., Marzocchi, W., Rhoades, D.A., Jackson, D.D., Milner, K., Field, E., Michael, A. 2020. Pseudoprospective Evaluation of UCERF3-ETAS Forecasts during the 2019 Ridgecrest Sequence. *Bulletin of the Seismological Society of America*.

Seif, S., Mignan, A., Zechar, J.D., Werner, M.J. & Wiemer, S. 2017. Estimating ETAS: The effects of truncation, missing data, and model assumptions. *Journal of Geophysical Research: Solid Earth* 122(1): 449–469

Sornette, D., Werner, M.J., 2005. Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences. *Journal of Geophysical Research: Solid Earth* 110.

Stepp, J.C. 1972 Analysis of completeness of the earthquake sample in the Puget Sound area and its effect on statistical estimates of earthquake hazard. In: *Proc. 1st Internat. Conf. Microzonazion*, Seattle, v.2, pp.897–910.

Stirling, M., McVerry, G., Gerstenberger, M., Litchfield, N., Van Dissen, R., Berryman, K., Barnes, P., Wallace, L., Villamor, P., Langridge, R., Lamarche, G., Nodder, S., Reyners, M., Bradley, B., Rhoades, D., Smith, W., Nicol, A., Pettinga, J., Clark, K., Jacobs, K. 2012. National Seismic Hazard Model for New Zealand: 2010 Update. *Bulletin of the Seismological Society of America* 102, 1514–1542.

Utsu, T., Ogata, Y., Matsu'ura S.R. 1995. The Centenary of the Omori Formula for a Decay Law of Aftershock Activity. *J,Phys,Earth* 43, 1–33.

Wells, D.L., Coppersmith, K.J., 1994. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America* 84, 974–1002.

Werner, M.J., Helmstetter, A., Jackson, D.D. & Kagan, Y.Y. 2011, High-Resolution Long-Term and Short-Term Earthquake Forecasts for California, *Bull. Seism. Soc. Am.*, 101(4).

Zechar, J.D., Gerstenberger, M.C., Rhoades, D.A., 2010. Likelihood-Based Tests for Evaluating Space-Rate-Magnitude Earthquake Forecasts. *Bulletin of the Seismological Society of America* 100, 1184–1195.

Zhuang, J. & Touati, S. 2015. Stochastic simulation of earthquake catalogs, *Community Online Resource for Statistical Seismicity Analysis.*

Zhuang, J., Werner, M.J., Hainzl, S., Harte, D. & Zhou, S. 2011, Basic models of seismicity: spatiotemporal models, *Community Online Resource for Statistical Seismicity Analysis*.

Zhuang, J., Ogata, Y., & Vere-Jones, D. 2002. Stochastic declustering of space-time earthquake occurrences, *Journal of the American Statistical Association*, 97(3), 369-380.