The Q -Score: A Magnitude-Weighted Goodness-of-Fit Score for Earthquake Forecasting

Citation for the original published paper (version of record):
Jansson Valter, J., Arjon, A., Serafini, F. et al (2026). The Q -Score: A Magnitude-Weighted Goodness-of-Fit Score for Earthquake Forecasting. Environmetrics, 37(5).
http://dx.doi.org/10.1002/env.70113

Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

The *Q*-Score: A Magnitude-Weighted Goodness-of-Fit Score for Earthquake Forecasting

Julia Jansson Valter[1] | Alejandra Arjon[2] | Francesco Serafini[3] | Frederic Schoenberg[2]

[1]Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Gothenburg, Sweden | [2]Department of Statistics & Data Science, UCLA, Los Angeles, USA | [3]School of Earth Sciences, University of Bristol, Bristol, UK

Correspondence: Frederic Schoenberg (frederic@stat.ucla.edu)

Received: 15 November 2025 | **Revised:** 3 June 2026 | **Accepted:** 22 June 2026

Keywords: CSEP | earthquakes | model evaluation | point process

ABSTRACT

Accurate forecasting of large earthquakes is of great importance, yet most current earthquake forecast evaluation metrics, such as the log-likelihood score, do not give additional weight to large-magnitude events. To this end, magnitude-weighted goodness-of-fit scores for earthquake forecasting have recently been introduced, such as potency-weighted log-likelihood and the Q-score. In this article, we investigate properties of the Q-score, which is a quotient emphasizing model fit for the largest 5% of earthquakes. We explore the theoretical properties of the Q-score, demonstrating that under certain null conditions, the expectations of the numerator and denominator are equal and thus the expectation of Q is 1 in a ratio sense. Additionally, the score satisfies a law of large numbers. We evaluated the Q-score of 21 next-day gridded earthquake forecasts for California, provided by the Collaborative for the Study of Earthquake Predictability (CSEP) for the years 2012, 2014, and 2017. We also calculate the log-likelihood scores of the forecasts to obtain a more comprehensive evaluation of how well different models perform, both for predicting the largest events and also in terms of overall model fit.

1 | Introduction

Although seismic regions such as California frequently experience earthquakes, the vast majority are minor and do not significantly disrupt daily life. Major earthquakes occur rarely, but the accurate forecasting of them is essential to facilitate preparedness, mitigate fatalities, and reduce infrastructure damage. Earthquakes can be modeled with spatial–temporal point processes, such as the Epidemic-Type-Aftershock-Sequence (ETAS) model (Ogata 1988). While effective at predicting aftershock sequences, these models unfortunately do not substantially outperform a baseline homogeneous Poisson model in the context of large-magnitude earthquake prediction (Schoenberg and Schorlemmer 2024).

The Collaboratory for the Study of Earthquake Predictability (CSEP) is a global-scale project designed to support a wide range of forecasting experiments by building community standards and protocols for comparative testing (Zechar et al. 2010; Michael and Werner 2018; Schorlemmer et al. 2018; Graham et al. 2024). CSEP is involved in the development of rigorous scores and hypothesis tests used to evaluate and compare earthquake forecasting models. The model evaluation scores currently employed are useful for assessing overall model fit to observed data, but do not emphasize the fit to the largest events. For example, the log-likelihood score, used in some of these tests, does not explicitly weight the larger magnitude events more heavily than the smaller ones (Schoenberg and Schorlemmer 2024). As a complement to existing evaluation methods, Schoenberg and Schorlemmer (2024) introduced the Q-score, which rewards models for accurately estimating the largest 5% of magnitude earthquakes.

In this article, the Q-score is defined, and its statistical properties are explored. It is shown that for a marked spatial–temporal point process with separability in the marks, the expectation of the numerator of Q and the expectation of the denominator of Q are equal. This means that the expectation of Q is 1 in a ratio sense. Further, we show a law of large numbers for Q, stating that Q approaches 1 asymptotically, under additional assumptions of stationarity. These assumptions lay out the conditions for a null hypothesis against which the Q-score for a particular model can be tested. Specifically, in applications, one may wish to set the null hypothesis that $Q = 1$, and inspect whether the observed value of Q for a given forecast exceeds 1 by a statistically significant margin. The Q-score is evaluated for several next-day gridded forecasts produced by 21 models operated fully prospectively by CSEP (Serafini, Bayona, Silva, et al. 2025). The evaluation is performed for the years 2012, 2014, and 2017, for all 21 models. The earthquake forecast models evaluated with the Q-score include ETAS and its variants, ensemble models of ETAS, and alternative models such as the Short-term Earthquake Probability (STEP) model, whose Q-scores are comparable with those of the ETAS models. We also evaluate the models using the log-likelihood, L-score, as a complement to the Q-score. The use of Q and L-scores in tandem allows us to evaluate simultaneously whether a model performed well at forecasting the largest events, and whether it fit well to the dataset as a whole.

The article is structured as follows. In Section 2, we provide theoretical background and details of a few models used in CSEP. Section 3 describes the Q-score and presents theoretical results. Application to California seismicity and results can be found in Section 4, followed by a discussion in Section 5.

2 | Background

A *point process* is a measurable mapping from a filtered probability space $(\Omega, \mathcal{F}, \mathcal{P})$ onto $\mathcal{N}$ (Daley and Vere-Jones 2003). Here, $\mathcal{N}$ is the set of $\mathbb{Z}^+$-valued random measures (counting measures) on a complete separable metric space (c.s.m.s.) $\mathcal{X}$, where $\mathbb{Z}^+$ denotes the set of positive integers. We will consider point processes that are *boundedly finite*, that is, processes having only a finite number of points inside any bounded set. Further, we consider a *spatial-temporal* point process, where $\mathcal{X}$ is a subset of $\mathbb{R}^+ \times \mathbb{R}^2$, where $\mathbb{R}^+$ and $\mathbb{R}^2$ represent the set of positive real numbers and the 2-dimensional Euclidean space, respectively. The point process is assumed to be adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ containing all information on the process N at all locations and all times up to and including time t. Continuing, we will assume the spatial domain $\mathcal{S}$ of the point process is a finite and bounded portion of the plane $\mathbb{R}^2$ and denote point i of the process as $\boldsymbol{\tau_i} = (t_i, x_i, y_i)$.

A process is $\mathcal{F}$-predictable if it is adapted to the filtration generated by the left-continuous processes $\mathcal{F}_{(-)}$. Assuming it exists, the $\mathcal{F}$-conditional intensity λ of N is an integrable, non-negative, $\mathcal{F}$-predictable process such that for $\boldsymbol{\tau} = (t, x, y)$

$$\lambda(\boldsymbol{\tau}) = \lim_{h,\delta\to 0} \frac{\mathbb{E}[N([t, t+h) \times \mathbb{B}_{(x,y),\delta}) | \mathcal{F}_{t-}]}{h\pi\delta^2}, \tag{1}$$

where $\mathbb{B}_{(x,y),\delta}$ is a ball centered at location (x, y) with radius δ, and $\mathcal{F}_{t-}$ represents the history of the process N up to but not including time t. Here, $\pi\delta^2$ is the area of $\mathbb{B}_{(x,y),\delta}$, and $N([t, t+h) \times \mathbb{B}_{(x,y),\delta})$ is the number of points occurring in the time interval $[t, t+h)$, and in the ball $\mathbb{B}_{(x,y),\delta}$. The conditional intensity is the expected number of points of the point process happening in a given subregion of the space-time domain, with the volume of the subregion approaching zero.

A point process is *simple* if, with probability one, no two points occur at the same time and location. In this article, we will only consider simple point processes. Since the conditional intensity λ uniquely determines the finite-dimensional distributions of any simple point process (Daley and Vere-Jones 2003, Proposition 7.2.IV), one typically models a simple spatial–temporal point process by specifying a model for λ.

We want to consider a marked spatial–temporal process, since we observe magnitudes, in addition to the space-time locations. A point process $\breve{N}$ on $\mathcal{X} \times \mathcal{M}$ with point i as $(\boldsymbol{\tau_i}, m_i)$, is called a *marked point process* with marks $m_i \in \mathcal{M}$ if the projection N with point i as $\boldsymbol{\tau_i}$, exists as a well-defined point process on $\mathcal{X}$. Here, $M(\boldsymbol{u}) \in \mathcal{M}$, $\boldsymbol{u} \in \mathcal{X}$, is the mark function/random field assigning marks to N, so that $m_i = M(\boldsymbol{\tau_i})$ for each point i in N. We consider the space $\mathcal{X} \times \mathcal{M}$ and points $z_i = (\boldsymbol{\tau_i}, m_i)$, and the marked point process $\breve{N}$ with the marked conditional intensity $\breve{\lambda}(\boldsymbol{\tau}, m)$, where in the context of magnitudes, we let the mark space be $\mathcal{M} = \mathbb{R}$. Similarly to (Daley and Vere-Jones 2008), we call the unmarked spatial–temporal process N the *ground process*, $\mathcal{X}$ the *ground space* and $\mathcal{M}$ the *mark space*. If the marks are i.i.d., that is each point $\boldsymbol{\tau_i}$ is assigned a mark m_i with the same probability distribution $\pi(m_i)$, independent of the other points, conditional on the ground process, it is meaningful to call $\pi(m)$ the *mark density*. Some models assume separability in the marks, that is $\hat{\lambda}(\boldsymbol{\tau}, m) = \lambda(\boldsymbol{\tau})\pi(m)$.

2.1 | CSEP Earthquake Forecast Models

The Collaboratory Study for Earthquake Predictability (CSEP) provides access to a database of next-day forecasts[1] containing daily forecasts produced by 21 models ranging from the dates of 01 August, 2008 to 31 August 2018 for the California region. In this time span, the models have been operated in a fully prospective fashion, independently of the modelers, at the SCEC testing center (Zechar et al. 2010) These grid-based forecasts cover a 0.1° × 0.1° latitude, longitude space grid with 0.1 magnitude bin intervals, which range from 3.95 to 8.95 and one open-end bin for magnitudes over 8.95 (Serafini, Bayona, Silva, et al. 2025). The depths are contained in a large bin ranging from 0km to 30km. The expected earthquake rate for the day forecast starts from 00:00:00 UTC (coordinated universal time) until 23:59:59 UTC. For each space-magnitude grid given the depth range, a daily rate is calculated.

In this section, we discuss two of the models that scored the highest, in terms of the Q-score, when their next-day gridded forecasts were evaluated. Ten of the earthquake forecasting models run by CSEP are a version of the Epidemic-Type Aftershock Sequence (ETAS) model. Another four are ensemble models, which are linear combinations of alternative models (to ETAS) with ETAS. The remaining seven of the 21 models in CSEP, for the time periods considered, are alternatives to ETAS. All 21 earthquake forecasting models considered base their magnitude distribution on the Gutenberg-Richter (Gutenberg and Richter 1944) relationship. This relationship captures the relative frequency of earthquakes in relation to their magnitude. The Omori-Utsu law (Omori 1894; Utsu 1961; Davidsen et al. 2015), is also used. This law describes the trend of expected number of aftershocks after a parent event.

2.1.1 | ETAS

ETAS was first introduced by Ogata (1988), as a marked temporal process with no spatial aspect. ETAS models belong to the family of marked Hawkes point processes (Hawkes 1971), which have conditional intensity obeying

$$\lambda(t|\mathcal{F}_t) = \mu + \sum_{i:t_i<t} \kappa(m_i) g(t - t_i), \tag{2}$$

where $\mathcal{F}_t$ is the history of all events up to time t. Here, μ is the background seismicity rate, $g(t - t_i)$ is the intensity of the process describing the occurrence of the offspring of the event occurred at time t_i, $\kappa(m) = e^{\beta(m-m_0)}$ is the aftershock productivity function, with m_0 being the magnitude of completeness or the minimum magnitude considered in the earthquake catalog, and the parameter β controls the effect of magnitude on the generation of aftershocks.

The forecast catalog used in Section 4 comes from the ETAS model formulation of Zhuang (2011), which assumes that any magnitude m that is greater than the magnitude of completeness m_0 has a rate of

$$\lambda(t, s, m|\mathcal{F}_t) = \pi(m)\left(\mu(s) + \sum_{(t_i,s_i,m_i)\in\mathcal{F}_t} \kappa(m_i) g(t - t_i) f(s - s_i, m_i)\right), \tag{3}$$

with $s = (x, y)$ the latitude and longitude spatial location, and $\pi(m)$ the magnitude density. Here, $\mu(s)$ is a spatially varying background rate that is assumed not to change over time, and $\kappa(m_i) = Ae^{\alpha(m_i-m_0)}$ for $m_i \geq m_0$, is the aftershock productivity function. The function $g(t - t_i)$ is the probability density function of the length of time between an aftershock and its parent, and the spatial triggering function $f(s - s_i; m_i)$ describes the spatial locations of aftershocks with scaling for magnitude, thus allowing wider spatial areas for higher magnitudes. The function $g(t - t_i)$ can be viewed as the intensity function of an inhomogeneous Poisson process describing the temporal locations (t) of aftershocks of a given earthquake occurring at time t_i, and similarly, $f(s - s_i; m_i)$ governs the aftershock locations given a parent at location s_i with magnitude m_i. In other words, ETAS can be seen as a superposition of various inhomogeneous Poisson processes, with one component for the background events, and another component representing the aftershocks.

2.1.2 | STEP

The STEP model, first described by Gerstenberger et al. (2005), is an alternative model to ETAS. The STEP model shares components with a model called the Reasenberg–Jones model (Reasenberg and Jones 1989), which in turn combines both the Gutenberg–Richter (Gutenberg and Richter 1944) relationship and the Omori–Utsu law (Omori 1894; Utsu 1961) to define the earthquake intensity rate as

$$\lambda(t, m) = 10^{a-b(m-m_p)} \cdot (t + c)^{-p} \tag{4}$$

with parent event of magnitude m_p.

STEP has two main components at play to model the spatial and clustering elements of aftershocks. It uses the Reasenberg–Jones model to describe the aftershock rate with an added component that models the long-term background seismicity with a spatial component. The second component is a clustering model with its complexity depending on the amount of aftershock data currently available (Gerstenberger et al. 2005). This clustering model is very active and localized during strong earthquake sequences, which leads to STEP having a narrow spatial distribution in the days after earthquake events. In ETAS, all events are considered capable of triggering aftershocks, while in STEP, only the largest events have aftershocks. Thus, instead of combining the background and all aftershock components by taking a sum as in ETAS, STEP essentially only takes the maximum of these aftershock components.

2.2 | Log-Likelihood Score

The log-likelihood (L) (Daley and Vere-Jones 2003) score is often used to evaluate earthquake forecasting models, such as those considered here in Section 4, as an overall measure of fit of the model to all the observed events. It is defined as

$$L = \sum_{i=1}^{n} \log \lambda_i - \int \lambda(t, x, y, m) \mathrm{d}t\mathrm{d}x\mathrm{d}y\mathrm{d}m,$$

given the set of observations (t_i, x_i, y_i, m_i), $i = 1, \ldots, n$ for all n observed events, with λ_i the mean intensity rate in the bin containing observation i. Here, $\lambda(t, x, y, m)$ denotes the model's forecast conditional intensity.

A proper scoring rule for model forecasting is one where the true model maximizes the expected score (Gneiting and Raftery 2007). In other words, no model cannot achieve a higher expected score than the true model. The Q-score is not proper because rather than rewarding accuracy, Q rewards a model for forecasting a high rate where the largest events occur, relative to the other events. If model 0 is correct, and model 1 tends to overestimate the rate in spatial–temporal locations where the largest events occur, relative to where the bulk of the events occur, then the expected Q-score for model 1 will exceed that of model 0, violating properness. Despite its lack of accuracy, such a model 1 would be extremely useful in practice. While the Q-score may be useful in highlighting models that forecast higher rates where the largest events occur, the Q-score is not a proper scoring rule and should not be used independently from other evaluation scores for a comprehensive evaluation of goodness of fit. In contrast, the L-score is a proper score. However, even though the L-score gives a general goodness-of-fit score between the model's forecast and the observed events, it does not explicitly weight large magnitude events more than the other events. Log-likelihood scores essentially reward the model equivalently for accurately forecasting a magnitude 3 or a magnitude 7 event. Events of greater magnitude are more hazardous and thus of much more interest to forecast, and since the Q-score rewards models with higher rates in locations with the highest magnitude observed events, the L- and Q-scores complement each other well and may be used in tandem as forecast model diagnostics.

3 | Q-Score

The Q-score is a recently proposed magnitude-weighted goodness-of-fit score for the α-largest earthquakes, where $\alpha \in [0, 1]$ is the proportion of the largest earthquakes. The Q-score was introduced in Schoenberg and Schorlemmer (2024) and Schoenberg (2025), with a recommendation of $\alpha = 5\%$.

Let $\hat{N}$ be a marked spatial–temporal point process with conditional intensity $\hat{\lambda}(\boldsymbol{\tau}, m)$, where $\boldsymbol{\tau} = (t, x, y)$. Let N denote the ground process, $\lambda_g(\boldsymbol{\tau}) = \int_{\mathcal{M}} \hat{\lambda}(\boldsymbol{\tau}, m)dm$ the conditional intensity integrated over the mark space, and $\pi(m)$ the mark density. Define $M^{[1-\alpha]}$ as the threshold value such that $P(m > M^{[1-\alpha]}) = \alpha$ for $m \in \mathcal{M}$. Now let $A_\alpha = \{\boldsymbol{\tau} : M(\boldsymbol{\tau}) > M^{[1-\alpha]}\}$ be all the space-time locations where the magnitude threshold $M^{[1-\alpha]}$ is exceeded. Then, the Q-score can be defined as

$$Q^\alpha = \frac{\left(\int \lambda_g(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)/\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)}{\left(\int \lambda_g(\boldsymbol{\tau})dN\right)/\left(\int dN\right)}.$$

Having observed data on the times, locations and magnitudes of n events, $\mathbf{z} = \{(\boldsymbol{\tau}_i, m_i)\}_{i=1}^n$, we obtain

$$Q^\alpha = \frac{\left(\sum_{j=1}^n \lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}\right)/\left(\sum_{j=1}^n \mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}\right)}{\sum_{j=1}^n \lambda_g(\boldsymbol{\tau}_j)/n}.$$

Here, $\int dN$ is the number of observed events, n. Let $n_\alpha = \sum_{j=1}^n \mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}$ be the number of observations with magnitude $m > M^{[1-\alpha]}$. This number will vary depending on the point pattern (observed earthquakes). The Q^α-score can be interpreted as the mean of the conditional intensity at the space-time locations of the largest earthquakes, divided by the mean of the conditional intensity over all the observed space-time locations. The Q-score thus represents the extent to which the model forecasts higher rates at spatial–temporal locations where the largest events happen to occur, relative to the spatial–temporal locations where the other events occur.

3.1 | Basic Properties

We now look into some properties of Q^α. Q^α depends critically on α. A special case is A_1, where $\alpha = 1$. In this case the selected subset of interest simply contains all the observations, so that $n_1 = n$. Thus

$$Q^1 = \frac{\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_1\}/n_1}{\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)/n} = \frac{\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)/n}{\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)/n} = 1.$$

In the other extreme, when $\alpha = 0$, Q^0 is undefined since $\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN = 0$. For small datasets and/or small values of α, this can also be an issue in practice, as A_α may contain no observations.

For the case of a homogeneous Poisson process, $Q^\alpha = 1$ for any $\alpha \in [0, 1]$, assuming A_α contains at least one observation. Indeed, for a homogeneous Poisson process with constant intensity ρ,

$$\Lambda(\boldsymbol{\tau}) = \int_{\mathcal{M}} \hat{\lambda}(\boldsymbol{\tau}, m)dm = \rho|\mathcal{M}|.$$

Then Q^α reduces to

$$\begin{aligned} Q^\alpha &= \frac{\sum_{j=1}^n \rho|\mathcal{M}|\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}/n_\alpha}{\sum_{j=1}^n \rho|\mathcal{M}|/n} = \frac{\rho|\mathcal{M}|\sum_{j=1}^n \mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}/n_\alpha}{\rho|\mathcal{M}|\sum_{j=1}^n 1/n} \\ &= \frac{\sum_{j=1}^n \mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}/n_\alpha}{1} = \frac{n_\alpha}{n_\alpha} = 1. \end{aligned}$$

According to Schoenberg and Schorlemmer (2024), Schoenberg (2025), the higher the value of Q^α, the better the model appears to be forecasting the spatial–temporal locations of the largest earthquakes. Note that the Q-score accounts for how much more intensity is assigned to large earthquakes relative to what is usually assigned, but does not give a satisfactory overview of overall model fit. The Q-score may also be misleading when there are very few observations, or when α is very low. When $Q^\alpha > 1$, the model is better than a homogeneous Poisson process at forecasting the largest earthquakes, whereas if $Q^\alpha < 1$, then the model forecasts the largest events worse than a homogeneous Poisson process. Section 4 presents examples of Q-scores for different earthquake models, with $Q^\alpha > 1$ for some models and $Q^\alpha < 1$ for others. It is difficult to calculate the distribution or even the expectation of the Q-score for a given model analytically, but under some assumptions, we can derive some properties for the null case where the model in question is separable in magnitude. Proposition 1 states that in such cases, the expectation of Q^α is 1 in a ratio sense, and Theorem 1 states that Q^α approaches 1 asymptotically.

3.2 | Expectation

The expectation of Q^α is

$$\mathbb{E}[Q^\alpha] = \mathbb{E}\left[\frac{\left(\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}\right)/n_\alpha}{\sum_{j=1}^n \Lambda(\boldsymbol{\tau}_j)/n}\right].$$

It is challenging to calculate the expectation directly, but by assuming separability in the marks, which is true for example, the ETAS model, the expression is simplified. Using this assumption, Proposition 1 shows that the numerator and denominator of Q^α are the same in expectation, but note that this does not necessarily mean that $\mathbb{E}[Q^\alpha] = 1$.

Proposition 1. *Assuming separability in the marks, that is,* $\hat{\lambda}(\boldsymbol{\tau}, m) = \lambda(\boldsymbol{\tau})\pi(m)$, *the numerator and denominator of* Q^α *are the same in expectation, which means that*

$$\frac{\mathbb{E}\left[\left(\int \lambda_g(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\left(\int dN\right)\right]}{\mathbb{E}\left[\left(\int \lambda_g(\boldsymbol{\tau})dN\right)\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\right]} = 1.$$

Proof. We first rewrite Q^α as

$$Q^\alpha = \frac{\left(\int \lambda_g(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)/\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)}{\left(\int \lambda_g(\boldsymbol{\tau})dN\right)/\left(\int dN\right)}$$
$$= \frac{\left(\int \lambda_g(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\left(\int dN\right)}{\left(\int \lambda_g(\boldsymbol{\tau})dN\right)\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)}.$$

Then, for the expectation of the denominator, we get

$$\mathbb{E}\left[\left(\int \lambda_g(\boldsymbol{\tau})dN\right)\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\right] = \mathbb{E}\left[\left(\sum_{j=1}^{n}\lambda_g(\boldsymbol{\tau}_j)\right)\left(\sum_{l=1}^{n}\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right)\right]$$
$$= \mathbb{E}\left[\sum_{j=1}^{n}\sum_{l=1}^{n}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right].$$

Using the law of total expectation, we obtain

$$\mathbb{E}\left[\sum_{j=1}^{n}\sum_{l=1}^{n}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right] = \sum_{k=0}^{\infty}P(n=k)\mathbb{E}\left[\sum_{j=1}^{k}\sum_{l=1}^{k}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right]$$
$$= \sum_{k=0}^{\infty}\sum_{j=1}^{k}\sum_{l=1}^{k}P(n=k)\mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right].$$

Now, looking only at the expectation term, one obtains

$$\mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_l \in A_\alpha\}\right] = \mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)|\{\boldsymbol{\tau}_l \in A_\alpha\}\right]P\left(\{\boldsymbol{\tau}_l \in A_\alpha\}\right)$$
$$= \mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)|m_l > M^{[1-\alpha]}\right]P\left(m_l > M^{[1-\alpha]}\right) = \mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)\right]P\left(m_l > M^{[1-\alpha]}\right).$$

Similarly, for the expectation of the numerator, we obtain

$$\mathbb{E}\left[\left(\int \lambda_g(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\left(\int dN\right)\right] = \mathbb{E}\left[\left(\sum_{j=1}^{n}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}\right)n\right]$$
$$= \mathbb{E}\left[\sum_{j=1}^{n}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}n\right] = \sum_{k=0}^{\infty}P(n=k)\mathbb{E}\left[\sum_{j=1}^{k}\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}k\right]$$
$$= \sum_{k=0}^{\infty}P(n=k)k\sum_{j=1}^{k}\mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)\mathbf{1}\{\boldsymbol{\tau}_j \in A_\alpha\}\right]$$
$$= \sum_{k=0}^{\infty}P(n=k)k\sum_{j=1}^{k}\mathbb{E}\left[\lambda_g(\boldsymbol{\tau}_j)\right]P\left(m_j > M^{[1-\alpha]}\right).$$

When a process is separable in the marks,

$$\lambda_g(\boldsymbol{\tau}) = \int_{\mathcal{M}}\hat{\lambda}(\boldsymbol{\tau},m)dm = \int_{\mathcal{M}}\pi(m)\lambda(\boldsymbol{\tau})dm = \lambda(\boldsymbol{\tau})\int_{\mathcal{M}}\pi(m)dm = \lambda(\boldsymbol{\tau}).$$

Thus, Q^α reduces to

$$Q^\alpha = \frac{\left(\int \lambda(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\left(\int dN\right)}{\left(\int \lambda(\boldsymbol{\tau})dN\right)\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)}.$$

Now, the expectation of the numerator is

$$\mathbb{E}\left[\left(\int \lambda(\boldsymbol{\tau})\mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\left(\int dN\right)\right]$$
$$= \sum_{k=0}^{\infty}P(n=k)\sum_{j=1}^{k}k\mathbb{E}\left[\lambda(\boldsymbol{\tau}_j)\right]P\left(m_j > M^{[1-\alpha]}\right),$$

and the expectation of the denominator is

$$\mathbb{E}\left[\left(\int \lambda(\boldsymbol{\tau})dN\right)\left(\int \mathbf{1}\{\boldsymbol{\tau} \in A_\alpha\}dN\right)\right]$$
$$= \sum_{k=0}^{\infty}P(n=k)\sum_{j=1}^{k}\sum_{l=1}^{k}\mathbb{E}\left[\lambda(\boldsymbol{\tau}_j)\right]P\left(m_l > M^{[1-\alpha]}\right).$$

Using the assumption of separability in the marks, we have that $P\left(m_l > M^{[1-\alpha]}\right) = \alpha$ for $l = 1, \ldots, k$.

Therefore, the expectation of the numerator is

$$\sum_{k=0}^{\infty}P(n=k)\sum_{j=1}^{k}k\mathbb{E}\left[\lambda(\boldsymbol{\tau}_j)\right]\alpha,$$

and the expectation of the denominator is

$$\sum_{k=0}^{\infty}P(n=k)\sum_{j=1}^{k}\sum_{l=1}^{k}\mathbb{E}\left[\lambda(\boldsymbol{\tau}_j)\right]\alpha = \sum_{k=0}^{\infty}P(n=k)\sum_{j=1}^{k}k\mathbb{E}\left[\lambda(\boldsymbol{\tau}_j)\right]\alpha.$$

Thus, the expectation of the numerator and the expectation of the denominator are equal. □

3.3 | Law of Large Numbers

In Theorem 1, we provide a law of large numbers for Q^α for a marked temporal point process $\check{N}$. In our context, $\check{N}$ is our marked temporal point process recording time t and magnitude m. An the end of the proof, we explain how to extend the proof for a marked spatial–temporal process.

Theorem 1 states that asymptotically, when time $T \to \infty$, Q_T^α, defined in Equation (5) approaches 1 almost surely and in L_1 norm. This result is stronger than the result in Proposition 1, which says that the ratio of the expectation of the numerator and the expectation of the denominator is 1. But Theorem 1 also requires more assumptions. In Proposition 1, it is enough to assume separability in the marks, while in Theorem 1 we need additional assumptions.

The first additional assumption is stationarity and metric transitivity of the ground process N. We also assume that the weighted random measure Ψ with $\Psi((0,T]) = \int_0^T \lambda_g(t)dN$ is stationary and metrically transitive. Note here that Ψ is a random measure but not necessarily a point process, since the point counts are scaled with the aggregated conditional intensity $\lambda_g(t)$. For stationary random measures, metric transitivity is equivalent to ergodicity (Daley and Vere-Jones 2008, Proposition 12.3.III). We refer to Daley and Vere-Jones (2008) for definitions of stationarity, metric transitivity and ergodicity. We use the definition for stationarity from (Daley and Vere-Jones 2008, Definition 12.1.II), and ergodicity from (Daley and Vere-Jones 2008, Definition 12.3.I(i)).

We further assume that both N and Ψ have finite mean density. For a stationary random measure ξ, the mean measure $\mathbb{E}[\xi(A)]$ is translation invariant, which means it only depends on the size of the set A and not on shifts in the space $\mathcal{X}$. If ξ has a finite mean density m, the mean measure can be written as $ml(A)$ where $0 \leq m < \infty$ and l is the Lebesgue measure.

Theorem 1. *Let $\check{N}$ be a marked temporal point process on $\mathbb{R}^+ \times \mathcal{M}$, with conditional intensity $\check{\lambda}(t,m)$ and ground intensity $\lambda_g(t) = \int_{\mathcal{M}}\check{\lambda}(t,m)dm$, for some c.s.m.s. $\mathcal{M}$. We also assume that $\check{N}$ is separable in the marks, that is $\check{\lambda}(t,m) = \lambda(t)\pi(m)$, where $\lambda(t)$ is the conditional intensity for N and $\pi(m)$ is the mark density. We assume that the ground process N is stationary and metrically transitive* *with finite mean density b. Now, let Ψ be the weighted random measure with $\Psi((0,T]) = \int_0^T \lambda_g(t)dN$, where $N((0,T]) = \int_0^T dN$. Further, assume that Ψ is also stationary and metrically transitive, with finite mean density $\bar{\lambda}b$. Let*

$$Q_T^\alpha = \frac{\left(\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN\right)/\left(\int_0^T \mathbf{1}\{t \in A_\alpha\}dN\right)}{\left(\int_0^T \lambda_g(t)dN\right)/\left(\int_0^T dN\right)}. \quad (5)$$

Then $\lim_{T\to\infty} Q_T^\alpha = 1$ almost surely and in L_1 norm.

Proof To simplify $\lim_{T\to\infty} Q_T^\alpha$ we want to use Slutsky's lemma (Ferguson 1996, Theorem 6) twice to obtain

$$\lim_{T\to\infty} Q_T^\alpha = \frac{\left(\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN\right)/\left(\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN\right)}{\left(\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)dN\right)/\left(\lim_{T\to\infty}\frac{1}{T}\int_0^T dN\right)}.$$

To be able to use Slutsky's lemma, we first have to verify that all denominator terms are constant, that is,

$$\lim_{T\to\infty}\frac{\frac{1}{T}\int_0^T \lambda_g(t)dN}{\frac{1}{T}\int_0^T dN} = c \quad \lim_{T\to\infty}\frac{1}{T}\int_0^T dN = c_1,$$
$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN = c_2,$$

where c, c_1, c_2 are constants. Note that we also require that $\lim_{T\to\infty}\frac{1}{T}\int_0^T \Lambda(t)dN$ is constant, since both $\lim_{T\to\infty}$ $\left(\frac{1}{T}\int_0^T \lambda_g(t)dN\right)/\left(\frac{1}{T}\int_0^T dN\right)$ and $\lim_{T\to\infty}\frac{1}{T}\int_0^T dN$ are constant.

We will now verify that all these terms are constant, starting with the term $\lim_{T\to\infty}\frac{1}{T}\int_0^T dN$. We want to use (Daley and Vere-Jones 2008, Corollary 12.2.V(a)), which is a corollary of (Daley and Vere-Jones 2008, Theorem 12.2.IV), so first we must verify that the assumptions hold.

Corollary (Corollary 12.2.V(a) from (Daley and Vere-Jones 2008)). *Let ξ be a stationary and metrically transitive random measure on $\mathbb{R}^d$ with finite mean density m. Then, for any convex averaging sequence $\{A_n\}$ on $\mathbb{R}^d$ it holds that $\xi(A_n)/l(A_n) \to m$ almost surely and in L1 norm.*

Let $A_T = (0,T]$, then it is easy to verify that $\{A_T\}$ is a convex averaging sequence on $\mathbb{R}^+$, see (Daley and Vere-Jones 2008, Definition 12.2.I) for the definition of a convex averaging sequence. It holds that $l(A_T) = T$, where l is the Lebesgue measure. We have assumed that N is a temporal point process which is stationary and metrically transitive with finite mean density b. Then, Daley and Vere-Jones (2008, Corollary 12.2.V(a)) gives us that

$$\frac{1}{T}\int_0^T dN = \frac{N(A_T)}{l(A_T)} \xrightarrow[T\to\infty]{} b$$

almost surely and in L_1 norm. Note that the limit is a constant, as required.

Let us now consider the term $\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)dN$, that is $\lim_{T\to\infty}\frac{1}{T}\Psi(A_T)$. Since we have assumed that Ψ is stationary and metrically transitive with finite mean density $\bar{\lambda}b$, Daley and Vere-Jones (2008, Corollary 12.2.V(a)) gives us that

$$\frac{1}{T}\int_0^T \lambda_g(t)dN = \frac{\Psi(A_T)}{l(A_T)} \xrightarrow[T\to\infty]{} \bar{\lambda}b$$

almost surely and in L_1 norm. We thus obtain

$$\frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)dN}{\lim_{T\to\infty}\frac{1}{T}\int_0^T dN} \xrightarrow[T\to\infty]{} \frac{\bar{\lambda}b}{b} = \bar{\lambda}$$

almost surely and in L_1 norm. The limit is a constant, as required.

Let us now consider

$$\frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN}{\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN}.$$

We want to show that this ratio converges to $\bar{\lambda}$, so that Q_T^α converges to 1.

We start with the term $\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN$. Due to the separability we have

$$\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN = \frac{1}{T}\sum_{t\in N\cap A_T}\mathbf{1}\{t \in A_\alpha\}$$
$$= \frac{1}{T}\sum_{(t,m)\in\breve{N}\cap A_T\times\mathcal{M}}\mathbf{1}\{m > M^{[1-\alpha]}\} = \frac{1}{T}\alpha N(A_T) = \alpha\frac{N(A_T)}{l(A_T)},$$

since the probability of having a mark larger than $M^{[1-\alpha]}$ is independent of the time t. Then, we can use the previous result to obtain

$$\frac{1}{T}\int_0^T \mathbf{1}\{t \in A_\alpha\}dN = \alpha\frac{N(A_T)}{l(A_T)} \xrightarrow[T\to\infty]{} \alpha b$$

almost surely and in L_1 norm. Here, the limit is also a constant, as required.

In order to complete the proof, we need only verify that the term $\frac{1}{T}\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN$ converges to $\alpha\bar{\lambda}b$, as we have then shown that $\lim_{T\to\infty} Q_T^\alpha = 1$ almost surely and in L_1 norm.

By separability in the marks, we have

$$\frac{1}{T}\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN = \frac{1}{T}\int_0^T\int_{\mathcal{M}}\check{\lambda}(t,m)dm\mathbf{1}\{t \in A_\alpha\}dN$$
$$= \frac{1}{T}\int_0^T\int_{\mathcal{M}}\check{\lambda}(t,m)\mathbf{1}\{t \in A_\alpha\}dmdN = \frac{1}{T}\int_0^T\int_{\mathcal{M}}\lambda(t)\pi(m)\mathbf{1}\{t \in A_\alpha\}dmdN$$
$$= \frac{1}{T}\int_0^T\int_{\mathcal{M}}\lambda(t)\pi(m)\mathbf{1}\{m > M^{[1-\alpha]}\}dmdN$$
$$= \frac{1}{T}\int_0^T\lambda(t)\int_{\mathcal{M}}\pi(m)\mathbf{1}\{m > M^{[1-\alpha]}\}dmdN$$
$$= \frac{1}{T}\int_0^T\lambda(t)\int_{M^{[1-\alpha]}}^\infty\pi(m)dmdN = \frac{1}{T}\int_0^T\lambda(t)P\left(M > M^{[1-\alpha]}\right)dN$$
$$= \alpha\frac{1}{T}\int_0^T\lambda(t)dN = \alpha\frac{1}{T}\int_0^T\Lambda(t)dN,$$

which implies that

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \lambda_g(t)\mathbf{1}\{t \in A_\alpha\}dN = \alpha\frac{1}{T}\int_0^T \lambda_g(t)dN \xrightarrow[T\to\infty]{} \alpha\bar{\lambda}b$$

almost surely and in L_1 norm, using the previous results.

To extend the proof for a spatial–temporal point process on $\mathbb{R}^+ \times \mathcal{S}$ where $\mathcal{S} \subset \mathbb{R}^2$ is bounded, consider the convex averaging sequence $A_T = (0, T] \times \mathcal{S}$. Here, $l(A_T) = T|\mathcal{S}|$, and the $1/|\mathcal{S}|$ term is an additional factor in all four terms in Q_T^{α}. This factor can be seen as a constant, as it does not vary when $T \to \infty$, and it is canceled out in both the ratios in the numerator and the denominator. Similarly, $N(A_T)$, and the other terms are scaled up with an additional factor, which does not vary when $T \to \infty$. □

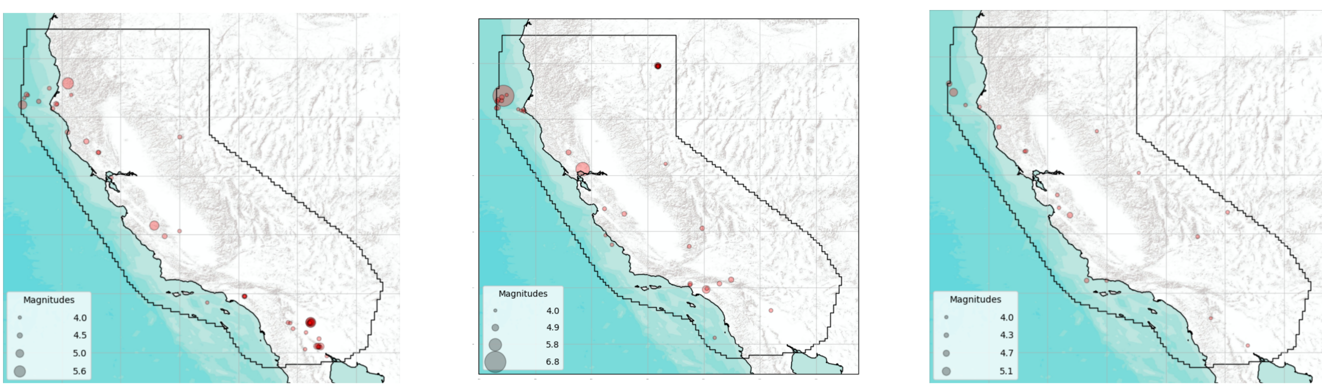
4 | Application to California Seismicity

The Collaboration for the Study of Earthquake Predictability (CSEP) provides access to next-day forecasts for California from 21 different models (Serafini, Bayona, Silva, et al. 2025), for which we calculated their Q- and L-scores to obtain an evaluation of their fit. An In-depth description of the models and the experiment conducted to generate their forecasts can be found in Graham et al. (2024). In this section, we present the annual Q-scores and L-scores of forecasts available during the years 2012, 2014, and 2017. The forecast catalogs for the models have different start and end run dates; therefore, we chose years that would allow us to evaluate the Q-scores and L-scores of most models provided by CSEP. Forecast catalogs cover different time periods lasting from August 1, 2007 to August 2018 (Serafini, Bayona, Silva, et al. 2025). All model forecast catalogs, as well as the observed catalog, have a minimum magnitude of 3.95. Figure 1 shows the observed event locations for those years, with point sizes proportional to their magnitudes. In 2012, there was the Brawley earthquake swarm; in 2014, there was the magnitude 6.0 South Napa earthquake (Brocher et al. 2015; Wei et al. 2015); and in 2017, the last full annual forecast was available.

In 2012, the three largest events had magnitudes 5.6, 5.41, and 5.32. In 2014, the three largest events had magnitudes 6.8, 6.02, and 5.1. In 2017, when only 18 events were recorded in California with magnitude of at least 3.95, only one event qualifies as a top 5% event, and it has magnitude 5.08. As a reference, Tables A1–A3 in the Appendix contain all Q- and L-scores calculated annually for all models depending on the forecast time period available, ranging from 2008 to 2017. As seen in Figure 2a,b, STEPJAVA had the highest Q-score in 2012 and the second-highest Q-score among all CSEP models in 2014. However, as seen in Figure 2c, STEPJAVA had a score of zero for the year 2017 since the forecast had a rate of zero on the space-magnitude bin where the only top 5% event occurred.

As seen in Figure 2, models with the highest Q-scores also typically had relatively high L-scores. However, the converse was not generally true, for example, the forecasts for the year 2014. During this year, the top observed event magnitude was 6.8, corresponding to the second-largest magnitude event documented within the observed catalog between 2008 and 2017. The four models that had the highest L-scores had Q-scores under 1, as seen in Table 1, indicating these models did not accurately forecast the locations of the largest events, even though they fit the rest of the observed events. However, models that had the highest Q-scores in 2014 also had relatively high L-scores, and thus offered relatively good fits to the observed catalog. As a side note, the highest magnitude event observed in the catalog was in the year 2010, with a magnitude of 7.2. In Table A1 in the Appendix, the model with the highest Q-score (STEP) for 2010 also had the highest L-score, while in this same year the models ETAS and KJSSOneDay had Q-scores near 1, which is the expected score for a homogeneous Poisson process.

One particularly exceptional case for these scores was STEPJAVA, whose log-likelihood diverges for all years. In each of the 3 years considered, STEPJAVA provided a forecast rate of zero for at least one bin where events actually occurred, which should theoretically never happen for a suitable forecast model. The STEPJAVA forecast seems to provide zero rates for spatial bins from the Northern California coastal area, as seen in the Appendix, Figure A2, where at least one 3.95 or larger magnitude earthquake typically occurs annually. This is compared to Figure A1, where other models provided a rate above zero for the same region. A possible solution to the problem of earthquakes occurring in areas with zero forecasted rate is the use of a "waterlevel," which is a minimum rate to be added to the zero rate bins. The selected value of the water level highly influences the L-score of the model's forecast, which makes it not an ideal solution. For instance, the L-score with the water level of 10^{-6} for STEPJAVA in the year 2012 was −296.76, for 2014 it was −351.91, and for 2017 it was −136.89. These were the values used in Figure 2. Comparing these values to the rest of the L-scores in the figure, we see that STEPJAVA is still not one of the best models in terms of



(a) (b) (c)

FIGURE 1 | ComCat catalog events above magnitude 3.95 in California. (a) 50 events in the year 2012, with a 5.6 magnitude of the largest event. (b) 44 events in the year 2014, with a 6.8 magnitude of the largest event. (c) 18 events in the year 2017, with a 5.08 magnitude of the largest event.

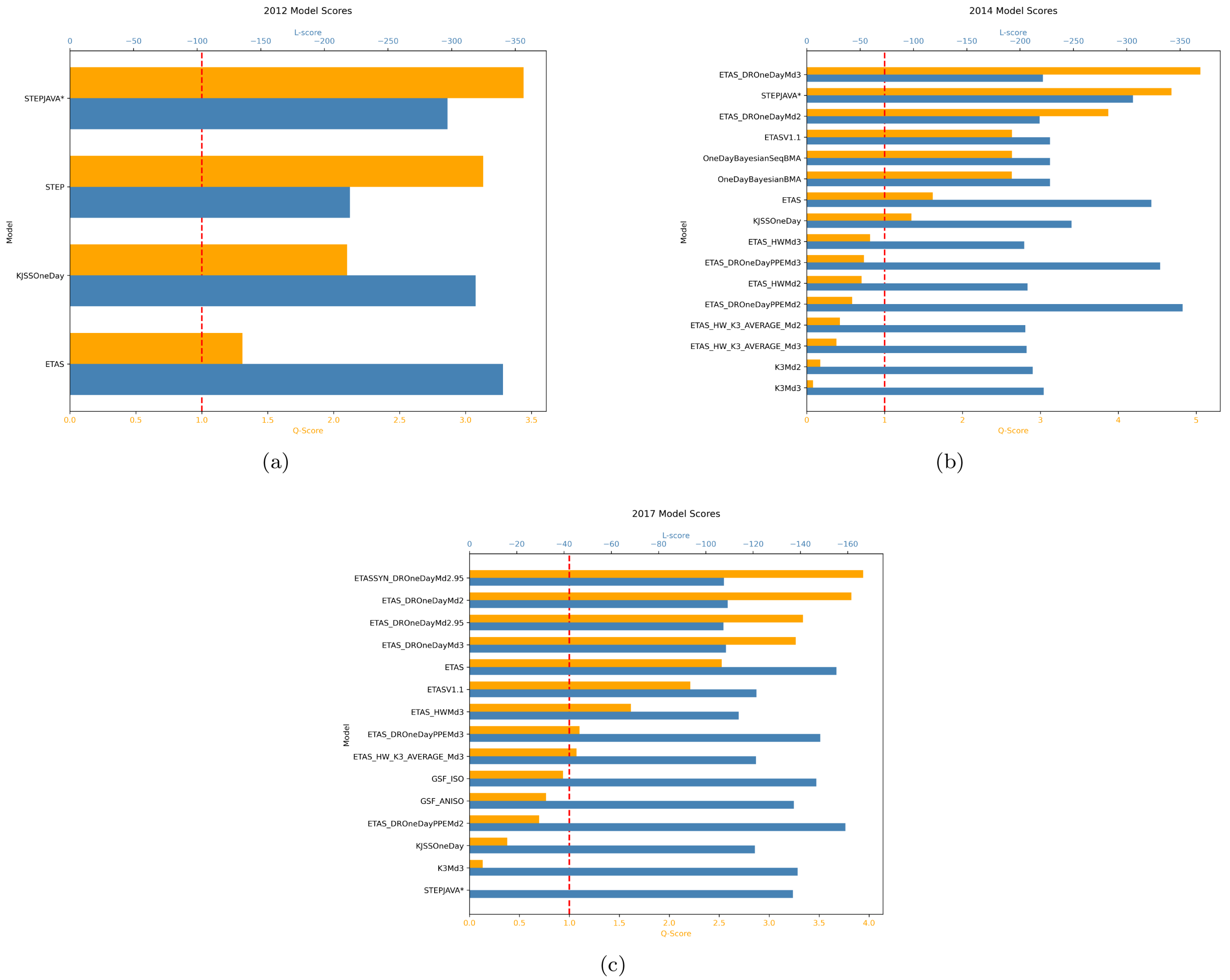


FIGURE 2 | Q-scores and L-scores of the models for the three different years. Models are in descending order based on the Q-score. The red dashed line indicates a Q-score of 1. The purple bars are the model's Q-score, and the orange bar its corresponding L-score. STEPJAVA* uses a water level to calculate the L-score. In reality, all loglikelihood scores for STEPJAVA diverge to $-\infty$. Models with a Q-score above 1, and an L-score close to 0 should be considered good fits to the observed data. Forecast model data was obtained from https://zenodo.org/records/15076187.

TABLE 1 | Top 4 Models with the highest L-scores in the year 2014.

Model	Q-score	L-score
ETAS_HWMd3	0.82	−161.50
ETAS_HWMd2	0.71	−164.51
ETAS_HW_K3_AVERAGE_Md3	0.38	−168.90
ETAS_HW_K3_AVERAGE_Md2	0.43	−169.72

L-score even with the water level correction, though it had one of the highest Q-scores in two of these years.

Even though STEPJAVA is the model with the highest Q-score in some of the time periods, it is the only model that we analyzed for which the L-score diverges every year. The STEP and STEPJAVA models, as seen in Appendix Figure A3, seem to forecast more magnitude 6 earthquakes than the rest of the models. However, STEP does seem to be a good fit to the observed data, evidenced by the relatively high L-scores, while also having a high Q-score. Conducting a more thorough evaluation using other scores in addition to the Q and L-scores needs to be conducted in order to explore if the magnitude distribution of these models is affecting these scores.

Aside from STEP and STEPJAVA, models that had high Q-scores for these three years were the ETAS_DROneDayMd models, which account for spatial inhomogeneity by implementing a spatial–temporal varying background rate through the Proximity to Past Earthquakes (PPE) model (Serafini, Bayona, Silva, et al. 2025). These ETAS_DROneDayMd models also had relatively high log-likelihood scores. In fact, the ETAS_DROneDayMd3 model had a Q-score of 5.06 for the year 2014, which is the highest Q-score observed. This means the ETAS_DROneDayMd3 model assigned forecasted rates at the largest magnitude earthquakes that were 5.06 times higher than the forecasted rates for all events in general. During the year 2014, there was a series of earthquakes off the Northern California coast that occurred on the same day, March 10. Four earthquakes of magnitude above 4 occurred on this day, with a magnitude 6.8 happening first, followed by three aftershocks within the hour. Another top 5% event of magnitude 5.1 also had a large aftershock within 14 hours. The use of the PPE model could be a factor in these ETAS models having such high Q-scores.

5 | Discussion

This article focuses on the Q-score, a recently proposed magnitude-weighted goodness-of-fit score to evaluate earthquake forecasts. The Q-score was first introduced for the 5% largest earthquakes in Schoenberg and Schorlemmer (2024) and Schoenberg (2025). In Section 3, the Q-score is defined more generally, allowing the proportion α of the largest earthquakes to vary, with $\alpha \in [0, 1]$. Proposition 1 states that the expectation of Q^α is 1 in a ratio sense, under the assumption of separability in the marks. Under additional assumptions, such as stationarity, Theorem 1 states that Q^α approaches 1 almost surely and in L1 norm, as time approaches infinity. These assumptions help form the conditions for a null hypothesis; in practice, one may set the null hypothesis that $Q = 1$, and evaluate if the value of Q for a given model and a given dataset exceeds 1 by a statistically significant margin. Since a homogeneous Poisson process has $Q = 1$, and for magnitude-separable models such as ETAS one expects Q to be approximately 1, such models form a natural basis of comparison for other models.

The interpretation of Proposition 1 is that the expectation of Q^α will tend to be close to 1, since the expectation of Q^α is 1 in a ratio sense, and the required assumption for Proposition 1 is separability in the marks, which holds for many of the models in CSEP, such as the ETAS model. Further, the interpretation of Theorem 1 is that Q^α will tend to be close to 1 for separable and stationary models when we look over long periods of time. However, the real purpose of the Q-score is to allow future models that forecast the largest events better than ETAS to properly distinguish themselves, models where the magnitude distribution is dependent on the spatial and/or temporal distribution. In such cases the expression for $E[Q^\alpha]$ will be less tractable, and Q will hopefully be much greater than 1. Another relevant aspect here is that we now know something about the expectation of Q^α but not about its variability. The issue of having very high variance of potency-weighted metrics is discussed in (Schoenberg 2025), so future work should focus on calculating or approximating the variance of Q^α, though this will probably lead to very complicated expressions.

In the application provided, many of the earthquake forecast models were some version of ETAS, or some combination of ETAS models with other models. Such models often had relatively high Q-scores and L-scores. However, there is still room for improvement when it comes to forecasting the largest events. In fact, some of the models with high L-scores, and thus good overall fit to the observed catalog, had Q-scores under 1. It would be interesting to analyze more time periods with high magnitude events to get a better understanding of the Q-score's performance in terms of evaluating earthquake forecasts. While Q-scores and L-scores can be used in tandem to identify models with good overall fit and good ability to forecast the largest events, care should be taken especially when models forecast zero rates in certain bins where events occurred. We recommend using other goodness-of-fit evaluation methods in addition to the Q-score.

Acknowledgments

The main part of this work was done during the visit of Julia Jansson Valter at the Department of Statistics and Data Science at UCLA. This research visit was supported by The Foundation Blanceflor and Adlerbertska Forskningsstiftelsen. Research supported by Southern California Earthquake Center Fund number 25166.

The data supporting the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.15076187. The earthquake catalogs and gridded forecast tools used in the calculations of the Q and L-scores are found in the Python package pyCSEP (Savran et al. 2022; Graham et al. 2024). Code repository for the evaluation results can be found on GitHub at: https://github.com/a-arjon/Q-score-a-magnitude-weighted-goodness-of-fit-score-for-earthquake-forecasting.

Funding

This work was supported by the Southern California Earthquake Center (Grant No. 25166).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in pyCSEP and at https://github.com/a-arjon/Q-score-a-magnitude-weighted-goodness-of-fit-score-for-earthquake-forecasting. These data were derived from the following resources available in the public domain:—Graham et al. (2024), https://github.com/a-arjon/Q-score-a-magnitude-weighted-goodness-of-fit-score-for-earthquake-forecasting.

Endnotes

[1]https://zenodo.org/records/15076187.

References

Brocher, T. M., A. S. Baltay, J. L. Hardebeck, et al. 2015. "The M_w 6.0 24 August 2014 South Napa Earthquake." *Seismological Research Letters* 86, no. 2A: 309–326. https://doi.org/10.1785/0220150004.

Daley, D. J., and D. Vere-Jones. 2003. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. 2nd ed. Springer-Verlag. https://doi.org/10.1007/b97277.

Daley, D. J., and D. Vere-Jones. 2008. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. 2nd ed. Springer-Verlag. https://doi.org/10.1007/978-0-387-49835-5.

Davidsen, J., C. Gu, and M. Baiesi. 2015. "Generalized Omori–Utsu Law for Aftershock Sequences in Southern California." *Geophysical Journal International* 201, no. 2: 965–978. https://doi.org/10.1093/gji/ggv061.

Ferguson, T. S. 1996. *A Course in Large Sample Theory*. Chapman & Hall. https://doi.org/10.1201/9781315136288.

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. A. Reasenberg. 2005. "Real-Time Forecasts of Tomorrow's Earthquakes in California." *Nature* 435, no. 7040: 328–331. https://doi.org/10.1038/nature03622.

Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102, no. 477: 359–378. https://doi.org/10.1198/016214506000001437.

Graham, K. M., J. A. Bayona, A. M. Khawaja, et al. 2024. "New Features in the pyCSEP Toolkit for Earthquake Forecast Development and Evaluation." *Seismological Research Letters* 95, no. 6: 3449–3463. https://doi.org/10.1785/0220240197.

Gutenberg, B., and C. F. Richter. 1944. "Frequency of Earthquakes in California." *Bulletin of the Seismological Society of America* 34, no. 4: 185–188. https://doi.org/10.1785/BSSA0340040185.

Hawkes, A. G. 1971. "Spectra of Some Self-Exciting and Mutually Exciting Point Processes." *Biometrika* 58, no. 1: 83–90. https://doi.org/10.1093/biomet/58.1.83.

Michael, A. J., and M. J. Werner. 2018. "Preface to the Focus Section on the Collaboratory for the Study of Earthquake Predictability (CSEP): New Results and Future Directions." *Seismological Research Letters* 89, no. 4: 1226–1228. https://doi.org/10.1785/0220180161.

Ogata, Y. 1988. "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association* 83, no. 401: 9–27. https://doi.org/10.1080/01621459.1988.10478560.

Omori, F. 1894. "On After-Shocks." *Seismological Journal of Japan* 19: 71–80.

Reasenberg, P. A., and L. M. Jones. 1989. "Earthquake Hazard After a Mainshock in California." *Science* 243, no. 4895: 1173–1176. https://doi.org/10.1126/science.243.4895.1173.

Savran, W. H., J. A. Bayona, P. Iturrieta, et al. 2022. "pyCSEP: A Python Toolkit for Earthquake Forecast Developers." *Seismological Society of America* 93, no. 5: 2858–2870. https://doi.org/10.1785/0220220033.

Schoenberg, F. 2025. "Magnitude-Weighted Goodness-Of-Fit Scores for Earthquake Forecasting." *Spatial Statistics* 67: 100895. https://doi.org/10.1016/j.spasta.2025.100895.

Schoenberg, F., and D. Schorlemmer. 2024. "Critical Questions About CSEP, in the Spirit of Dave, Yan, and Ilya." *Seismological Research Letters* 95, no. 6: 3617–3625. https://doi.org/10.1785/0220240213.

Schorlemmer, D., M. J. Werner, W. Marzocchi, et al. 2018. "The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities." *Seismological Research Letters* 89, no. 4: 1305–1313. https://doi.org/10.1785/0220180053.

Serafini, F., J. A. Bayona, F. Silva, et al. 2025. "A benchmark Database of Ten Years of Prospective Next- Day Earthquake Forecasts in California From the Collaboratory for the Study of Earthquake Predictability." *Scientific Data* 12, no. 1: 1501. https://doi.org/10.1038/s41597-025-05766-3.

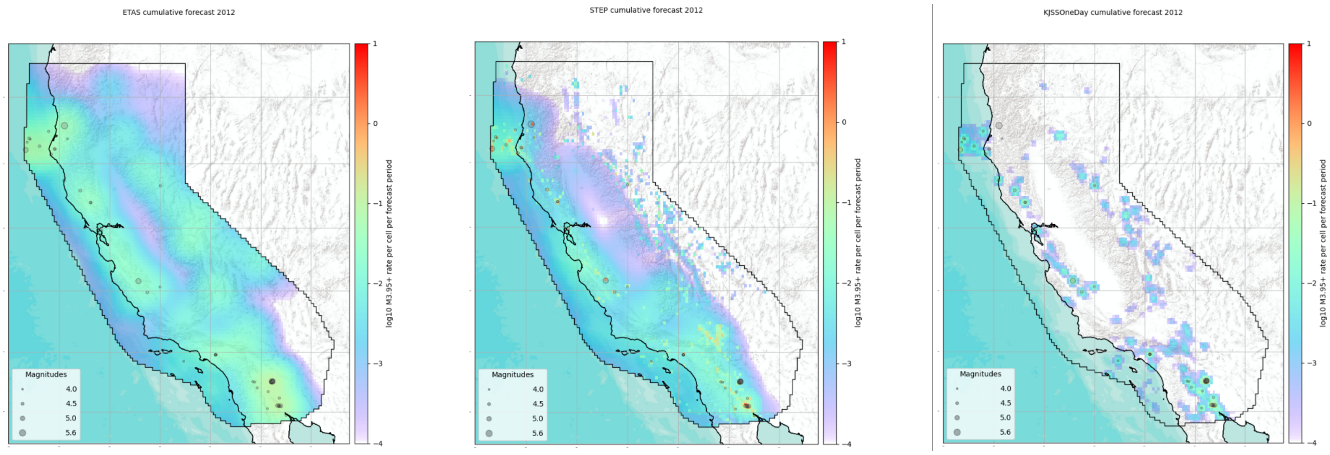
Utsu, T. 1961. "A Statistical Study on the Occurrence of Aftershocks." *Geophysical Magazine* 30: 521–605.

Wei, S., S. Barbot, R. Graves, et al. 2015. "The 2014 M w 6.1 South Napa Earthquake: A Unilateral Rupture With Shallow Asperity and Rapid Afterslip." *Seismological Research Letters* 86, no. 2A: 344–354. https://doi.org/10.1785/0220140249.

Zechar, J. D., D. Schorlemmer, M. Liukis, et al. 2010. "The Collaboratory for the Study of Earthquake Predictability Perspective on Computational Earthquake Science." *Concurrency and Computation: Practice and Experience* 22, no. 12: 1836–1847. https://doi.org/10.1002/cpe.1519.

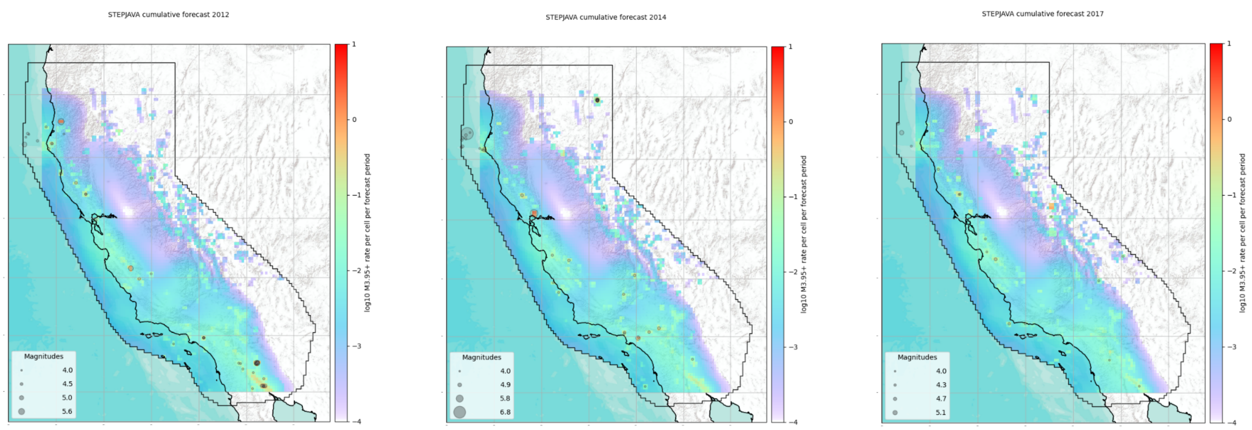
Zhuang, J. 2011. "Next-Day Earthquake Forecasts for the Japan Region Generated by the ETAS Model." *Earth, Planets and Space* 63: 207–216. https://doi.org/10.5047/eps.2010.12.010.

Appendix A



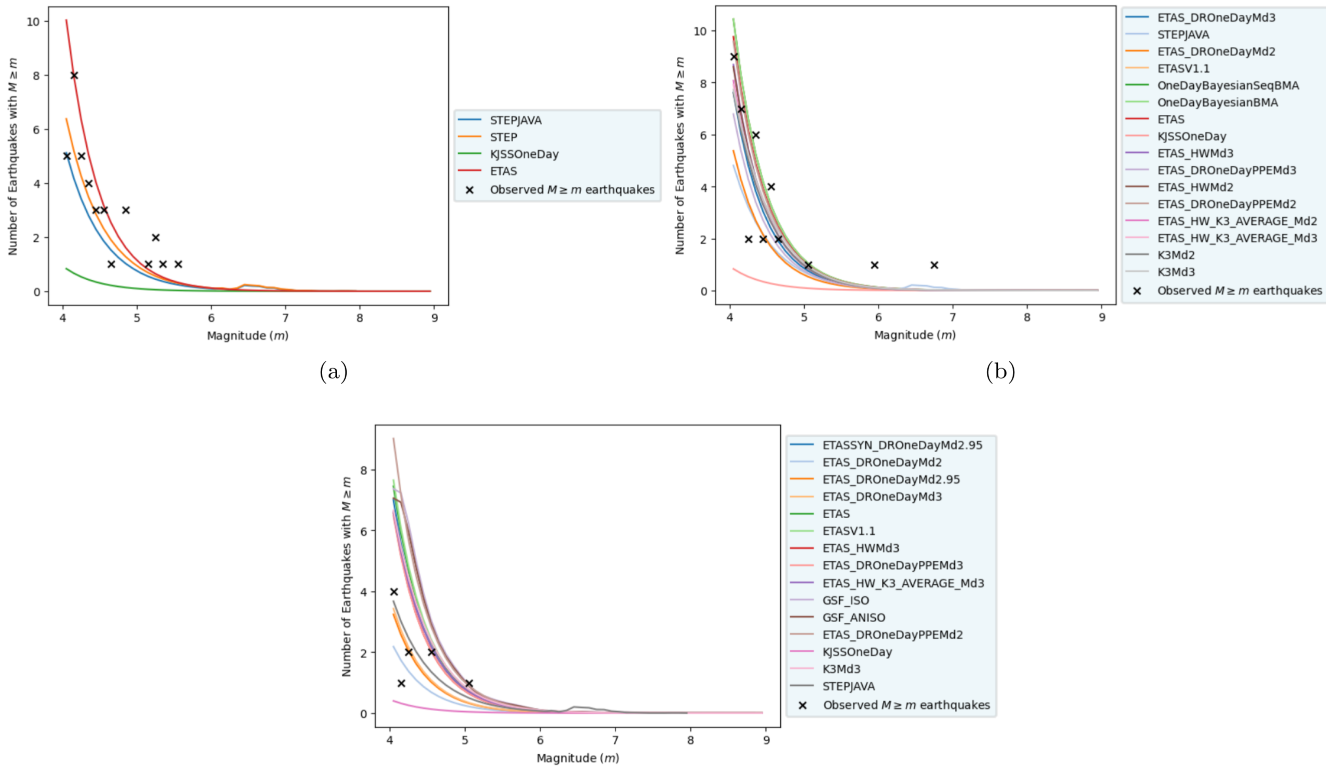
(a) (b) (c)

FIGURE A1 | Plot of the cumulative rate for different models during the year 2012 in California. (a) is for the ETAS model, (b) is for the STEP model, and (c) is for the KJSSOneDay model.



(a) (b) (c)

FIGURE A2 | Plot of the cumulative rate for the STEPJAVA model during the specified time period in California. (a) is for the forecast during the year 2012, (b) is for the year 2014, and (c) is for the year 2017. The days with zero rates in 2012 are (month/day): 01/17, 06/30, 07/21, 08/09, 09/24, and 10/14. In 2014 the days are: 01/23, 03/10, 03/14, 03/16, 03/18, and 12/04. In the year 2017, the days are: 03/06 and 07/29.



(a) (b) (c)

FIGURE A3 | Plot of the annual magnitude distribution for the models available. (a) is for the forecasts during the year 2012, (b) is for the year 2014, and (c) is for the year 2017. The "X" markers indicate the observed number of events.

TABLE A1 | Table shows the Q and log-likelihood scores of the models run during the years 2008 through 2011 (Serafini, Bayona, Silva, et al. 2025).

	2008		2009		2010		2011	
Model	**Q-score**	***L*-score**	**Q-score**	***L*-score**	**Q-score**	***L*-score**	**Q-score**	***L*-score**
ETAS	0.47	−330.78	1.66	−276.55	1.06	−667.65	0.94	−295.47
STEP	2.49	−211.03	1.93	−165.32	4.01	−468.13	1.77	−210.34
STEPJAVA							1.96	-inf (−254.19)
KJSSOneDay			1.58	-inf (−224.68)	1.16	−634.06	0.83	−248.38

Note: Highlighted in green are the highest scores for that year. Models that didn't have a forecast for the year are left blank. Models whose L-score diverges also have an L-score calculated with the water level of 10^{-6} within the parentheses. The last three models are an alternative to ETAS.

TABLE A2 | Table shows the Q and log-likelihood scores of the models for the years 2012 through 2014 (Serafini, Bayona, Silva, et al. 2025).

	2012		2013		2014	
Model	***Q*-score**	***L*-Score**	***Q*-score**	***L*-Score**	***Q*-score**	***L*-Score**
ETAS	1.31	−295.56	1.09	−242.07	1.62	−281.87
ETASV1.1			2.80	−170.21	2.64	−186.91
ETAS_DROneDayMd2			3.33	−193.04	3.87	−177.07
ETAS_DROneDayMd3			3.40	−161.71	5.06	−180.13
ETAS_HWMd2			2.73	−164.77	0.71	−164.51
ETAS_HWMd3			2.92	−164.84	0.82	−161.50
STEP	3.13	−178.90				
STEPJAVA	3.44	-inf (−259.42)	3.33	-inf (−229.90)	4.68	-inf (−278.75)
KJSSOneDay	2.10	−274.86	2.84	−197.22	1.35	−207.67
K3Md2			2.44	−195.63	0.17	−184.08
K3Md3			2.14	−202.26	0.08	−193.63
ETAS_DROneDayPPEMd2			1.03	−253.79	0.58	−311.05
ETAS_DROneDayPPEMd3			0.98	−237.91	0.74	−289.89
ETAS_HW_K3_AVERAGE_Md2			2.61	−174.20	0.43	−169.72
ETAS_HW_K3_AVERAGE_Md3			2.68	−176.59	0.38	−168.90
OneDayBayesianBMA			2.80	−170.22	2.63	−186.92
OneDayBayesianSeqBMA			2.80	−170.22	2.63	−186.92

Note: Highlighted in green are the highest scores for that year. Models that didn't have a forecast for the year are left blank. Models whose L-score diverges also have an L-score calculated with the water level of 10^{-6} within the parentheses. The first 6 models are ETAS models, the next 7 models are an alternative to ETAS, and the last 4 models are ensemble models.

TABLE A3 | Table shows the Q and log-likelihood scores of the models for years 2015 through 2017 (Serafini, Bayona, Silva, et al. 2025).

	2015		2016		2017	
Model	***Q*-score**	***L*-score**	***Q*-score**	***L*-score**	***Q*-score**	***L*-score**
ETAS	1.75	−246.75	1.41	−246.06	2.52	−147.36
ETASV1.1	0.88	−154.13	1.74	−167.68	2.21	−113.58
ETAS_DROneDayMd2	0.83	−136.95	3.15	−158.02	3.82	−101.23
ETAS_DROneDayMd3	0.67	−137.05	1.74	−158.25	3.27	−100.63
ETAS_DROneDayMd2.95					3.34	−99.53
ETASSYN_DROneDayMd2.95					3.94	−99.94
ETAS_HWMd2	0.20	−139.46				
ETAS_HWMd3	0.37	−134.88	1.91	−159.55	1.62	−105.52
GSF_ISO					0.94	−144.62
GSF_ANISO					0.77	−134.85
STEPJAVA	0	-inf (−212.33)	1.90	-inf (−184.97)	0	-inf (−133.12)
KJSSOneDay	0.51	−181.48	2.54	−224.14	0.38	−112.88
K3Md2	0.08	−159.34				
K3Md3	0.05	−161.37	0.66	−206.38	0.13	−134.32
ETAS_DROneDayPPEMd2	1.85	−253.16	0.24	−248.95	0.70	−151.11
ETAS_DROneDayPPEMd3	1.71	−232.92	0.15	−234.25	1.10	−140.44
ETAS_HW_K3_AVERAGE_Md2	0.14	−146.63				
ETAS_HW_K3_AVERAGE_Md3	0.20	−143.15	1.56	−172.86	1.07	−113.54
OneDayBayesianBMA	0.88	−154.13				
OneDayBayesianSeqBMA	0.88	−154.13				

Note: The first 10 models are ETAS models, the next 6 models are an alternative to ETAS, and the last 4 models are ensemble models.