Annals of the Institute of Statistical Mathematics (2024) 76:47–71
https://doi.org/10.1007/s10463-023-00875-5

Comparative evaluation of point process forecasts

Jonas R. Brehmer[1] · Tilmann Gneiting[1,2] · Marcus Herrmann[3] · Warner Marzocchi[3] · Martin Schlather[4] · Kirstin Strokorb[5]

Received: 4 May 2022 / Revised: 13 April 2023 / Accepted: 9 May 2023 /
Published online: 15 June 2023

Abstract
Stochastic models of point patterns in space and time are widely used to issue forecasts or assess risk, and often they affect societally relevant decisions. We adapt the concept of consistent scoring functions and proper scoring rules, which are statistically principled tools for the comparative evaluation of predictive performance, to the point process setting, and place both new and existing methodology in this framework. With reference to earthquake likelihood model testing, we demonstrate that extant techniques apply in much broader contexts than previously thought. In particular, the Poisson log-likelihood can be used for theoretically principled comparative forecast evaluation in terms of cell expectations. We illustrate the approach in a simulation study and in a comparative evaluation of operational earthquake forecasts for Italy.

Keywords Consistent scoring function · Elicitability · Forecast evaluation · Proper scoring rule · Statistical seismology

1 Introduction

In many situations, scientific forecasts of uncertain future quantities provide critical input to societally relevant decision making. For example, criminologists develop methods for forecasts of criminal offences (Mohler et al. 2011; Flaxman et al. 2019; Zhuang and Mateu 2019), epidemiologists assess when and where people catch diseases (Meyer and Held 2014; Schoenberg et al. 2019), and seismologists use statistical models to study and forecast earthquake behaviour (Ogata 1988, 1998; Zhuang et al. 2002; Bray and Schoenberg 2013). The relevant events in these examples—criminal offences, infections, and seismic events—occur as random point patterns in space and time. In probabilistic terms, they are modelled as realizations of point processes (Daley and Vere-Jones 2003). Beyond the development of new point process

Extended author information available on the last page of the article

models for these phenomena, there is a growing demand for theoretically principled evaluation methods.

Model evaluation and forecast assessment are subjects of a vast body of scientific literature. Among a plethora of approaches, a simple distinction can be made between the assessment of absolute and relative performance. Evaluating absolute performance, or assessing goodness-of-fit, means checking whether the assumed model is consistent with the data and rejecting it if this is not the case. If two or more models are available, it is desirable to assess their relative performance and check whether a model outperforms its competitors. Consistent scoring functions and proper scoring rules are widely used and well-studied tools that serve this purpose, see e.g. Gneiting and Raftery (2007) and Gneiting (2011). The central objective of our paper is to demonstrate that this idea and associated statistical methods transfer to point process forecasts and, consequently, provide practical, yet theoretically principled tools for comparative forecast evaluation in this setting.

A scoring function or scoring rule assigns a real number to each pair of a forecast and the respective realized observation of a random variable Y. If the forecast is expressed as a statistical property, such as the mean or a quantile of the (possibly, implicit) predictive distribution, this mapping is called *scoring function*, whereas the term *scoring rule* is used when an entire predictive distribution is reported. In either case, the key requirement to be satisfied is that forecasting the truth yields the best score in expectation: A scoring function is *consistent* for a statistical property if the value of this property for a distribution, F is a minimizer of the expected score with respect to F. Likewise, a scoring rule is *proper* if the expected score with respect to F is minimized by forecasting F. In addition to forecast comparison, propriety and consistency allow for regression and M-estimation (Gneiting and Raftery 2007).

Thus far, statistical seismology has been a driving force in the development of methods to evaluate point process models, see e.g. Bray and Schoenberg (2013) for a review. In particular, the regional earthquake likelihood models (RELMs) initiative (Field 2007) and its successor, the Collaboratory for the Study of Earthquake Predictability (CSEP) (Zechar et al. 2010b; Schorlemmer et al. 2018), have set up forecast experiments for the prospective evaluation of models based on a number of statistical tests. Bray and Schoenberg (2013, p. 518) point out the connection between some of these tests and the scoring literature by stating that "numerical tests, such as the L-test, can be viewed as examples of scoring rules [...]". The paper by Heinrich-Mertsching et al. (2021) makes this connection explicit and derives consistent scoring functions to compare forecasts in the point process setting. We complement their simulation-based approach and develop an alternative, computationally much less intense framework, in which we work with distributional properties for which closed form expressions under the posited point process model are available. This yields a flexible approach to forecast comparison, which incorporates existing methods, and admits new perspectives on the strengths and weaknesses of the CSEP methodology for earthquake forecast evaluation.

The remainder of the paper is structured as follows: Section 2 recalls fundamentals on scoring functions and their role in forecast evaluation and model selection. Section 3 rigorously introduces scoring functions for point patterns and compares to the approach of Heinrich-Mertsching et al. (2021). The use of consistent scoring

functions for the intensity is illustrated in finite sample simulation experiments in Sect. 4. In Sect. 5, we evaluate operational earthquake forecasts for Italy and discuss how scoring functions relate to extant methods in seismology. The paper closes with a discussion in Sect. 6.

The main article concentrates on scoring functions for the intensity—the most fundamental first order property of a point process. Scoring functions and simulation experiments for further standard properties such as moment measures are addressed in the Supplementary Material.

2 Scoring functions and forecast evaluation

The following overview of consistent scoring functions and their role in comparative forecast evaluation is primarily based on Gneiting (2011).

Let O and A be subsets of a real vector space, and let $\mathscr{F}$ be a collection of probability distributions on the Borel-σ-algebra of O. We interpret $x \in \mathsf{A}$ as a forecast in terms of a single-valued *functional* $T : \mathscr{F} \to \mathsf{A}$ that is to be compared to an outcome in O. A function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called *scoring function* if for all $x \in \mathsf{A}$ the mapping $S(x, \cdot)$ is F-integrable for all $F \in \mathscr{F}$. The literature usually distinguishes point forecasts ($A \subseteq \mathbb{R}^n$) and probabilistic forecasts ($A = \mathscr{F}$ and T is the identity) and uses the term *scoring rule* in the latter setting. We do not make this distinction and exclusively use the term scoring function.

The key concept, which motivates the use of scoring functions, is consistency, meaning that a perfect forecast should achieve the lowest score in expectation. Specifically, a scoring function S is *consistent* for a functional $T : \mathscr{F} \to \mathsf{A}$ if for all $x \in \mathsf{A}$ and $F \in \mathscr{F}$ we have

$$\mathbb{E}_F S(x, Y) \geq \mathbb{E}_F S(T(F), Y), \tag{1}$$

where the expectation $\mathbb{E}_F$ refers to the random variable Y following the distribution F. It is *strictly consistent* for T if in addition equality in (1) implies $x = T(F)$. A central question is which functionals T are *elicitable*, i.e. possess a strictly consistent scoring function. Many elicitable functionals and corresponding classes of strictly consistent scoring functions are known, e.g. expectations, quantiles, and expectiles (Gneiting 2011; Dawid and Musio 2014; Frongillo and Kash 2015, 2021). For $\mathsf{A} = \mathscr{F}$ the most relevant functionals are the identity and restrictions to the tails (Gneiting and Raftery 2007; Gneiting and Ranjan 2011; Lerch et al. 2017; Holzmann and Klar 2017).

A fundamental result is that expectations of integrable functions are elicitable. For instance, $\mathbb{E}_F(x - Y)^2$ is uniquely minimized by $x = \mathbb{E}_F Y$; thus, the *quadratic* score $S(x, y) = (x - y)^2$ is a strictly consistent scoring function for the expectation functional. To state a general theorem on the elicitability of expectations (Savage 1971; Gneiting 2011; Frongillo and Kash 2015), let $\mathsf{A}, \mathsf{O} \subseteq \mathbb{R}^k$ and let $\nabla f(x)$ denote the subderivative of a convex function $f : \mathsf{A} \to \mathbb{R}^k$ at $x \in \mathbb{R}^k$. The subderivative or subgradient is a generalization of the derivative that applies to any convex function, and the two concepts coincide if the derivative exists (Rockafellar 1970). The function $b : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ defined by

$$b(x, y) = -f(x) - \nabla f(x)^\top (y - x) \tag{2}$$

is called a *Bregman function* for f. If f is strictly convex, we call b *strict*.

Theorem 1 (elicitability of expectations) *Let $h : \mathsf{O} \to \mathbb{R}^k$ be F-integrable for all $F \in \mathscr{F}$. Then, the functional $T : \mathscr{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ defined via*

$$T(F) = \int h(y)\, \mathrm{d}F(y) = \mathbb{E}_F\, h(Y) = \left(\mathbb{E}_F h_1(Y), \ldots, \mathbb{E}_F h_k(Y)\right)^\top$$

is elicitable, and consistent scoring functions $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ are given by $S(x, y) = b(x, h(y))$, where b is a Bregman function. If b is strict, then S is strictly consistent for T.

In general, bijective transformations of the domain A preserve the elicitability of a functional, a fact which is usually called *revelation principle* (Gneiting 2011, Theorem 4). Likewise, if we consider transformations of the observation domain O, we can state the following simple result, which resembles, but differs from, findings on weighted functionals as discussed in Gneiting and Ranjan (2011) and Gneiting (2011, Theorem 5). The proof is a straightforward consequence of integration with respect to the pushforward measure and thus omitted.

Proposition 1 (transformation principle) *Let $T : \mathscr{F} \to \mathsf{A}$ be an elicitable functional and $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ a (strictly) consistent scoring function for T. Let $g : \mathsf{O}' \to \mathsf{O}$ be measurable, and let $\mathscr{F}'$ be a set of distributions on O', such that $\{F' \circ g^{-1} \mid F' \in \mathscr{F}'\} \subseteq \mathscr{F}$. Then, the functional $T' : \mathscr{F}' \to \mathsf{A}$ defined via $T'(F') := T(F' \circ g^{-1})$ is elicitable with (strictly) consistent scoring function $S'(x, y) = S(x, g(y))$.*

In case multiple forecasts in terms of an elicitable functional T are available, their predictive performance can be assessed in a natural way: If S is a strictly consistent scoring function for T, then a forecast is considered superior to its competitor if it achieves a lower expected score with respect to S. This allows for a choice between two forecasts based on their difference in expected scores, without further assumptions on the data-generating process.

To illustrate the idea, we introduce a simple point process scenario, which is motivated by our earthquake forecasting case study (Sect. 5). Let $\boldsymbol{\Phi}$ be a spatial point process, which models the locations of earthquake epicentres in a specified region during a period of seven days. Let S be a scoring function such that $S(r, \boldsymbol{\Phi})$ is the score of the forecast report $r \in \mathsf{A}$, and assume that S is strictly consistent for a statistical property of point processes, e.g. the intensity measure (see Sect. 3.3). In this situation, two intensity forecasts r and r^* can be compared based on $\mathbb{E}(S(r, \boldsymbol{\Phi}) - S(r^*, \boldsymbol{\Phi}))$, where, due to the consistency of S, negative values support r, while positive values support r^*.

In typical applications, we face forecasts r_t, r_t^* and corresponding realizations $\boldsymbol{\Phi}_t$ of the point process for time points $t = 1, \ldots, N$. With these values, the expected

score difference can be estimated via the realized average score difference. Substantial deviations from zero then indicate differences in the predictive performance of the forecast sequences (r_t) and (r_t^*). To estimate the uncertainty inherent in the score differences, it is common to use the Diebold–Mariano test (Diebold and Mariano 1995) or extensions of this testing framework, see e.g. Nolde and Ziegel (2017) and Hering and Genton (2011).

Although we here focus on the specific scenario of a discretely observed spatial point process, strictly consistent scoring functions can be used in many other point process settings, as discussed in Section S1 of the Supplementary Material.

3 Consistent scoring functions for point patterns

We now turn our attention to the situation, where each observation is a finite point pattern. We first connect to existing theory (Sect. 2) and then derive scoring functions for the distribution and the intensity measure. Scoring functions for further point process characteristics are discussed in Section S2 of the Supplementary Material.

3.1 Technical context

We follow the common convention that a finite *point process* $\boldsymbol{\Phi}$ is a random element in the space $\mathbb{M}_0 = \mathbb{M}_0(\mathcal{X})$ of finite counting measures on the Borel set $\mathcal{X} \subseteq \mathbb{R}^d$ and refer to Daley and Vere-Jones (2003) for details. We denote a set of probability measures on $\mathbb{M}_0$ by $\mathscr{P}$ and the distribution of $\boldsymbol{\Phi}$ by $P_{\boldsymbol{\Phi}}$. Any forecast is issued for a *functional* $\Gamma : \mathscr{P} \to \mathsf{A}$ and is to be compared to an outcome in $\mathbb{M}_0$. We call a mapping $S : \mathsf{A} \times \mathbb{M}_0 \to \mathbb{R}$ a *scoring function* if $\mathbb{E}_P S(a, \boldsymbol{\Phi}) = \int S(a, \varphi)\, \mathrm{d}P(\varphi)$ exists for all $a \in \mathsf{A}$ and $P \in \mathscr{P}$. *Elicitability* of Γ as well as *(strict) consistency* of S is then defined as above via inequality (1), i.e. S is strictly consistent for Γ if $\mathbb{E}_P S(a, \boldsymbol{\Phi}) \geq \mathbb{E}_P S(\Gamma(P), \boldsymbol{\Phi})$ for all $a \in A$ and $P \in \mathscr{P}$ and equality implies $a = \Gamma(P)$. For ease of presentation and practical implementation, we will usually state how the score of a realization $\varphi = \sum_{i=1,\dots,n} \delta_{y_i} \in \mathbb{M}_0$ is computed from an enumeration of its points, i.e. from the set $\{y_1, \dots, y_n\}$, where $n = |\varphi|$ is the total mass of the counting measure $\varphi \in \mathbb{M}_0$. To make this meaningful, we will ensure that for spatial processes all scoring functions are independent of the enumeration of points (Daley and Vere-Jones 2003, Chapter 5).

In light of Theorem 1, constructing simple examples for elicitable functionals of point processes is straightforward: Point processes induce real-valued random variables in many ways and the expectations of these random variables (provided they are finite) will be elicitable functionals.

Example 1 (expected number of points) Given a set $B \in \mathscr{B}(\mathcal{X})$, the ($\mathbb{N}_0$-valued) random variable $\boldsymbol{\Phi}(B)$ denotes the number of points of $\boldsymbol{\Phi}$ in B. According to Theorem 1, the functional $\Gamma_B : \mathscr{P} \to \mathbb{R}$ given by $\Gamma_B(P) = \mathbb{E}_P \boldsymbol{\Phi}(B)$ is elicitable with Bregman scoring function

$$S_B(x, \varphi) = b(x, \varphi(B)) = -f(x) - \nabla f(x)^\top (\varphi(B) - x),$$

where $f : [0, \infty) \to \mathbb{R}$ is a strictly convex function.

This construction is not limited to the expected number of points in a set, but works for any combination of elicitable functional (e.g. expectation) and point process feature (e.g. number of points): Let O be an observation domain and $g : \mathbb{M}_0 \to \mathsf{O}$ a measurable mapping. The transformation principle (Proposition 1) then implies that the functional $\Gamma(P) := T(P \circ g^{-1})$ is elicitable whenever $T : \{P \circ g^{-1} \mid P \in \mathscr{P}\} \to \mathsf{A}$ is elicitable. We recover Example 1 by choosing $T(F) = \mathbb{E}_F Y$ and $g(\varphi) = \varphi(B)$. The elicitability of other "simple" properties such as finite-dimensional distributions and void probabilities is a straightforward consequence of Proposition 1 and deferred to the Supplementary Material.

Different choices for T and g in Proposition 1 lead to a wide variety of different functionals and consistent scoring functions. The core idea in Heinrich-Mertsching et al. (2021) is to choose T as the identity on $\{P \circ g^{-1} \mid P \in \mathscr{P}\}$. Two distributional models $P, Q \in \mathscr{P}$ of the process $\boldsymbol{\Phi}$ can then be compared based on realizations by comparing $P \circ g^{-1}$ and $Q \circ g^{-1}$ via a consistent scoring function for distributions. The mapping $g : \mathbb{M}_0 \to \mathsf{O}$ is selected to be an estimator of some quantity of interest, e.g. a kernel-based intensity estimator. Since the distributions of such estimators will usually not be explicitly available, approximating the scoring functions via simulations becomes necessary. Moreover, as different $P \in \mathscr{P}$ may lead to the same law $P \circ g^{-1}$, this approach hinges on the ability of g to discriminate between two distributions P and Q.

Instead of following this approach, we focus on common point process characteristics $\Gamma : \mathscr{P} \to \mathsf{A}$ and develop strictly consistent scoring functions for them. This allows for a direct comparison of the characteristic Γ, which includes distributional models $P \in \mathscr{P}$ as a special case. In contrast, comparison in Heinrich-Mertsching et al. (2021) always depends on specific aspects of the distributions in $\mathscr{P}$ which are determined via the estimator choice g. This arguably leads to a good discrimination ability, as the whole point process distribution is taken into account, whereas comparison in our approach focuses on how similar the property values $\Gamma(P)$ and $\Gamma(Q)$ (e.g. the intensity measures) are. However, this also means that knowledge of the distribution P is not needed in our setting, as long as $\Gamma(P)$ is available. In cases where Γ can be computed explicitly for models in $\mathscr{P}$, this avoids point process simulations, which might be prohibitive in routine applications. Furthermore, this simplifies reporting, since forecasters do not need to come up with a fully specified point process distribution. For these reasons the methodology proposed here complements the approach developed by Heinrich-Mertsching et al. (2021), and which is more suitable depends on the setting at hand.

3.2 Distribution and density

In this subsection, we construct consistent scoring functions for the identity functional $\Gamma = \mathrm{id}_{\mathscr{P}}$, i.e. for the entire point process distribution. To this end, we need to specify how we represent the law $P_{\boldsymbol{\Phi}}$ of the finite point process $\boldsymbol{\Phi}$ on $\mathscr{X}$. One way to do so is via sequences $(p_k)_{k\in\mathbb{N}_0}$ and $(\Pi_k)_{k\in\mathbb{N}}$. Each p_k specifies the probability of finding k points in a realization, and Π_k are symmetric probability measures on $\mathscr{X}^k$ which describe the distribution of any ordering of points, given k points are realized (Daley and Vere-Jones 2003, Chapter 5.3). Although this representation already allows for the construction of consistent scoring functions for $P_{\boldsymbol{\Phi}}$, we focus on the case where densities are available, since these are often more convenient to deal with, especially when multivariate distributions are of interest.

Gneiting and Raftery (2007) formalize density forecasting as follows: Let $(\Omega, \mathscr{A}, \mu$ be a σ-finite measure space and for $\alpha > 1$ let $\mathscr{L}_\alpha$ consist of all (equivalence classes of) densities p of probability measures P that are absolutely continuous with respect to μ and such that $\|p\|_\alpha := \left(\int_\Omega p(\omega)^\alpha \,\mathrm{d}\mu(\omega)\right)^{1/\alpha}$ is finite. In this setting, important examples of strictly consistent scoring functions $S : \mathscr{L}_\alpha \times \Omega \to \mathbb{R}$ are the *pseudospherical* and the *logarithmic* score, defined via

$$\mathrm{PseudoS}(p,\omega) = -p(\omega)^{\alpha-1}/\|p\|_\alpha^{\alpha-1} \quad \text{and} \quad \mathrm{LogS}(p,\omega) = -\log p(\omega), \tag{3}$$

respectively. The logarithmic score is the (appropriately scaled) limiting case of the pseudospherical score as $\alpha \to 1$.

Returning to point processes we follow, Daley and Vere-Jones (2003, Chapters 5.3 and 7.1) and let P_0 denote the distribution of the Poisson point process with unit rate on some bounded domain $\mathscr{X} \subset \mathbb{R}^d$. If $P \in \mathscr{P}$ is absolutely continuous with respect to P_0, then the Radon–Nikodým density $\mathrm{d}P/\mathrm{d}P_0$ exists and can be regarded as the density of P. It can be computed via the identity

$$\frac{\mathrm{d}P}{\mathrm{d}P_0}(\varphi) = \exp(|\mathscr{X}|)\frac{j_k(y_1,\dots,y_k)}{k!},$$

where $|\mathscr{X}|$ denotes the Lebesgue measure of $\mathscr{X}$, $y_1,\dots,y_k$ are the points of $\varphi \in \mathbb{M}_0$, and the (symmetric) function j_k given by

$$j_k(x_1,\dots,x_k)\,\mathrm{d}x_1 \cdots \mathrm{d}x_k = k! p_k \,\mathrm{d}\Pi_k(x_1,\dots,x_k) \tag{4}$$

is the k-th *Janossy density* of $\boldsymbol{\Phi}$. For $k = 0$ this is interpreted as $j_0 = p_0$. The value $j_k(x_1,\dots,x_k)$ can be understood as the *likelihood* of k points materializing, one of them in each of the distinct locations $x_1,\dots,x_k \in \mathscr{X}$.

In principle, plugging the Janossy densities into (3) allows us to obtain scoring functions for the point process distribution P. However, two important difficulties need to be addressed in the point process setting. First, explicit expressions for $(j_k)_{k\in\mathbb{N}_0}$ are usually hard to determine and known only for some models, see Daley and Vere-Jones (2003, Chapter 7.1) and Example 2 below. Second, even if explicit expressions are available, calculating the norm $\|\mathrm{d}P/\mathrm{d}P_0\|_\alpha$ amounts to computing $(k!)^{-1}\int j_k(x_1,\dots,x_k)^\alpha \,\mathrm{d}x_1 \dots \mathrm{d}x_k$ for all $k \in \mathbb{N}$, which may be prohibitive. This

complicates the use of scoring functions relying on $\|\cdot\|_\alpha$, such as the pseudospherical score (3). We will thus only consider the logarithmic score here and discuss a further choice in the Supplementary Material.

Assume that for all distributions $Q \in \mathscr{P}$ the corresponding Janossy densities $(j_k^Q)_{k\in\mathbb{N}_0}$ are well-defined. Due to the strict consistency of the logarithmic score, the function $S : \mathscr{P} \times \mathbb{M}_0 \to \mathbb{R}$ defined via

$$S((j_k^Q)_{k\in\mathbb{N}_0}, \{y_1, \dots, y_n\}) = -\log(j_n^Q(y_1, \dots, y_n)) \tag{5}$$

for $n \in \mathbb{N}$ and $S((j_k^Q)_{k\in\mathbb{N}_0}, \emptyset) := -\log(j_0^Q)$ is a strictly consistent scoring function for the distribution of the point process $\boldsymbol{\Phi}$. The term $-|\mathscr{X}| + \log(n!)$ can be omitted, since it is independent of the forecast report $(j_k^Q)_{k\in\mathbb{N}_0}$. This choice recovers the log-likelihood of the point process distribution Q from the perspective of consistent scoring functions.

Example 2 (Poisson point process) Let $\boldsymbol{\Phi}$ be an inhomogeneous Poisson point process with intensity $\lambda : \mathscr{X} \to [0, \infty)$. It is well-known that $\boldsymbol{\Phi}$ admits the densities

$$j_{n,\lambda}(y_1, \dots, y_n) = \Big(\prod_{i=1}^n \lambda(y_i)\Big) \exp\left(-\int_{\mathscr{X}} \lambda(y)\,\mathrm{d}y\right)$$

for $n \in \mathbb{N}$. In case $n = 0$ the product is interpreted as one. When reporting the Poisson point process distribution $P_{\boldsymbol{\Phi}}$ via its Janossy densities, (5) gives the score

$$S(P_{\boldsymbol{\Phi}}, \{y_1, \dots, y_n\}) = -\sum_{i=1}^n \log \lambda(y_i) + \int_{\mathscr{X}} \lambda(y)\,\mathrm{d}y \tag{6}$$

for $n \in \mathbb{N}$ and $S(P_{\boldsymbol{\Phi}}, \emptyset) = \int_{\mathscr{X}} \lambda(y)\,\mathrm{d}y$.

Before turning to the intensity measure, we briefly discuss temporal point processes, which demand a special treatment since the dimension "time" possesses a natural ordering. The instantaneous rate of points occurring in the point process $\boldsymbol{\Phi}$ is usually described via the *conditional intensity*

$$\lambda^*(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}\big(\boldsymbol{\Phi}((t, t+\Delta t)) \mid \mathscr{H}_t\big)}{\Delta t}, \tag{7}$$

where $(\mathscr{H}_t)_{t\in\mathbb{R}}$ is the filtration generated by the history of $\boldsymbol{\Phi}$ (Reinhart 2018; Daley and Vere-Jones 2003, Chapter 7). Although $\lambda^*(t)$ is random, it is deterministic conditional on $\boldsymbol{\Phi}$, thus a measurable mapping linking it to $\boldsymbol{\Phi}$ allows for modelling as well as evaluation via consistent scoring functions.

Specifically, let $\boldsymbol{\Phi}$ be a point process on $\mathbb{R}$ and consider an observation window $\mathscr{X} := [0, T]$ for some $T > 0$. Given a realization $0 < t_1 < \cdots < t_n$ of $\boldsymbol{\Phi}$ the realized values of the conditional intensity can be computed for all $t \in \mathscr{X}$. More precisely, for a $t \in \mathscr{X}$ with $t_1 < \cdots < t_i \le t < t_{i+1}$ we denote the realized value of λ^* at t via $\lambda^*(t \mid t_1, \dots, t_i)$. Since the collection of all mappings $t \mapsto \lambda^*(t \mid t_1, \dots, t_i)$ for all $i = 0, \dots, n$ and all possible realizations $t_1, \dots, t_n$ uniquely determines the

distribution of $\boldsymbol{\Phi}$ (Daley and Vere-Jones 2003), comparing forecasts for the conditional intensity is equivalent to a comparison of forecasts for the distribution. This connection is made explicit by the representation of the likelihood of $t_1, \dots, t_n$ occurring in $[0, T]$ via

$$j_n(t_1, \dots, t_n) = \Big(\prod_{i=1}^{n} \lambda^*(t_i) \Big) \exp\left(- \int_0^T \lambda^*(u)\, \mathrm{d}u \right), \tag{8}$$

where the product is interpreted as one if no points occur. Consequently, (strictly) consistent scoring functions for the conditional intensity can be obtained by arguments similar to above.

Example 3 (Recovery of log-likelihood of a temporal point process) Plugging (8) into the logarithmic score (5), we see that the scoring function

$$S(\lambda^*, \{t_1, \dots, t_n\}) = - \sum_{i=1}^{n} \log\big(\lambda^*(t_i)\big) + \int_0^T \lambda^*(u)\, \mathrm{d}u,$$

is strictly consistent for the conditional intensity. This recovers the log-likelihood of a temporal point process (Daley and Vere-Jones 2003; Reinhart 2018). If $\boldsymbol{\Phi}$ is a Poisson point process on $\mathbb{R}$, its conditional intensity λ^* agrees with its intensity λ, and S coincides with (6).

3.3 Intensity measure

One of the key characteristics of a point process $\boldsymbol{\Phi}$ is its intensity measure $\Lambda : B \mapsto \mathbb{E}\boldsymbol{\Phi}(B)$ that quantifies the expected number of points in any set $B \in \mathscr{B}(\mathscr{X})$ (Daley and Vere-Jones 2003; Chiu et al. 2013). Analogous to the first moment of a univariate random variable, it describes the average behaviour of the point process $\boldsymbol{\Phi}$. For a fixed Borel set B, we have already identified the expected number of points $\Lambda(B) = \mathbb{E}\boldsymbol{\Phi}(B)$ as an elicitable functional (Example 1). Here, we focus on constructing scoring functions for the full measure Λ as a functional on $\mathscr{P}$ with values in a set of finite measures $\mathscr{M}_\mathrm{f}$ on $\mathscr{X}$. To this end, we call $\Lambda^* := \Lambda/|\Lambda|$, where $|\Lambda| := \Lambda(\mathscr{X})$ is the total mass of Λ, the *normalized measure* of a finite measure $\Lambda \in \mathscr{M}_\mathrm{f}$.

Proposition 2 *Set* $\mathscr{F} := \{\Lambda^* \mid \Lambda \in \mathscr{M}_\mathrm{f}\}$ *and let* $S' : \mathscr{F} \times \mathscr{X} \to \mathbb{R}$ *be a* (*strictly*) *consistent scoring function for* $\mathrm{id}_{\mathscr{F}}$. *Let* $b : [0, \infty) \times [0, \infty) \to \mathbb{R}$ *be a* (*strict*) *Bregman function*, *as in* (2). *The scoring function* $S : \mathscr{M}_\mathrm{f} \times \mathbb{M}_0 \to \mathbb{R}$ *defined via*

$$S(\Lambda, \{y_1, \dots, y_n\}) := \sum_{i=1}^{n} S'(\Lambda^*, y_i) + cb\big(|\Lambda|, n\big)$$

for $n \in \mathbb{N}$ *and* $S(\Lambda, \emptyset) = cb(|\Lambda|, 0)$ *for* $c > 0$, *is consistent for the intensity measure. It is strictly consistent if* S' *is strictly consistent and* b *is strict.*

Proof Let $W \in \mathscr{M}_\mathrm{f}$ and $\boldsymbol{\Phi}$ be a point process with intensity measure $\Lambda \in \mathscr{M}_\mathrm{f}$ and distribution $P \in \mathscr{P}$. The difference in expected scores is

$$\mathbb{E}_P[S(W, \boldsymbol{\Phi}) - S(\Lambda, \boldsymbol{\Phi})] = \int \sum_{x \in \varphi} S'(W^*, x) - S'(\Lambda^*, x) \, \mathrm{d}P(\varphi) + c\mathbb{E}_P(b(|W|, |\boldsymbol{\Phi}|) - b(|\Lambda|, |\boldsymbol{\Phi}|))$$

and the last term is nonnegative since b is a Bregman function. Using Campbell's theorem, the second expression equals

$$\int_{\mathscr{X}} S'(W^*, x) - S'(\Lambda^*, x) \, \mathrm{d}\Lambda(x) = |\Lambda| \int_{\mathscr{X}} S'(W^*, x) - S'(\Lambda^*, x) \, \mathrm{d}\Lambda^*(x),$$

and is also nonnegative, due to the consistency of S'. If the score difference is zero, b is strict, and S' is strictly consistent, this gives $W^* = \Lambda^*$ and $|W| = |\Lambda|$, showing that S is strictly consistent for the intensity measure. □

In principle, it is possible to define scoring functions which only depend on normalized measures, by using arguments in Hendrickson and Buehler (1971) who discuss a connection to homogeneous functions on the cone induced by a set of probability measures. As we are interested in the full intensity measure, we combine the total mass $|\Lambda| = \mathbb{E}\boldsymbol{\Phi}(\mathscr{X})$, which is an elicitable property of $\boldsymbol{\Phi}$ (Example 1), with Λ^* to obtain a consistent scoring function.

Example 4 As an important special case, assume that each $\Lambda \in \mathscr{M}_\mathrm{f}$ admits a density λ with respect to Lebesgue measure. Using the common quadratic score for b and the logarithmic score (3) for S', the strictly consistent scoring function of Proposition 2 becomes

$$S(\Lambda, \{y_1, \dots, y_n\}) = -\sum_{i=1}^{n} \log(\lambda(y_i)) + n \log |\Lambda| + c\,(|\Lambda| - n)^2$$

for some $c > 0$. Simulation experiments in Sect. 4 illustrate how S can be used to compare intensity forecasts.

The choice of the constant $c > 0$ in Proposition 2 is irrelevant for (strict) consistency of the scoring function S. However, since S evaluates both the shape and the total mass of the intensity, judicious choices of c serve to balance the scoring components.

4 Simulation study

In this section, we investigate finite sample properties of scoring function-based model evaluation via mean score differences, with focus on intensity forecasting for spatial point processes. All calculations are performed with R (R Core Team 2021),

including point process simulations with the `spatstat` package (Baddeley and Turner 2005; Baddeley et al. 2015).

We compare different intensity reports for a point process $\boldsymbol{\Phi}$ on the window $[0,1]^2$ based on $N \in \mathbb{N}$ realizations, where N could reflect a number of different time windows, e.g. $N = 52$ for one year of weekly data. We draw $N = 100$ i.i.d. samples φ_i from $\boldsymbol{\Phi}$ and use the mean score

$$\bar{s}_j := \frac{1}{N} \sum_{i=1}^{N} S(f_j, \varphi_i)$$

as an estimator of the expected score $\mathbb{E}S(f_j, \boldsymbol{\Phi})$ of a given forecast intensity f_j in the population. We use the scoring function S from Example 4 with scaling factor $c = 1/10$ such that the logarithmic and squared terms vary at the same order of magnitude. The simulations are repeated $M = 500$ times to assess the variation in mean scores.

We consider four different data-generating processes for $\boldsymbol{\Phi}$, all of which have (approximate) intensity $f_0(x, y) = 6\sqrt{x^2 + y^2}$, which leads to four different simulation experiments. In the first experiment, $\boldsymbol{\Phi}$ is an inhomogeneous Poisson point process. In the second, $\boldsymbol{\Phi}$ is a determinantal point process (DPP) with Gaussian covariance such that its points exhibit moderate inhibition. In the remaining two simulation experiments $\boldsymbol{\Phi}$ inclines to clustering. For the third one, we choose a log-Gaussian Cox process (LGCP) with exponential covariance and log-expectation μ such that its intensity equals f_0. In the last experiment, $\boldsymbol{\Phi}$ is an inhomogeneous Thomas process, i.e. a cluster process which arises from an inhomogeneous Poisson process as parent and a random number of cluster points which are drawn from a normal distribution centred at its parent point. Due to this clustering, the intensity of the Thomas process is only approximately equal to f_0. For details on the processes see Lavancier et al. (2015), Illian et al. (2008, Chapter 6) and Section S3 of the Supplementary Material.

The study compares six different intensity forecasts, namely f_0 and

$$\begin{aligned}
f_1(x, y) &= 7.8\sqrt{(x - 0.2)^2 + (y - 0.1)^2}, \\
f_2(x, y) &= 2.3(x + 3y), \\
f_3(x, y) &= 10\sqrt{(x - 0.2)^2 + (y - 0.1)^2}, \\
f_4(x, y) &= 7.5 \exp\left[-3\left\{(x - 0.6)^2 + (y - 0.6)^2\right\}\right], \\
f_5(x, y) &= 2\left\{\frac{1}{\sqrt{1.2 - x}} + 2(1 - y)\right\}.
\end{aligned}$$

these choices are motivated as follows. Intensity f_1 has the correct shape, up to a small shift, and f_3 is a version of f_1 with too high total mass. Intensity f_2 is similar to f_0 but linear, while f_4 and f_5 have completely different shape, as illustrated by Figure S1 in the Supplementary Material. Except for f_3, all intensities put roughly identical mass on $[0,1]^2$. This allows for an assessment of how the scoring function reacts to misspecifications in shape instead of total mass.

Figure 1 shows the mean score differences between the five different forecasts $f_1, \ldots, f_5$ and the optimal forecast f_0 for all experiments. The four experiment show a similar pattern, namely f_1 is close to the optimal forecast, f_2 and f_3 less so, and the mean score differences of the misspecified functions f_4 and f_5 are far from zero. The fourth experiment shows an increase in variance, which likely stems from the strong clustering tendency of the process. Moderate clustering or inhibition, as present in the third and second experiment, seem to have almost no impact on the score differences. Overall, varying the intensity forecasts leads to pronounced differences in realized average scores, highlighting differences in forecast performance. Further experiments with different scoring functions as well as tests for superior predictive ability are given in Section S3 of the Supplementary Material.

5 Case study: Earthquake forecasting

In this case study, we illustrate how consistent scoring functions can be used to compare earthquake forecasting models, and we shed new light on extant evaluation methods in seismology. All calculations are performed with R (R Core Team 2021).

5.1 Earthquake forecasting experiments

Over the past decades it has become consensus that earthquake forecasts ought to be probabilistic, i.e. instead of specifying whether or not an earthquake will occur, they provide a respective predictive distribution or aspects thereof (Jordan et al. 2011). Statistical models to issue such forecasts are based on spatiotemporal point processes. They are usually specified via a conditional intensity (see (7)) that exhibits self-exciting behaviour, reflecting the conjecture that earthquakes trigger each other and cluster in space and time. An important example is the epidemic-type aftershock sequence (ETAS) model, see e.g. Kagan and Knopoff (1987) and Ogata (1988, 1998).

The Collaboratory for the Study of Earthquake Predictability (CSEP, see Introduction) evaluates earthquake forecasts prospectively in several regional testing centres with standardized testing routines. The prospective approach uses only forecasts submitted in real time before the respective outcomes are realized, which guarantees independence of the forecasts from actual observations. An important part of these routines is the earthquake likelihood model testing approach of Kagan and Jackson (1995) and Schorlemmer et al. (2007), which we discuss in Sect. 5.3. Our case study relies on data from the operational earthquake forecasting system in Italy (OEF-Italy, Marzocchi et al. (2014)), which is based on the three independent short-term forecasting models that were tested prospectively in a CSEP testing centre for the Italian testing region (Taroni et al. 2018). See Fig. 2 for an illustration.

The three independent models comprise LM (Lombardi and Marzocchi 2010) and FMC (Falcone et al. 2010), which are ETAS-based models with distinct structure and calibration choices, and LG (Woessner et al. 2010), which is based on the short-term earthquake probability (STEP) model of Gerstenberger et al. (2005) and composed of sub-models. We refer to the original references for more details about the individual

models. OEF-Italy also includes an aggregated or ensemble forecast, namely SMA, which predicts a weighted average of the above three models using the score model averaging (SMA) rule (Marzocchi et al. 2012), with models being weighted inversely proportional to the log-likelihood of observed data. The SMA model is updated continuously based on new observations and was successfully applied to track the evolution of the recent earthquake sequence in central Italy in real time (Marzocchi et al. 2017).

Our study considers earthquakes of magnitude greater or equal to four (M4+) between April 2005 and May 2020 (5520 days) that fall into the Italian CSEP testing region (Fig. 2). The testing region is divided into 8993 grid cells. On each day, the four models produce forecasts for the expected number of M4+ earthquakes in the subsequent seven-day period for each grid cell. The forecasts are thus nonnegative values $x_{i,t}^{(j)}$ where j denotes the model, i the cell, and t the day. They can then be compared to the observed number of events in each cell for that upcoming week. Since the forecasts concern seven-day periods, this number is only known seven days after a forecast was issued. For the same reason, the number of days available for evaluation reduces to 5514.

5.2 Model comparison and results

Since the models we consider produce mean forecasts, we have to employ (strictly) consistent scoring functions for the expectation functional for a sound comparison, see also Example 1. Such functions are of the Bregman form (2) and a natural choice is the quadratic score $S_{\text{quad}}(x, y) = (x - y)^2$. However, the quadratic score focuses on no particular forecast cases in the sense of elementary scores (Ehm et al. 2016). As an alternative that puts more emphasis on small forecast values and connects to the CSEP methods (see Sect. 5.3) we use the *Poisson* scoring function $S_{\text{pois}} : (0, \infty) \times \mathbb{N}_0 \to \mathbb{R}$ defined via

$$S_{\text{pois}}(x, y) = -y \log(x) + x. \tag{9}$$

It is strictly consistent since it is a Bregman function corresponding to the strictly convex function $f(x) = x(\log(x) - 1)$. Note that (9) can be interpreted as a discrete analogue to the Dawid–Sebastiani score (Dawid and Sebastiani 1999), but with the normal distribution replaced by the Poisson distribution (Brehmer 2021). To obtain a daily score of the forecast models, the individual scores for the 8993 grid cells are summed up. The daily scores and the mean score of model j are thus given by

$$s_{j,t} := \sum_{i=1}^{8993} S\big(x_{i,t}^{(j)}, \varphi_t(B_i)\big) \qquad \text{and} \qquad \bar{s}_j := \frac{1}{5514} \sum_{t=1}^{5514} s_{j,t}, \tag{10}$$

respectively, where $\varphi_t(B_i)$ is the observed number of events in cell B_i over the period from day t to $t + 6$. The mean score $\bar{s}_j$ estimates the expected score of model j and is thus a measure of the relative forecast performance of this model. Figure 3 depicts the daily scores (10) based on S_{pois} for the four different models. It uses a logarithmic scale, because the values are much larger on days when events occur, in comparison to days without events. The FMC model consistently achieves the lowest

Table 1 Summarized performance of the four models according to the mean score over the testing period $\bar{s}_j$ from (10). The scoring functions used for evaluation are the Poisson ("pois") and the quadratic ("quad") score. Lowest values in each column are in boldface

Model	pois	quad
LM	**2.68**	**0.8218**
FMC	2.76	0.8269
LG	2.98	0.8275
SMA	2.70	0.8248

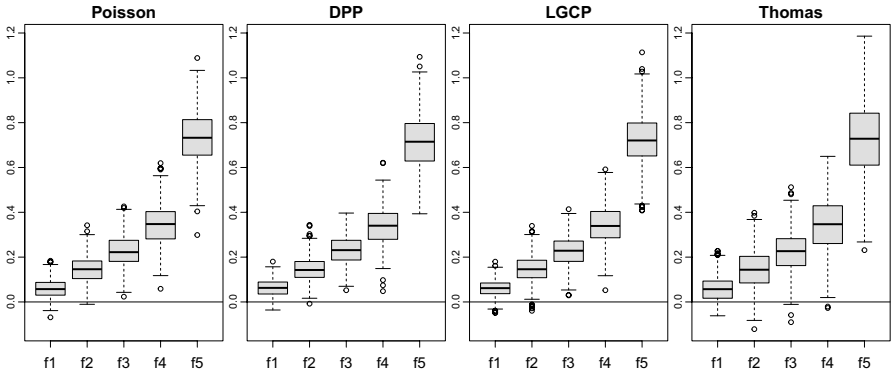

Fig. 1 Boxplots of the difference in mean scores $\bar{s}_j - \bar{s}_0$ for $j = 1, \ldots, 5$ and the scoring function S from Example 4. From left to right, $\boldsymbol{\Phi}$ is a Poisson point process, a Gaussian determinantal point process, a log-Gaussian Cox process, and an inhomogeneous Thomas process. Means are based on $N = 100$ realizations, boxplots on $M = 500$ replicates

scores on days without earthquakes, since it consistently forecasts the lowest number of events. However, overall the LM model shows the best performance in terms of mean scores over the whole testing period (10), as can be seen in Table 1. This conclusion applies under both the Poisson and the quadratic score.

To understand why the overall scores indicate superior predictive ability of the LM model, we compute the mean score difference between model j and model j' for each grid cell i via

$$\Delta_i^{(j,j')} := \frac{1}{5514} \sum_{t=1}^{5514} \left(S_{\text{pois}}(x_{i,t}^{(j)}, \varphi_t(B_i)) - S_{\text{pois}}(x_{i,t}^{(j')}, \varphi_t(B_i)) \right). \tag{11}$$

The left part of Fig. 4 plots $\Delta_i^{(1,2)}$, i.e. the mean score differences between the LM and the FMC model per grid cell. It illustrates that the lower mean score of the LM model stems from its good performance in central Italy in comparison to the FMC model. The right part illustrates aggregated performance, i.e. each pixel shows the performance when the forecasts and observed values within a square neighbourhood centred at this pixel are added up. In this case, the neighbourhood has an edge length of 11 pixels. Again, better predictive ability of the LM model is most pronounced

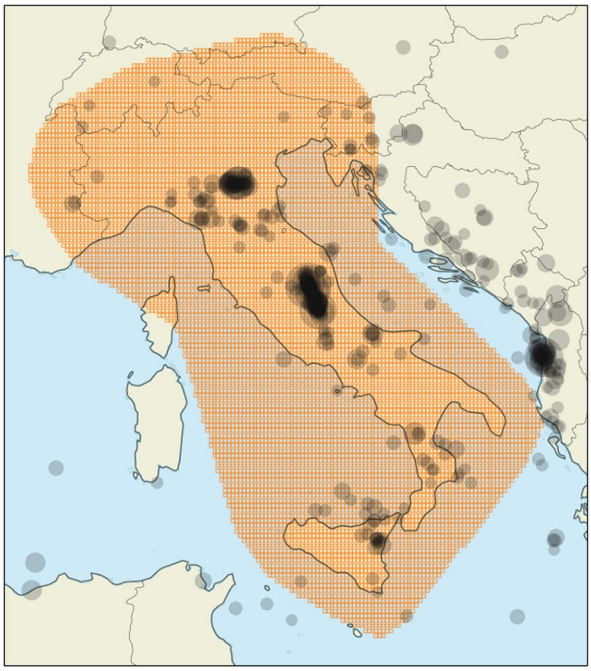

Fig. 2 Testing region of the Italian CSEP experiment. Gray circles represent locations of M4+ earthquakes. Figure reproduced from Herrmann and Marzocchi (2023)

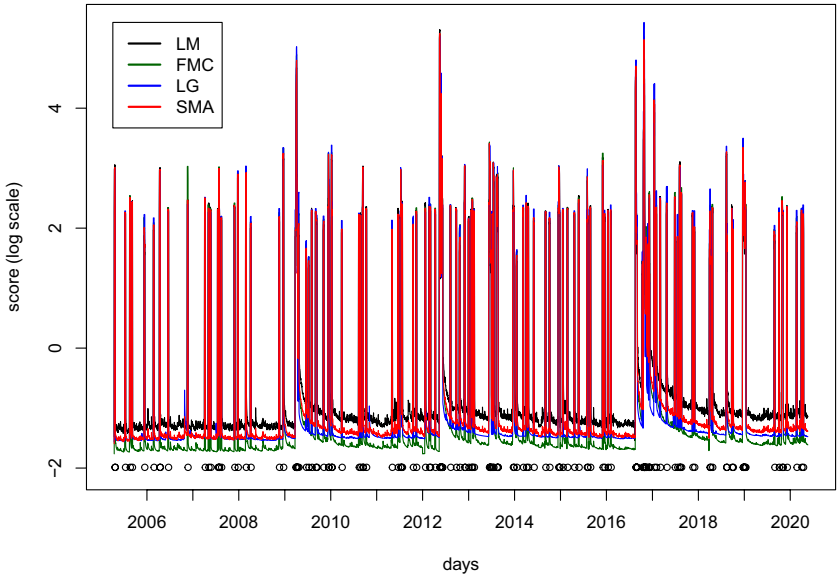

Fig. 3 Daily scores $s_{j,t}$ from (10) based on S_{pois} for the four forecasting models from 2005 to 2020, logarithmic scale. The circles indicate the days of M4+ earthquakes and the tickmarks on the horizontal axis mark the first day of each year

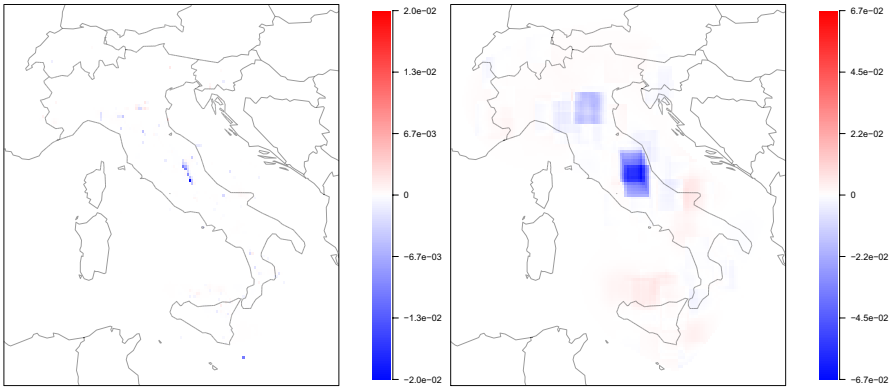

Fig. 4 Mean score difference based on S_{pois} (11) between the LM and the FMC model, without (left) and with (right) aggregation. Negative values (blue) indicate that the LM model has superior forecast performance, and positive values (red) vice versa

in central Italy and to a lesser extent in the north, i.e. in areas where earthquake sequences occurred during the study period. The opposite is true for marine regions around Sicily.

Often, lack of data complicates the forecasting of point processes as well as the proper testing of proposed forecasting models. This circumstance raises the question of how much data are needed to reach valid conclusions on superior predictive ability. As noted above, a commonly used tool is the Diebold–Mariano test (Diebold and Mariano 1995), which is a one-sample t-test applied to the score differentials, with adaptations to time series settings. Standard power calculations for t-tests apply to independent samples, where rules of thumb for the calculation of a required sample size or a detectable difference are available (Lehr 1992; van 2008). In time series settings, rules of this type also require adaptation, as exemplified in Section S4 of the Supplementary Material, which contains details on the analyses in this section.

5.3 A new perspective on earthquake likelihood model testing

An important element of the CSEP forecast experiments is a model evaluation approach introduced by Kagan and Jackson (1995) and Schorlemmer et al. (2007), to which we refer as earthquake likelihood model testing (ELMT). Further conceptual and computational improvements are due to Zechar et al. (2010a), Rhoades et al. (2011), and Ogata et al. (2013).

Put simply, ELMT represents earthquakes by points in some region $\mathscr{X} \subset \mathbb{R}^k$, which is partitioned into grid cells $B_1, \dots, B_N$ for some $N \in \mathbb{N}$, see e.g. Fig. 2. The data consist of values $x_1, \dots, x_N \in \mathbb{N}_0$ which count the earthquakes falling in each cell. A forecast or "model" is given by values $\lambda_1, \dots, \lambda_N \in (0, \infty)$ and its "log-likelihood" (Schorlemmer et al. 2007) is defined as a sum of Poisson log-likelihoods, i.e. via

$$\ell(\lambda_1, \dots, \lambda_N, x_1, \dots, x_N) = \sum_{i=1}^{N} \big(x_i \log \lambda_i - \log(x_i!) - \lambda_i\big). \tag{12}$$

This terminology is motivated by the fact that, for a Poisson point process with intensity measure Λ such that $\Lambda(B_i) = \lambda_i$, for $i = 1, \dots, N$, (12) is the log-likelihood of the observation $x_1, \dots, x_N$. Based on (12), Schorlemmer et al. (2007) propose different tests. Here, we only consider the test designed to compare forecasts.

The *R-test*, or ratio test, compares two forecasts A and B specified by their grid cell values λ_i^A and λ_i^B for $i = 1, \dots, N$, and aims to check whether model A is at least as good as model B. The R-test considers the "log-likelihood ratio" based on (12), i.e.

$$R(A, B, x_1, \dots, x_N) = \ell(\lambda_1^A, \dots, \lambda_N^A, x_1, \dots, x_N) - \ell(\lambda_1^B, \dots, \lambda_N^B, x_1, \dots, x_N), \tag{13}$$

and then compares the realized value $z := R(A, B, x_1, \dots, x_N)$ to the distribution of the random variable $Z := R(A, B, X_1, \dots, X_N)$, where $X_1, \dots, X_N$ are independent Poisson random variables with parameters λ_i^A for $i = 1, \dots, N$. If z lies in the lower tail of the distribution of Z, then model A is deemed worse than model B. As the distributional assumptions on $X_1, \dots, X_N$ demonstrate, there is an asymmetry inherent in the R-test: If model A is tested against model B, then the X_i are assumed to have parameters λ_i^A and if B is tested against A, then λ_i^B are assumed for X_i. As noted by Rhoades et al. (2011) this implies that the R-test is not really a comparative test, but rather a goodness-of-fit test. This explains seemingly contradictory results observed in practice, where R-tests deem A worse than B and vice versa, see also Bray and Schoenberg (2013) for a discussion. As a remedy, Rhoades et al. (2011) propose two modifications of the R-test, which do not rely on a Poisson assumption to determine the distribution of Z.

As pointed out by Harte (2015), ELMT suffers from several drawbacks. First, relying on a partition leads to a loss of information, since the behaviour of models inside cells does not affect the evaluation. Moreover, assuming independence across cells as well as a Poisson distribution leads to a likelihood mis-specification under general point process models. This prohibits the testing of model characteristics other than cell expectations, since by reporting $(\lambda_i)_{i=1,\dots,N}$, every forecast is treated like a Poisson point process. However, as mentioned by Bray and Schoenberg (2013), it is unclear how big the impact of the Poisson assumption is on the testing results.

Taking the perspective of consistent scoring functions, we can answer this question and clarify the role of the testing assumptions. To formalize ELMT in our setting, assume that the bounded domain $\mathcal{X}$ is partitioned into k_n grid cells $\mathcal{I}_n = \{B_1, \dots, B_{k_n}\}$. Based on (12) and (13), we define the *cell scoring function* $S_{\text{cell}}^{\mathcal{I}_n} : (0, \infty)^{k_n} \times \mathbb{M}_0 \to \mathbb{R}$ via

$$S_{\text{cell}}^{\mathcal{I}_n}(\lambda_1, \dots, \lambda_{k_n}, \varphi) = \sum_{i=1}^{k_n} -\varphi(B_i) \log(\lambda_i) + \lambda_i \tag{14}$$

for each partition $\mathscr{T}_n$, $n \in \mathbb{N}$. If $k_n = N$ and $x_i = \varphi(B_i)$ for $i = 1, \dots, N$, then (13) can be understood as the score difference between the forecasts λ_i^A and λ_i^B with respect to $S_{\text{cell}}^{\mathscr{T}_n}$. Since it applies the scoring function (9) to each grid cell, $S_{\text{cell}}^{\mathscr{T}_n}$ is strictly consistent for the collection of cell expectations $\mathbb{E}\boldsymbol{\Phi}(B_i)$, $B_i \in \mathscr{T}_n$, cf. Example 1. This shows that the Poisson log-likelihood in (13) can be used for a sound comparison of cell expectations, since the true expectations obtain the minimal expected score. We emphasize that this conclusion holds regardless of whether or not the data or the forecasts are based on Poisson point processes. Moreover, dependence among cells is irrelevant for this fact, since (strict) consistency concerns only *expected* scores. Hence, the validity of statistical methods which rely on the expected scores of (14) is not limited to Poisson models nor to Poisson point process data. In a nutshell, these methods assess forecast performance in terms of cell expectations only, since the scoring function (9) is strictly consistent for the expectation. For instance, the symmetric modifications of the R-test due to Rhoades et al. (2011) can be seen as Diebold–Mariano (DM) tests (Diebold and Mariano 1995) based on $S_{\text{cell}}^{\mathscr{T}_n}$. Hence, they test whether one model is better than its competitor in forecasting the mean number of earthquakes in the cells. Note that although such methods are valid for arbitrary point processes, considerable spatial or temporal dependencies will affect significance levels and deteriorate their ability to detect differences in forecast performance in finite samples.

It remains to discuss the role of the partitioning of $\mathscr{X}$ into grid cells. To understand its implications, note that just as the Poisson distribution leads to the scoring function (9) for the expectation, the Poisson point process can be used to obtain a scoring function for the intensity (Sect. 3.3). The reason is that every intensity report induces a Poisson point process with this intensity and these processes can then be compared via the logarithmic score (5), which attains the value (6) for Poisson densities. In the setting of Sect. 3.3, we can formalize as follows.

Proposition 3 *Let every element of $\mathscr{M}_{\text{f}}$ admit a density λ with respect to Lebesgue measure. Then, the scoring function $S : \mathscr{M}_{\text{f}} \times \mathbb{M}_0 \to \mathbb{R}$ defined by*

$$S(\Lambda, \{y_1, \dots, y_n\}) = -\sum_{i=1}^{n} \log \lambda(y_i) + \int_{\mathscr{X}} \lambda(y)\,\mathrm{d}y \tag{15}$$

for $n \in \mathbb{N}$, and $S(\Lambda, \emptyset) = \int_{\mathscr{X}} \lambda(y)\,\mathrm{d}y$, is a strictly consistent scoring function for the intensity.

Proof The scoring function (15) corresponds to S from Proposition 2 when choosing the logarithmic score for S', the Bregman function (9) for b and $c = 1$. Since S' is strictly consistent and b is strict, S is strictly consistent for the intensity. □

The scoring function (15) can be interpreted as a point process analogon to the Dawid–Sebastiani score (Dawid and Sebastiani 1999). While the Dawid–Sebastiani score relies on the first and second moments of the predictive distribution, this scoring function depends on the intensity only.

The next result shows that the cell scoring function $S_{\text{cell}}^{\mathscr{T}_n}$ serves as an approximation to the scoring function (15). Essentially, if a forecaster does not report an intensity λ, but only the integrals $\lambda_i^{(n)}$ of λ over the collection of grid cells $\mathscr{T}_n$, then forecast comparison using the cell scoring function $S_{\text{cell}}^{\mathscr{T}_n}$ is on par with a comparison based on the scoring function (15), provided the partition is sufficiently fine. The correction term in (16) does not affect the evaluation, as it is independent of the reported integrals. To make this precise, we follow Daley and Vere-Jones (2003) and call a sequence of partitions $(\mathscr{T}_n)_{n\in\mathbb{N}}$ *dissecting* if it is nesting and asymptotically separates every pair of points.

Proposition 4 *Let $\lambda : \mathscr{X} \to (0,\infty)$ be an intensity and $(\mathscr{T}_n)_{n\in\mathbb{N}}$ a dissecting system of measurable partitions of $\mathscr{X}$ which generates the Borel σ-algebra on $\mathscr{X}$. Let $P_0 \in \mathscr{P}$ be the distribution of the unit rate Poisson point process on $\mathscr{X}$ and define partition integrals*

$$\lambda_i^{(n)} = \int_{B_i^{(n)}} \lambda(y)\,\mathrm{d}y,$$

for all $i = 1,\dots,k_n$, $B_i^{(n)} \in \mathscr{T}_n$, and $n \in \mathbb{N}$. Then,

$$S_{\text{cell}}^{\mathscr{T}_n}\big(\lambda_1^{(n)},\dots,\lambda_{k_n}^{(n)},\varphi\big) + \sum_{i=1}^{k_n} \mathbb{1}(\varphi(B_i^{(n)}) > 0)\log(|B_i^{(n)}|) \longrightarrow S(\Lambda,\varphi), \tag{16}$$

for P_0-a.e. $\varphi \in \mathbb{M}_0$ as $n \to \infty$, where S is the scoring function (15).

Proof Let $\varphi = \{y_1,\dots,y_m\}$ with $m \in \mathbb{N}_0$ be a point process realization. For a large $n \in \mathbb{N}$ every set $B_i^{(n)}$ contains at most one point of φ, and we let $i_n(j)$ denote the index of the set such that $y_j \in B_{i_n(j)}^{(n)}$ for $j = 1,\dots,m$. Then, the left-hand side of (16) equals

$$\begin{aligned}
&-\sum_{i=1}^{k_n}\left(\varphi(B_i^{(n)})\log\left(\int_{B_i^{(n)}}\lambda(y)\,\mathrm{d}y\right) - \mathbb{1}(\varphi(B_i^{(n)}) > 0)\log(|B_i^{(n)}|) - \int_{B_i^{(n)}}\lambda(y)\,\mathrm{d}y\right)\\
&\quad= -\sum_{j=1}^{m}\log\left(|B_{i_n(j)}^{(n)}|^{-1}\int_{B_{i_n(j)}^{(n)}}\lambda(y)\,\mathrm{d}y\right) + \int_{\mathscr{X}}\lambda(y)\,\mathrm{d}y\\
&\quad\longrightarrow -\sum_{j=1}^{m}\log(\lambda(y_j)) + \int_{\mathscr{X}}\lambda(y)\,\mathrm{d}y
\end{aligned}$$

for $n \to \infty$ and P_0-a.e. $\varphi \in \mathbb{M}_0$. The last line follows from an approximation result for the Radon–Nikodým derivative λ (Daley and Vere-Jones 2003, Lemma A1.6.III). □

Propositions 3 and 4 show that comparisons based on the Poisson log-likelihood (13) can be understood as approximations to a comparison of intensity forecasts with the scoring function (15). In particular, we can conclude that partitioning is not essential for model evaluation: A straightforward generalization of ELMT

relies on models that produce intensities $\lambda : \mathscr{X} \to (0, \infty)$ on the testing region, which can then be compared via consistent scoring functions (Sect. 3.3), with (15) giving one possible choice. However, in some situations partitioning might be desirable, e.g. when no explicit expression for the intensity is available. This also applies to our case study, where only the expected numbers per grid cell were produced by the forecasting models. In light of Proposition 4, our evaluation is essentially a comparison of the point process intensities forecasted by the four competing models.

6 Discussion

Assessing forecast accuracy and comparing the performance of several competing forecasts is a non-trivial task that poses challenges across disciplines and sectors. In this paper, we have demonstrated that consistent scoring functions allow for the comparative evaluation of point process forecasts. Our methods are complementary to the simulation-based approach of Heinrich-Mertsching et al. (2021), encompass existing techniques for model comparison, and yield a novel understanding of earthquake likelihood model testing. In particular, we have shown that the Poisson log-likelihood can be used for theoretically principled comparative forecast evaluation in terms of cell expectations. This is an important finding, as it supports current practice in comparisons between Poisson models, for which the interpretation in terms of log-likelihood is useful and welcome, and other types of models, which might generate cell expectations only. When one ignores the possibility of multiple events in a cell, the cell expectation equals the probability of an event, and we are in the setting studied by Serafini et al. (2022).

To conclude our study, we continue the discussion of methods for model comparison that are based on the log-likelihood, i.e. the model log-density evaluated at the observations, distinguish relative and absolute performance assessment, and hint at future work.

The *entropy score* considers the log-likelihood of probability forecasts induced by a point process model (Daley and Vere-Jones 2004; Harte and Vere-Jones 2005). It can be interpreted as an application of the logarithmic score to probabilistic predictions in terms of numbers of events. The expected value of the entropy score difference between a model of interest and a reference model yields the *information gain*. Daley and Vere-Jones (2004) note that the information gain is an inherent characteristic of a point process model that quantifies predictability and relates closely to entropy. A detailed discussion of the relationships between proper scoring rules, entropy, and divergences is available in Section 2.2 of Gneiting and Raftery (2007).

Information criteria such as AIC or BIC assess the relative quality of competing models, and can be applied to point process models, provided that densities are available, see e.g. Chen et al. (2018). They connect naturally to consistent scoring functions through their goodness-of-fit component, which usually consists of a log-likelihood and thus finds the logarithmic score (3) for the model at hand. The penalty component, which depends on the number of fitted parameters, is a necessary correction when operating in-sample, i.e. relying on the same data

as used for model fitting. In contrast, comparative forecast evaluation via scoring functions is tailored to out-of-sample settings, as in our case study.

In Bayesian settings, a standard approach to model comparison is the use of *Bayes factors* of a model vs. a competitor, as employed by Marzocchi et al. (2012) in earthquake likelihood model testing. Similar to information criteria, Bayes factors are closely connected to the logarithmic score (Gneiting and Raftery 2007, Section 7).

A further likelihood-based method for point processes uses deviance residuals, as proposed by Clements et al. (2011). In general, point process residuals form an empirical process arising from fitting a conditional intensity to data (Schoenberg 2003; Baddeley et al. 2005; Bray et al. 2014). Residuals can be used to assess goodness-of-fit and especially indicate in which regions a model fits well or poorly. Clements et al. (2011) propose a graphic comparison of models for the conditional intensity by plotting the log-likelihood ratio across a partition of the spatial domain, which can be interpreted as visualizing local differences in the logarithmic score.

Consistent scoring functions, as well as the just discussed methods, compare competing models or forecasts. This contrasts with many existing point process model evaluation tools, which focus on absolute performance, e.g. based on calibration (Thorarinsdottir 2013) and goodness-of-fit. Although this is important in model building, a selection among the available competitors has to be done eventually, and measures of absolute performance are not designed, and hence tend to be poorly positioned, for this task. Moreover, as pointed out by Nolde and Ziegel (2017), focusing on absolute performance may lead to misguided incentives in designing candidate models.

Earthquake likelihood model testing, a central element of the CSEP forecasting experiments, is tacitly based on strictly consistent scoring functions for expectations. A principled use of these functions, as illustrated in our case study, provides valid comparisons of forecasted intensities. Importantly, common assumptions in the context of CSEP tests are not needed for such an evaluation: Neither the forecasting models, nor the data, need to follow any Poisson or independence assumption, and with suitably adapted models, partitioning the testing region can be avoided. As these conclusions apply to intensity forecasts, a natural next step is to employ consistent scoring functions to compare earthquake forecasts in terms of other statistical properties. In particular, dependence properties or full distributions are natural candidates for forecast evaluation in the CSEP framework (Schorlemmer et al. 2018; Nandan et al. 2019). The choice and implementation of consistent scoring functions in settings of this type pose challenges for future work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00875-5.

Acknowledgements Jonas Brehmer and Tilmann Gneiting are grateful for support by the Klaus Tschira Foundation. Jonas Brehmer gratefully acknowledges support by the German Research Foundation (DFG) through Research Training Group RTG 1953. Part of this research came to fruition during mutual visits of Kirstin Strokorb at the University of Mannheim and Jonas Brehmer and Martin Schlather at Cardiff University during a workshop funded by the London Mathematical Society. We thank our hosting institutions for their generous hospitality. The authors would also like to thank Claudio Heinrich-Mertsching, Christopher Dörr and Alexander Jordan for helpful discussions, and Kristof Kraus for code review.

Likewise, we are grateful to the anonymous reviewers for their comments that helped improve the clarity of this paper.

References

Baddeley, A., Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software, 12*, 1–42.

Baddeley, A., Turner, R., Møller, J., Hazelton, M. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*, 617–666.

Baddeley, A., Rubak, E., Turner, R. (2015). *Spatial point patterns: Methodology and applications with R*. London: Chapman and Hall/CRC Press.

Bray, A., Schoenberg, F. P. (2013). Assessment of point process models for earthquake forecasting. *Statistical Science, 28*, 510–520.

Bray, A., Wong, K., Barr, C. D., Schoenberg, F. P. (2014). Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *Annals of Applied Statistics, 8*, 2247–2267.

Brehmer, J. R. (2021). A construction principle for proper scoring rules. *Proceedings of the American Mathematical Society Series B, 8*, 297–301.

Brehmer, J. R. (2023). Reproduction material for "Comparative evaluation of point process forecasts". Available at https://github.com/jbrehmer42/pp_evaluation.

Chen, J., Hawkes, A. G., Scalas, E., Trinh, M. (2018). Performance of information criteria for selection of Hawkes process models of financial data. *Quantitative Finance, 18*, 225–235.

Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. (2013). *Stochastic geometry and its applications* (3rd ed.). Chichester: Wiley.

Clements, R. A., Schoenberg, F. P., Schorlemmer, D. (2011). Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. *Annals of Applied Statistics, 5*, 2549–2571.

Daley, D. J., Vere-Jones, D. (2003). *An introduction to the theory of point processes* (2nd ed., Vol. I). NewYork: Springer.

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

Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron, 72*, 169–183.

Dawid, A. P., Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics, 27*, 65–81.

Diebold, F. X., Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*, 253–263.

Ehm, W., Gneiting, T., Jordan, A., Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 78*, 505–562.

Falcone, G., Console, R., Murru, M. (2010). Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST. *Annals of Geophysics, 53*, 41–50.

Field, E. H. (2007). Overview of the working group for the development of regional earthquake likelihood models (RELM). *Seismological Research Letters, 78*, 7–16.

Flaxman, S., Chirico, M., Pereira, P., Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-Time Crime Forecasting Challenge''. *Annals of Applied Statistics, 13*, 2564–2585.

Frongillo, R., Kash, I. A. (2015). Vector-valued property elicitation. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 40*, 1–18.

Frongillo, R., Kash, I. A. (2021). Elicitation complexity of statistical properties. *Biometrika, 108*, 857–879.

Gerstenberger, M. C., Wiemer, S., Jones, L. M., Reasenberg, P. A. (2005). Real-time forecasts of tomorrow's earthquakes in California. *Nature, 435*, 328–331.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*, 746–762.

Gneiting, T., Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*, 359–378.

Gneiting, T., Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics, 29*, 411–422.

Harte, D. (2015). Log-likelihood of earthquake models: evaluation of models and forecasts. *Geophysical Journal International, 201*, 711–723.

Harte, D., Vere-Jones, D. (2005). The entropy score and its uses in earthquake forecasting. *Pure and Applied Geophysics, 162*, 1229–1253.

Heinrich-Mertsching, C., Thorarinsdottir, T. L., Guttorp, P., Schneider, M. (2021). Validation of point process predictions with proper scoring rules. Preprint. arXiv:2110.11803.

Hendrickson, A. D., Buehler, R. J. (1971). Proper scores for probability forecasters. *Annals of Mathematical Statistics, 42*, 1916–1921.

Hering, A. S., Genton, M. G. (2011). Comparing spatial predictions. *Technometrics, 53*, 414–425.

Herrmann, M., Marzocchi, W. (2023). Maximizing the forecasting skill of an ensemble model. *Geophysical Journal International*. https://doi.org/10.1093/gji/ggad020

Holzmann, H., Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *Annals of Applied Statistics, 11*, 2404–2431.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. Chichester: Wiley.

Jordan, T. H., Chen, Y. T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., Zschau, J. (2011). Operational earthquake forecasting: State of knowledge and guidelines for utilization. *Annals of Geophysics, 54*, 4.

Kagan, Y. Y., Jackson, D. D. (1995). New seismic gap hypothesis: Five years after. *Journal of Geophysical Research: Solid Earth, 100*, 3943–3959.

Kagan, Y. Y., Knopoff, L. (1987). Statistical short-term earthquake prediction. *Science, 236*, 1563–1567.

Lavancier, F., Møller, J., Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 77*, 853–877.

Lehr, R. (1992). Sixteen *s*-squared over *d*-squared: A relation for crude sample size estimates. *Statistics in Medicine, 11*, 1099–1102.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science, 32*, 106–127.

Lombardi, A. M., Marzocchi, W. (2010). The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment. *Annals of Geophysics, 53*, 155–164.

Marzocchi, W., Zechar, J. D., Jordan, T. H. (2012). Bayesian forecast evaluation and ensemble earthquake forecasting. *Bulletin of the Seismological Society of America, 102*, 2574–2584.

Marzocchi, W., Lombardi, A. M., Casarotti, E. (2014). The establishment of an operational earthquake forecasting system in Italy. *Seismological Research Letters, 85*, 961–969.

Marzocchi, W., Taroni, M., Falcone, G. (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence. *Science Advances, 3*, e1701239.

Meyer, S., Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics, 8*, 1612–1639.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association, 106*, 100–108.

Nandan, S., Ouillon, G., Sornette, D., Wiemer, S. (2019). Forecasting the full distribution of earthquake numbers is fair, robust, and better. *Seismological Research Letters, 90*, 1650–1659.

Nolde, N., Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics, 11*, 1833–1874.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association, 83*, 9–27.

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

Ogata, Y., Katsura, K., Falcone, G., Nanjo, K., Zhuang, J. (2013). Comprehensive and topical evaluations of earthquake forecasts in terms of number, time, space, and magnitude. *Bulletin of the Seismological Society of America, 103*, 1692–1708.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science, 33*, 299–318.

Rhoades, D., Schorlemmer, D., Gerstenberger, M., Christophersen, A., Zechar, J. D., Imoto, M. (2011). Efficient testing of earthquake forecasting models. *Acta Geophysica, 59*, 728–747.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.
Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association, 66*, 783–801.
Schoenberg, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association, 98*, 789–795.
Schoenberg, F. P., Hoffmann, M., Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics, 71*, 1271–1287.
Schorlemmer, D., Gerstenberger, M. C., Wiemer, S., Jackson, D. (2007). Earthquake likelihood model testing. *Seismological Research Letters, 78*, 17–29.
Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P. J., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., Zhuang, J. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters, 89*, 1305–1313.
Serafini, F., Naylor, M., Lindgren, F., Werner, M. J., Main, I. (2022). Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment. *Geophysical Journal International, 230*, 1419–1440.
Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M. J., Wiemer, S., Zechar, J. D., Heiniger, L., Euchner, F. (2018). Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy. *Seismological Research Letters, 89*, 1251–1261.
Thorarinsdottir, T. L. (2013). Calibration diagnostic for point process models via the probability integral transform. *Stat, 2*, 150–158.
van Belle, G. (2008). *Statistical rules of thumb. Wiley series in probability and statistics* (2nd ed.). Chichester: Wiley.
Woessner, J., Christophersen, A., Zechar, J. D., Monelli, D. (2010). Building self-consistent, short-term earthquake probability (STEP) models: Improved strategies and calibration procedures. *Annals of Geophysics, 53*, 141–154.
Zechar, J. D., Gerstenberger, M. C., Rhoades, D. A. (2010a). Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts. *Bulletin of the Seismological Society of America, 100*, 1184–1195.
Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J., Jordan, T. H. (2010b). The collaboratory for the study of earthquake predictability perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience, 22*, 1836–1847.
Zhuang, J., Mateu, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 182*, 919–942.
Zhuang, J., Ogata, Y., Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association, 97*, 369–380.

Authors and Affiliations

Jonas R. Brehmer[1] · Tilmann Gneiting[1,2] · Marcus Herrmann[3] · Warner Marzocchi[3] · Martin Schlather[4] · Kirstin Strokorb[5]

✉ Jonas R. Brehmer
jonas.brehmer@h-its.org

Tilmann Gneiting
tilmann.gneiting@h-its.org

Marcus Herrmann
marcus.herrmann@unina.it

Warner Marzocchi
warner.marzocchi@unina.it

Martin Schlather
schlather@math.uni-mannheim.de

Kirstin Strokorb
strokorbk@cardiff.ac.uk

[1] Computational Statistics Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

[2] Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Englerstraße 2, 76131 Karlsruhe, Germany

[3] Department of Earth, Environmental, and Resources Sciences, University of Naples Federico II, Complesso di Monte Sant'Angelo, Via Vicinale Cupa Cintia, 21, 80126 Naples, Italy

[4] Institute of Mathematics, University of Mannheim, B 6, 26, 68159 Mannheim, Germany

[5] School of Mathematics, Cardiff University, Senghennydd Road, Cardiff CF24 4AG, UK