Supplement to: Comparative evaluation of point process forecasts

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

[1]Heidelberg Institute for Theoretical Studies (HITS)
[2]Institute for Stochastics, Karlsruhe Institute of Technology (KIT)
[3]Department of Earth Sciences, University of Naples, Federico II
[4]Institute for Mathematics, University of Mannheim
[5]School of Mathematics, Cardiff University

April 12, 2023

Contents

S1 Discussion of point process scenarios

This section extends the discussion at the end of Section 2.

In the main manuscript we focus on the setting where a spatial point process Φ on some domain $\mathcal{X} \subset \mathbb{R}^d$ is observed at fixed points in time. For example, the case study (Section 5) considers daily observations of locations of earthquakes in Italy. However, forecasting for point processes appears in a variety of other situations, and the use of strictly consistent scoring functions adapts readily. To clarify this idea, we distinguish three different point process scenarios. Although motivated by commonly encountered applications, there might be settings where the distinction is artificial.

Scenario A (purely spatial) In this scenario, the process is defined on either a single spatial domain (Scenario A1), or several non-overlapping subdomains (Scenario A2). Examples include points fixated by observers of images (Barthelmé et al., 2013) and locations of trees in a forest (Stoyan and Penttinen, 2000). Stationarity is a common simplifying assumption in this context.

Scenario B (purely temporal) In this scenario, there is no spatial component and the process concerns points in time only. Examples are arrival times of e-mails (Fox et al., 2016) and times of infection with a disease (Schoenberg et al., 2019). In this special setting the directional character of time allows for a distinct interpretation and treatment.

Scenario C (spatio-temporal) In addition to the spatial component, processes in this scenario possess a temporal component, which could be discrete (Scenario C1) or continuous (Scenario C2). Examples include locations and times of crimes in a city (Mohler et al., 2011) and earthquakes observed over time in a specific region (Ogata, 1998; Zhuang et al., 2002). The main manuscript focuses on Scenario C1.

In order to compare forecasts in each of these scenarios, we can in principle proceed as in Sections 4 and 5: Choose a strictly consistent scoring function S for a statistical property of point processes, e.g. the intensity, and find the mean score difference

$$\frac{1}{n}\sum_{i=1}^{n}\left(S(r_i,\varphi_i) - S(r_i^*,\varphi_i)\right)$$

for forecast reports r_i and r_i^* and associated observed point patterns φ_i, where the index $i = 1,\ldots,n$ represents repeated observations. Then negative values support forecast r, while positive values support r^*. The mean score difference is an estimator of the expected score difference $\mathbb{E}\left(S(r,\Phi) - S(r^*,\Phi)\right)$, and implementation details vary across scenarios, also impacting the assessment of the uncertainty inherent in the estimate, which is of particular importance when tests for superior predictive performance are sought. To illustrate the key ideas we distinguish whether the point process has a continuous or discrete time component.

Discrete time Assume that the point process is sampled at fixed points in time, i.e. it can be modelled by a sequence $(\Phi_t)_{t\in\mathbb{N}}$ adapted to a filtration $(\mathcal{H}_t)_{t\in\mathbb{N}}$. This setting includes the special case of i.i.d. realizations and relates to Scenario C1 as well as variants of Scenario A with repeated observations. Given two forecast sequences $(R_t)_{t\in\mathbb{N}}$ and $(R_t^*)_{t\in\mathbb{N}}$ the score differences $(S(R_t,\Phi_t) - S(R_t^*,\Phi_t))_{t\in\mathbb{N}}$ form a sequence of real-valued random variables, thus the common Diebold–Mariano (DM) tests (Diebold and Mariano, 1995) are directly applicable. We briefly discuss the more general forecast comparison framework of Nolde and Ziegel (2017) in our setting. Let S be strictly consistent for a point process statistic $\Gamma : \mathcal{P} \to \mathsf{A}$ and assume that forecasts in terms of Γ applied to the conditional distribution $\Phi_t \mid \mathcal{H}_{t-1}$ are given. These forecasts can be regarded as random sequences $R = (R_t)_{t\in\mathbb{N}}$ and $R^* = (R_t^*)_{t\in\mathbb{N}}$ such that R_t and R_t^* are $\mathcal{H}_{t-1}$-measurable. Their forecast performance can be compared via the *mean score difference*

$$\Delta_n(R,R^*) := \frac{1}{n}\sum_{t=1}^{n} S(R_t,\Phi_t) - \frac{1}{n}\sum_{t=1}^{n} S(R_t^*,\Phi_t) = \frac{1}{n}\sum_{t=1}^{n}\left(S(R_t,\Phi_t) - S(R_t^*,\Phi_t)\right), \quad (1)$$

which is an estimator for the difference in expected scores. Based on the law of large numbers and the strict consistency of S, a positive value supports the hypothesis that R^* is superior to R, while a negative value supports the opposite hypothesis. A further step is to test whether $\Delta_n(R,R^*)$ is significantly different from zero. In the simple situation of an i.i.d. sequence $(\Phi_t)_{t\in\mathbb{N}}$, the forecast sequences reduce to $r, r^* \in \mathsf{A}$, i.e. they are constant in time. We can then test for significant differences in expected scores based on the asymptotic normality of the well-known t-statistic $t_n := \sqrt{n}\Delta_n(r,r^*)/\sqrt{\hat{\sigma}_n^2}$, where $\hat{\sigma}_n^2$ estimates the variance of $S(r,\Phi) - S(r^*,\Phi)$. For dependent time series $(\Phi_t)_{t\in\mathbb{N}}$, $(R_t)_{t\in\mathbb{N}}$, and $(R_t^*)_{t\in\mathbb{N}}$ we refer to Nolde and Ziegel (2017), where tests for equal forecast performance rely on suitable asymptotic results developed in Giacomini and White (2006).

Continuous time If we consider point processes in Scenario C2 or Scenario B, then temporal dependence between the points of Φ becomes an essential feature of the process and can also be object of the forecast. For instance, the statistic Γ might consist of temporal features of the point process. Also, dependencies need to be accounted for in estimation and testing, as they affect asymptotic distributions. To illustrate this, assume for simplicity that Φ is a purely temporal process observed over a time period $[0,T]$ with $0 < t_1 < \cdots < t_k < T$ denoting the corresponding arrival times. Moreover, let R_i and R_i^* be reports issued at time t_{i-1} based on the previous arrivals $t_1, \dots, t_{i-1}$. This yields a realized score difference

$$\Delta_T(R, R^*) = \sum_{i=1}^{n(T)} \left(S(R_i, t_i) - S(R_i^*, t_i) \right), \tag{2}$$

where $n(T) := \Phi((0, T])$ is the random number of points in $[0, T]$. In contrast to (1) we do not consider averages since $n(T)$ is a random variable depending on Φ and dividing by it will interfere with the consistency of S. The score difference $\Delta_T(R, R^*)$ is a sum of a random number of random variables, usually called a random sum. This perspective connects the estimation of score differences to the theory of total claim amount in insurance, see e.g. Mikosch (2009) and Embrechts et al. (1997).

Asymptotic results for the score difference (2) for $T \to \infty$ are desirable to assess how uncertainty affects forecast evaluation and transfer the DM test to the continuous time setting. One possible approach to this problem relies on limit theorems for randomly indexed processes due to Anscombe (1952), in particular random central limit theorems: If the number of points $n(T)$ satisfies a weak law of large numbers, then under Anscombe's condition, we only need to ensure that the sequence $(S(R_i, t_i) - S(R_i^*, t_i))_{i \in \mathbb{N}}$ satisfies a central limit theorem in order to obtain asymptotic normality for (2). Such results are available for strong mixing (Lee, 1997), ψ-weakly dependent (Hwang and Shin, 2012), and m-dependent (Shang, 2012) sequences. Working these into tests for superior forecast performance for (spatio-)temporal point processes is an avenue for future work.

S2 Further scoring functions for point processes

The technical context of this section is the same as in Section 3.

S2.1 Simple examples

The subsequent examples are applications of the transformation principle (Proposition 1).

Example S1 (void probability)**.** For any fixed set $B \in \mathcal{B}(\mathcal{X})$ the functional Γ defined via $\Gamma(P) = P(\{\varphi \mid \varphi \cap B = \emptyset\})$ is elicitable. This follows from Proposition 1 with $T(F) = \mathbb{E}_F Y$ and $g(\varphi) = \mathbb{1}(\varphi(B) = 0)$. Strictly consistent scoring functions for Γ are of the Bregman form (2), see also Example 1.

Example S2 (point process integrals)**.** Fix measurable functions $f_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \dots, m$ for $m \in \mathbb{N}$. Define $g : \mathbb{M}_0 \to \mathbb{R}^m$ via

$$g(\varphi) = \left(\int_{\mathcal{X}} f_1 \, \mathrm{d}\varphi, \dots, \int_{\mathcal{X}} f_m \, \mathrm{d}\varphi \right)^\top = \left(\sum_{x_i \in \varphi} f_1(x_i), \dots, \sum_{x_i \in \varphi} f_m(x_i) \right)^\top,$$

set $g(\mathcal{P}) := \{P \circ g^{-1} \mid P \in \mathcal{P}\}$ and let $T = \mathrm{id}_{g(\mathcal{P})}$. Then the finite-dimensional distribution functional $\Gamma_{f_1,\dots,f_m}(P) = T(P \circ g^{-1})$ is an elicitable property of the point process Φ. Consistent scoring functions for Γ are obtained by applying consistent scoring functions for distributions (Gneiting and Raftery, 2007) to the m-variate distribution $P \circ g^{-1}$, see also Heinrich-Mertsching et al. (2021).

S2.2 Distribution and density

This material extends Section 3.2.

General result for the full distribution The law P_Φ of a finite point process on $\mathcal{X}$ can be equivalently represented by two 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. The Π_k are symmetric probability measures on $\mathcal{X}^k$ which describe the distribution of any ordering of points, given k points are realized, see Daley and Vere-Jones (2003, Chapter 5.3) for details.

To state the next result, we introduce the notion of *symmetric* scoring functions, where $S : \mathsf{A} \times \mathbb{R}^n \to \mathbb{R}$ is called symmetric if $S(a, y_1, \dots, y_n) = S(a, y_{\pi(1)}, \dots, y_{\pi(n)})$ for all $a \in \mathsf{A}$, $y \in \mathbb{R}^n$ and permutations π. Symmetry ensures that the scoring functions in the subsequent proposition are independent of the enumeration of the realization of Φ.

Proposition S1. *Let $\mathcal{P}$ be a class of distributions of finite point processes, with $Q \in \mathcal{P}$ decomposed into $(\Pi_k^Q)_{k \in \mathbb{N}}$ and $(p_k^Q)_{k \in \mathbb{N}_0}$. Set $\mathcal{F}_k := \{\Pi_k^Q \mid Q \in \mathcal{P}\}$ and let $S_k : \mathcal{F}_k \times \mathcal{X}^k \to \mathbb{R}$ be a symmetric consistent scoring function for $\mathrm{id}_{\mathcal{F}_k}$ for all $k \in \mathbb{N}$. Let S_0 be a consistent scoring function for distributions on $\mathbb{N}_0$. Then the function $S : \mathcal{P} \times \mathbb{M}_0 \to \mathbb{R}$ defined via*

$$S(((\Pi_k^Q)_{k \in \mathbb{N}}, (p_k^Q)_{k \in \mathbb{N}_0}), \{y_1, \dots, y_n\}) = S_n(\Pi_n^Q, y_1, \dots, y_n) + S_0((p_k^Q)_{k \in \mathbb{N}_0}, n)$$

for $n \in \mathbb{N}$ and $S(((\Pi_k^Q)_{k \in \mathbb{N}}, (p_k^Q)_{k \in \mathbb{N}_0}), \emptyset) := S_0((p_k^Q)_{k \in \mathbb{N}_0}, 0)$ is a consistent scoring function for the distribution of the point process Φ. It is strictly consistent if S_0 and $(S_k)_{k \in \mathbb{N}}$ are strictly consistent.

Proof. The result follows by decomposing the expectation $\mathbb{E}_P S(Q, \Phi)$ into expectations on the sets $\{\Phi = n\}$ for $n \in \mathbb{N}$ and using the (strict) consistency of S_n on each set. □

Hyvärinen score Assume that a point process model admits explicit expressions for the Janossy densities $(j_k)_{k\in\mathbb{N}_0}$ (see Section 3.2), however, only up to an unknown normalizing constant. In this situation, 0-homogeneous consistent scoring functions for densities can be of use, as they allow for the consistent evaluation of an unnormalized density. The most relevant example is the *Hyvärinen score* defined via

$$\mathrm{HyvS}(f,y) := \Delta \log f(y) + \frac{1}{2}\|\nabla \log f(y)\|^2,$$

where ∇ denotes the gradient, Δ is the Laplace operator, and f is a twice differentiable density on $\mathbb{R}^d$. To ensure strict consistency on a class of probability densities $\mathcal{L}$ its members have to be positive almost everywhere and for all $f,g \in \mathcal{L}$ it must hold that $\nabla \log(f(y))g(y) \to 0$ as $\|y\| \to \infty$, see Hyvärinen (2005), Parry et al. (2012), and Ehm and Gneiting (2012) for details.

Similar to the logarithmic score, we can transfer the Hyvärinen score to the point process setting. To do this we assume that for all $Q \in \mathcal{P}$ and $k \in \mathbb{N}$, j_k^Q is defined on $(\mathbb{R}^d)^k$ and satisfies the aforementioned regularity conditions. Then the function $S : \mathcal{P} \times \mathbb{M}_0 \to \mathbb{R}$ defined via

$$S((j_k^Q)_{k\in\mathbb{N}_0}, \{y_1,\dots,y_n\}) = \mathrm{HyvS}(j_n^Q, y_1,\dots,y_n) \tag{3}$$

for $n \in \mathbb{N}$ and $S((j_k^Q)_{k\in\mathbb{N}_0}, \emptyset) := 0$ is a consistent scoring function for the distribution of the point process Φ. Observe that we cannot achieve strict consistency for S, since the probability of $|\Phi| = n$ is proportional to j_n and thus not accessible to the Hyvärinen score.

Example S3 (Gibbs point process)**.** Stemming from theoretical physics, Gibbs processes are a popular tool to model particle interactions. They are defined via their Janossy densities

$$j_n(y_1,\dots,y_n) = C(\theta)\exp\left(-\theta U(y_1,\dots,y_n)\right),$$

where U represents point interactions, θ is a parameter often referred to as temperature, and C is the partition function, which ensures that the collection $(j_k)_{k\in\mathbb{N}_0}$ is properly normalized, see e.g. Daley and Vere-Jones (2003, Chapter 5.3) and Chiu et al. (2013, Chapter 5.5). It is in general difficult to find closed form expressions for C, or even to approximate it, hence the Hyvärinen score might seem attractive to evaluate models based on $(j_k)_{k\in\mathbb{N}_0}$. Plugging j_n into (3) gives

$$S((j_k)_{k\in\mathbb{N}_0}, \{y_1,\dots,y_n\}) = \theta\left(-\Delta U(y_1,\dots,y_n) + \frac{\theta}{2}\|\nabla U(y_1,\dots,y_n)\|^2\right)$$

for $n \in \mathbb{N}$, where the derivatives are computed with respect to the coordinates of the vector $(y_1, \ldots, y_n) \in (\mathbb{R}^d)^n$. The simplest choice for interactions is to restrict U to first- and second-order terms

$$U(y_1, \ldots, y_n) := \sum_{i=1}^{n} l(y_i) + \sum_{i,j=1}^{n} \psi\left(\|y_i - y_j\|^2\right)$$

for $l : \mathbb{R}^d \to \mathbb{R}$ and $\psi : [0, \infty) \to [0, \infty)$ with $\psi(0) = 0$, see e.g. Daley and Vere-Jones (2003, Chapter 5.3). To apply the Hyvärinen score in this setting, l and ψ have to satisfy regularity conditions detailed above and in Hyvärinen (2005), and in particular admit second order derivatives almost everywhere. The soft-core models for ψ introduced in Ogata and Tanemura (1984) satisfy this condition, while their hard-core model for ψ is not even continuous. An additional technical issue is that Ogata and Tanemura (1984) consider point processes on a finite domain $\mathcal{X}$ and use a constant l. To make the Hyvärinen score applicable in this setting a possible solution is to approximate their models via twice differentiable densities on $(\mathbb{R}^d)^n$.

S2.3 Moment measures

Moment measures can be interpreted as the point process analogue to the moments of a univariate random variable. Strictly consistent scoring functions for these measures can be constructed in the same way as for the intensity, see Proposition 3.4.

For $n \in \mathbb{N}$, let $\mathcal{M}_{\mathrm{f}}^n = \mathcal{M}_{\mathrm{f}}(\mathcal{X}^n)$ be the set of finite Borel measures on $\mathcal{X}^n$. For positive measurable functions $f : \mathcal{X}^n \to (0, \infty)$ the *n-th moment measure* $\mu^{(n)}$ and the *n-th factorial moment measure* $\alpha^{(n)}$ are defined via the relations

$$\mathbb{E}\left(\sum_{x_1,\ldots,x_n \in \Phi} f(x_1, \ldots, x_n)\right) = \int_{\mathcal{X}^n} f(x_1, \ldots, x_n)\, \mathrm{d}\mu^{(n)}(x_1, \ldots, x_n),$$

and

$$\mathbb{E}\left(\sum_{x_1,\ldots,x_n \in \Phi}^{\neq} f(x_1, \ldots, x_n)\right) = \int_{\mathcal{X}^n} f(x_1, \ldots, x_n)\, \mathrm{d}\alpha^{(n)}(x_1, \ldots, x_n),$$

respectively, see e.g. Chiu et al. (2013) and Daley and Vere-Jones (2003). Here $\Sigma^{\neq}$ denotes summation over all n-tuples that contain distinct points of Φ. Using the notion of *factorial product* defined via

$$m^{[n]} := \begin{cases} m(m-1)(m-2)\cdots(m-n+1) & , m \geq n \\ 0 & , m < n \end{cases}$$

for $m, n \in \mathbb{N}$ we obtain the concise representations $\mu^{(n)}(B^n) = \mathbb{E}\Phi(B)^n$ and $\alpha^{(n)}(B^n) = \mathbb{E}\Phi(B)^{[n]}$ for Borel sets $B \in \mathcal{B}(\mathcal{X})$, see e.g. Daley and Vere-Jones (2003, Chapter 5).

Proposition S2. *Set $\mathcal{F}^n := \{P^* \mid P \in \mathcal{M}_{\mathrm{f}}^n\}$, let $S : \mathcal{F}^n \times \mathcal{X}^n \to \mathbb{R}$ be a consistent scoring function for $\mathrm{id}_{\mathcal{F}^n}$ and $b : [0,\infty) \times [0,\infty) \to \mathbb{R}$ a Bregman function.*

(i) The function $S_1 : \mathcal{M}_{\mathrm{f}}^n \times \mathbb{M}_0 \to \mathbb{R}$ defined via

$$S_1(\mu, \{y_1, \dots, y_m\}) = \sum_{x_1,\dots,x_n \in \{y_1,\dots,y_m\}} S(\mu^*, x_1, \dots, x_n) + cb(\mu(\mathcal{X}^n), m^n)$$

for $m \in \mathbb{N}$, and $S_1(\mu, \emptyset) = cb(\mu(\mathcal{X}^n), 0)$ for $c > 0$, is a consistent scoring function for the n-th moment measure.

(ii) The function $S_2 : \mathcal{M}_{\mathrm{f}}^n \times \mathbb{M}_0 \to \mathbb{R}$ defined via

$$S_2(\alpha, \{y_1, \dots, y_m\}) = \sum_{x_1,\dots,x_n \in \{y_1,\dots,y_m\}}^{\neq} S(\alpha^*, x_1, \dots, x_n) + cb(\alpha(\mathcal{X}^n), m^{[n]})$$

for $m \geq n$ and $S_2(\alpha, \{y_1, \dots, y_m\}) = cb(\alpha(\mathcal{X}^n), 0)$ for $m < n$ and with $c > 0$ is a consistent scoring function for the n-th factorial moment measure.

Both S_1 and S_2 are strictly consistent if S is strictly consistent and b is strict.

In many cases of interest $\alpha^{(n)}$ is absolutely continuous with respect to Lebesgue measure on $\mathcal{X}^n$ and its density $\varrho^{(n)}$ is called *product density*, see e.g. Chiu et al. (2013). A (strictly) consistent scoring function for $\varrho^{(n)}$ can be obtained from Proposition S2 (ii) by choosing S to be a (strictly) consistent scoring function for densities.

Example S4. Let $n = 2$ and for simplicity consider the product density $\varrho^{(2)}$ of a stationary and isotropic point process. In this situation, $\varrho^{(2)}$ depends on the point distances only, i.e. it can be represented via $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$ for some $\varrho_0^{(2)} : [0,\infty) \to [0,\infty)$. Analogous to Example 4, we can use the quadratic score for b and the logarithmic score for S in Proposition S2 (ii). This gives the strictly consistent scoring function

$$\begin{aligned} S(\varrho^{(2)}, \{y_1, \dots, y_m\}) = &- \sum_{x_1,x_2 \in \{y_1,\dots,y_m\}}^{\neq} \log(\varrho_0^{(2)}(\|x_1 - x_2\|)) \\ &+ m^{[2]} \log|\varrho^{(2)}| + c\,(|\varrho^{(2)}| - m^{[2]})^2, \end{aligned}$$

where $c > 0$ is some scaling constant. Simulation experiments in Section S3.2 show how S compares different product density forecasts.

S2.4 Summary statistics

Summary statistics of point processes are central tools to quantify point interactions such as clustering or inhibition. This subsection constructs strictly consistent scoring functions for the frequently used K-function. Throughout we assume that Φ is a *stationary* point process on $\mathbb{R}^d$, i.e. any translation of the process by $x \in \mathbb{R}^d$, which we denote via Φ_x, has the same distribution as Φ. This implies that the intensity measure of Φ is a multiple of Lebesgue measure and can be represented via some $\lambda > 0$, see e.g. Chiu et al. (2013, Chapter 4.1).

A common way to describe a stationary point process is to consider its properties in the neighbourhood of $x \in \mathbb{R}^d$, given that x is a point in Φ. Due to stationarity, the location of x is irrelevant and thus it is usually referred to as the "typical point" of Φ. The technical tool to describe the behaviour around this point is the *Palm distribution* of Φ, denoted via $\mathbb{P}_0$ for probabilities and $\mathbb{E}_0$ for expectations. It satisfies the defining identity

$$\lambda \, |W| \, \mathbb{E}_0 f(\Phi) = \mathbb{E}\left(\sum_{x \in \Phi \cap W} f(\Phi_{-x}) \right)$$

for all measurable functions $f : \mathbb{M}_0 \to \mathbb{R}$ such that the expectations are finite, and it is independent of the observation window $W \in \mathcal{B}(\mathbb{R}^d)$ (Illian et al., 2008, Chapter 4). When we need to highlight the distribution of the point process, we write $\mathbb{E}_{P,0}$ for the Palm expectation given Φ has distribution $P \in \mathcal{P}$.

Denote the d-dimensional ball of radius $r > 0$ around zero via $B_r = B(0, r)$. The *K-function* of Φ is defined via

$$K : (0, \infty) \to [0, \infty), \quad r \mapsto \frac{\mathbb{E}_0 \Phi \left(B_r \backslash \{0\} \right)}{\lambda},$$

and it quantifies the mean number of points in a ball around the "typical point" of Φ, see e.g. Chiu et al. (2013) and Illian et al. (2008) for details. Deriving strictly consistent scoring functions for the K-function appears challenging since it combines the Palm distribution and the intensity. However, in many situations both of these quantities are of interest. We thus derive a result which defines scoring functions for joint reports of the K-function and the intensity. Our point process property of interest is thus $\Gamma(P) := (\lambda_P, K_P)$, where the subscript denotes the dependence of the quantities on the distribution $P \in \mathcal{P}$ of the process Φ. Since observation windows are always finite, we fix some $r^* > 0$ and let K_P be the restriction of the K-function to the interval $(0, r^*)$.

To derive consistent scoring functions let us fix some $r \in (0, r^*)$ and assume for now that λ_P is known and that instead of data we directly observe the Palm distribution of Φ. In this simplified situation, $K_P(r)$ is just an expectation with respect to $\mathbb{P}_0$, hence "consistent scoring functions" for it are of the Bregman form

$$S(x, \varphi) = -f(\lambda_P x) - f'(\lambda_P x)\big(\varphi(B_r \backslash \{0\}) - \lambda_P x\big), \tag{4}$$

for a convex function $f : (0, \infty) \to \mathbb{R}$, see Theorem 1 and Example 1. This is because $\mathbb{E}_{P,0} b(x, \Phi) \geq \mathbb{E}_{P,0} b(K_P(r), \Phi)$ holds for all $x \geq 0$ and $P \in \mathcal{P}$. To arrive at a strictly consistent scoring function for the functional Γ three steps remain: Firstly, we have to include a consistent scoring function for the first component of Γ, i.e. the intensity. Moreover, we need to integrate (4) with respect to r in order to evaluate the K-function on the entire interval $(0, r^*)$. Finally, we have to account for the fact that we can not observe $\mathbb{P}_0$, but only points of Φ on some closed and bounded observation window $W \subset \mathbb{R}^d$. Hence, we need to compute the expected score $\mathbb{E}_0 S(x, \Phi)$ via an expectation of Φ on W. Such problems lead to edge corrections, i.e. additional terms to account for the fact that (unobserved) points outside of W affect the estimation near the boundary of W, see e.g. Chiu et al. (2013, Chapter 4.7) for details. Since (4) is linear in φ, edge corrections for the expected score are equivalent to edge corrections for the expectation $\mathbb{E}_0 \Phi(B_r \backslash \{0\})$, which are well-known in the context of K-function estimation. Before we formalize these three steps in a proposition, we state a result needed for the proof, see Gneiting (2011, Theorem 4).

Lemma S1 (revelation principle)**.** *Let A, A' be some sets and $g : \mathsf{A} \to \mathsf{A}'$ a bijection with inverse g^{-1}. Let $T : \mathcal{F} \to \mathsf{A}$ and $T_g : \mathcal{F} \to \mathsf{A}'$ defined via $T_g(F) := g(T(F))$ be functionals. Then T is elicitable if and only if T_g is elicitable. A function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is a (strictly) consistent scoring function for T if and only if $S_g : \mathsf{A}' \times \mathsf{O} \to \mathbb{R}$, $(x, y) \mapsto S_g(x, y) := S(g^{-1}(x), y)$ is a (strictly) consistent scoring function for T_g.*

Proposition S3. *Let $b_1, b_2 : [0, \infty) \times [0, \infty) \to \mathbb{R}$ be Bregman functions and $w : (0, \infty) \to [0, \infty)$ a weight function. Define $\mathcal{C} := \{K_P \mid P \in \mathcal{P}\}$, a set of possible K-functions, and let κ satisfy $\mathbb{E}_P \kappa(B_r, \Phi \cap W) = \lambda_P \mathbb{E}_{P,0} \Phi(B_r \backslash \{0\})$ for all $P \in \mathcal{P}$ and $r \in (0, r^*)$. Then the function $S : ((0, \infty) \times \mathcal{C}) \times \mathbb{M}_0 \to \mathbb{R}$ defined via*

$$S((\lambda, K), \varphi) = b_1(\lambda, \varphi(W)|W|^{-1}) + \int_0^{r^*} b_2(\lambda^2 K(r), \kappa(B_r, \varphi)) w(r) \, \mathrm{d}r$$

is consistent for the point process property $\Gamma(P) := (\lambda_P, K_P)$, where the second component is restricted to $(0, r^*)$. It is strictly consistent if b_1 and b_2 are strict and w is strictly positive.

Proof. Using Theorem 1, the Fubini-Tonelli theorem, and

$$\mathbb{E}_P \kappa(B_r, \Phi) = \lambda_P \mathbb{E}_{P,0} \Phi(B_r \backslash \{0\}) = \lambda_P^2 K_P(r)$$

for $r \in (0, r^*)$, standard arguments show that the scoring function

$$S'((\lambda, h), \varphi) := b_1(\lambda, \varphi(W)|W|^{-1}) + \int_0^{r^*} b_2(h(r), \kappa(B_r, \varphi))w(r)\,\mathrm{d}r,$$

where $h : (0, \infty) \to (0, \infty)$ is an increasing function, is consistent for the property $\Gamma'(P) := (\lambda_P, \lambda_P^2 K_P(r))$. An application of the revelation principle (Lemma S1) gives (strict) consistency for Γ. □

Similar to Proposition 2, this result blends two scoring components, namely the expected number of points and their distances. Hence, choosing suitable Bregman functions b_1 and b_2 in applications, again leads to issues of balancing the magnitudes of different scoring components. A similarly intricate question is the choice of κ. Relevant choices result from the construction of estimators for the K-function, which are often based on dividing κ by an estimator for λ^2. A common choice is

$$\kappa_{\mathrm{st}}(B_r, \varphi) := \sum_{x_1, x_2 \in \varphi \cap W}^{\neq} \frac{\mathbb{1}_{B_r}(x_2 - x_1)}{|W_{x_1} \cap W_{x_2}|},$$

where $W_z := \{x + z \mid x \in W\}$ is the shifted observation window and r is such that $|W \cap W_z|$ is positive for all $z \in B_r$, see e.g. Illian et al. (2008, Chapter 4.3) and Chiu et al. (2013, Chapter 4.7). An alternative arises via minus-sampling, i.e. by reducing the observation window W in order to reduce edge effects. This yields

$$\kappa_{\mathrm{minus}}(B_r, \varphi) := \frac{1}{|W|} \sum_{x_1, x_2 \in \varphi \cap W,\, x_2 \in W \ominus r}^{\neq} \mathbb{1}_{B_r}(x_2 - x_1),$$

where $W \ominus r := \{x \mid B(x, r) \subset W\}$ is the reduced observation window and $r < \mathrm{diam}(W)/2$. For other choices of κ, most notably for isotropic point processes, see Chiu et al. (2013, Chapter 4.7).

Practitioners usually rely on the L-function, a modification of the K-function, which is defined via $L(r) = \sqrt[d]{K(r)/\beta_d}$ for $r \geq 0$, where $\beta_d := |B_1|$. It satisfies $L(r) = r$ for the Poisson point process, and thus normalizes the K-function such that it is independent of the dimension d for a Poisson point process (Chiu et al., 2013). A (strictly) consistent scoring function for the L-function follows immediately from Proposition S3 and another application of the revelation principle. The explicit formula follows by replacing the first component of b_2 by $\lambda^2 L(r)^d \beta_d$ in Proposition S3. The idea underlying the construction of scoring functions for the K- and L-function presented here can be transferred to other summary statistics for stationary point processes.

S3 Extended simulation study

S3.1 Intensity

This subsection extends Section 4. We give more details on the used point processes and provide a closer analysis of the simulation experiments in the main paper. We then perform additional simulations with a different scoring function and study the approximation derived in Proposition 4.

All experiments rely on the six intensities defined in Section 4, see Figure S1 for an illustration. We consider two strictly consistent scoring functions for the intensity. The first choice is used in Section 4 and given by

$$S_1(\Lambda, \{y_1, \dots, y_n\}) = -\sum_{i=1}^{n} \log(\lambda(y_i)) + n \log |\Lambda| + c\,(|\Lambda| - n)^2, \tag{5}$$

see also Example 3.5. Our second choice is

$$S_2(\Lambda, \{y_1, \dots, y_n\}) = -\sum_{i=1}^{n} \log \lambda(y_i) + \int_{\mathcal{X}} \lambda(y)\,\mathrm{d}y, \tag{6}$$

which is defined in Proposition 5.1 and appears as the limit scoring function in earthquake likelihood model testing, see Section 5.3. The scaling factor $c > 0$ in (5) is set to $c = 1/10$. We draw $N = 100$ i.i.d. samples and repeat $M = 500$ times.

Details on the point process models We consider four different data-generating processes for Φ on $[0,1]^2$, all of which have (approximate) intensity $f_0(x,y) = 6\sqrt{x^2+y^2}$. The models are specified as follows:

1. An inhomogeneous Poisson point process with intensity f_0.
2. A thinned Gaussian determinantal point process (DPP), see e.g. Hough et al. (2006) and Lavancier et al. (2015). In general, a DPP is a locally finite point process with product densities (see Section S2.3) given by

   $$\varrho^{(n)}(x_1, \dots, x_n) = \det\left(C(x_i, x_j)\right)_{i,j=1,\dots,n}$$

   for $n \in \mathbb{N}$, where $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a covariance. As a result, the DPP's intensity function is $x \mapsto C(x,x)$ and it is stationary and isotropic whenever its covariance is. We choose $C(x_1, x_2) = C_0(\|x_1 - x_2\|)$, where $C_0 : [0, \infty) \to \mathbb{R}$ is the Gaussian covariance function

   $$C_0(r) = \sigma^2 \exp\left\{-\left(\frac{r}{s}\right)^2\right\}, \tag{7}$$

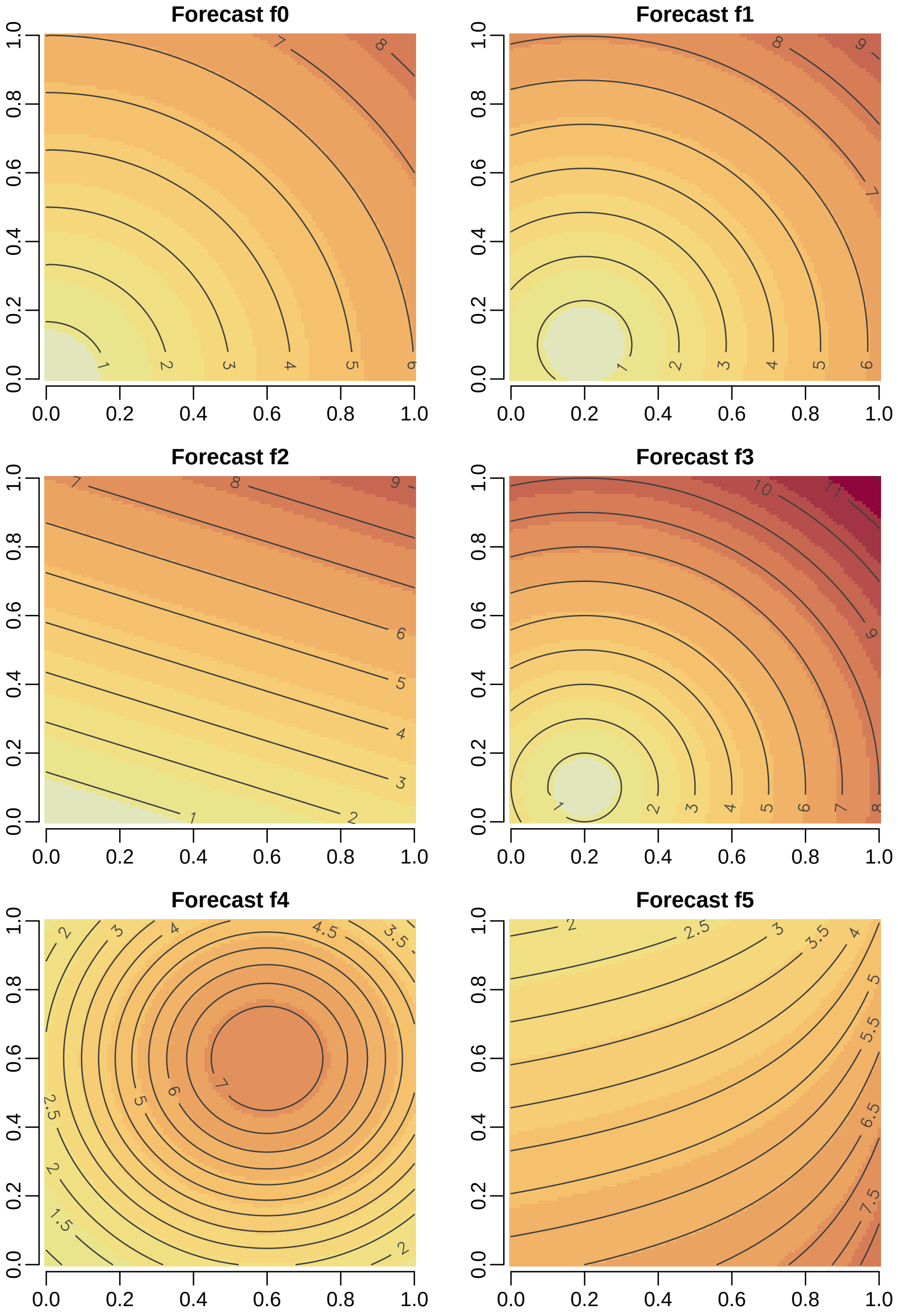

Figure S1: Heat maps of the intensity forecasts $f_0, \dots, f_5$, see Section 4.

with variance $\sigma^2 = \max_{x,y\in[0,1]} f_0(x,y)$ and scale $s = 6/100$. We then apply independent thinning to the homogeneous Gaussian DPP in order to obtain the final point process with intensity function f_0.

3. An inhomogeneous log-Gaussian Cox process (LGCP), see e.g. Illian et al. (2008, Chapter 6). A LGCP is a Poisson point process conditional on a random intensity function arising from a log-Gaussian random field. If $\mu : \mathbb{R}^d \to \mathbb{R}$ is the mean and $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the covariance of the random field, then the LGCP has intensity function

$$x \mapsto \exp\left(\mu(x) + \frac{1}{2}C(x,x)\right).$$

   We choose $C(x_1, x_2) = C_0(\|x_1 - x_2\|)$, where $C_0 : [0,\infty) \to \mathbb{R}$ is the exponential covariance function

$$C_0(r) = \sigma^2 \exp\left(-\frac{r}{s}\right), \tag{8}$$

   with variance $\sigma^2 = 1/4$ and scale $s = 1/5$. The mean is set to $\mu(x) = \log(f_0(x)) - 1/8$ such that the intensity equals f_0.

4. An inhomogeneous Thomas cluster process, see e.g. Illian et al. (2008, Chapter 6). This is a cluster process which arises from an inhomogeneous Poisson point process as parent and a random number of cluster points which are drawn from a normal distribution centred at its parent point. As intensity of the parent process we choose $2f_0/3$ and the number of points per cluster follows a Poisson distribution with parameter $3/2$. The location of each cluster point is determined by a normal distribution which is centred at the parent point and where the components are uncorrelated and have standard deviation 0.05. As a result of the clustering, the intensity of the Thomas process is only approximately equal to f_0.

Further details for the experiments of Section 4 Section 4 presents four simulation experiments based on the scoring function S_1. Table S1 shows the results of DM tests (see Diebold and Mariano (1995) and Section S1) for these experiments. For each of the $M = 500$ realizations we test whether forecast f_i (row) achieves the same expected score as forecast f_j (column). The rejection frequencies in favour of f_0 against f_j, $j = 1, \ldots, 5$ (first row of each table) are generally in line with the mean score differences in Figure 1. Moreover, the results of the DM tests are similar for all four simulation experiments. In the third and fourth experiment (lower part of Table S1) the frequencies of rejection in favour of the optimal forecast f_0 (first row of each table) decrease slightly

Table S1: Fraction of replicates where the “row forecast” was preferred over the “column forecast” by a standard DM test with level $\alpha = 0.05$ based on the scoring function S_1 (5) and $M = 500$ replicates

Poisson

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.45	0.81	0.96	0.99	1.00
f_1	0.00		0.40	0.88	0.82	0.99
f_2	0.00	0.00		0.28	0.69	0.98
f_3	0.00	0.00	0.01		0.24	0.91
f_4	0.00	0.00	0.00	0.01		0.97
f_5	0.00	0.00	0.00	0.00	0.00	

DPP

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.52	0.83	0.97	0.99	1.00
f_1	0.00		0.39	0.91	0.80	1.00
f_2	0.00	0.00		0.27	0.67	0.97
f_3	0.00	0.00	0.01		0.22	0.93
f_4	0.00	0.00	0.00	0.01		0.98
f_5	0.00	0.00	0.00	0.00	0.00	

LGCP

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.48	0.80	0.93	0.99	1.00
f_1	0.00		0.39	0.85	0.81	1.00
f_2	0.00	0.00		0.27	0.66	0.97
f_3	0.00	0.00	0.01		0.19	0.92
f_4	0.00	0.00	0.00	0.01		0.97
f_5	0.00	0.00	0.00	0.00	0.00	

Thomas

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.24	0.52	0.76	0.89	1.00
f_1	0.00		0.26	0.60	0.56	0.91
f_2	0.00	0.00		0.18	0.44	0.81
f_3	0.00	0.00	0.01		0.14	0.68
f_4	0.00	0.00	0.00	0.01		0.78
f_5	0.00	0.00	0.00	0.00	0.00	

for the LGCP and substantially for the Thomas process. An intuitive reason for this is that clustering, which is a feature of both processes, complicates the distinction between different intensity forecasts.

Experiments with a different scoring function We now investigate how the forecast comparison changes when using the scoring function S_2 instead of the scoring function S_1 from Section 4. Boxplots of mean score differences are given in Figure S2 and they are generally similar to the ones presented in Figure 1.

The same conclusion holds for the results of DM tests given in Table S2 resemble those in Table S1. This suggests that in our experiments the choice of $c = 1/10$ for S_1 leads to a similar balance of shape and total mass of the intensity as with S_2. However, in other forecast settings, or with a different choice of c, the two scoring functions may lead to differing conclusions. As in the previous experiments, the clustering of the LGCP and the Thomas process leads to less conclusive decisions between the forecasts. In contrast, the inhibition of the Gaussian DPP seems to facilitate the comparison between the forecasts.

A further sequence of experiments considers the speed of convergence in Proposition 4, i.e. how well score differences based on $S_{\mathrm{cell}}^{\mathcal{T}_n}$, as defined in (14), approximate

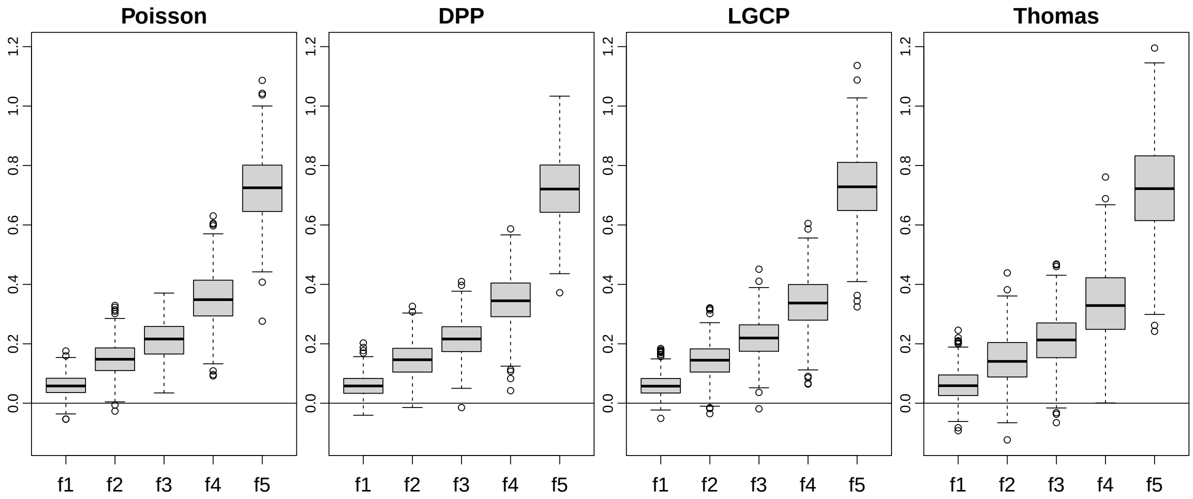

Figure S2: Boxplot of difference in mean scores $\bar{s}_j - \bar{s}_0$ for $j = 1, \dots, 5$ and scoring function S_2 (6). From left to right, Φ is a Poisson point process, a Gaussian determinantal point process, a log-Gaussian Cox process, or an inhomogeneous Thomas process. Means are based on $N = 100$ realizations, boxplots on $M = 500$ replicates.

score differences based on S_2 (6). We select a family of partitions $(\mathcal{T}_n)_{n \in \mathbb{N}}$ of $[0,1]^2$ which arises from dyadic partitions of both axes. Specifically, each grid cell $B_{ij}^{(n)} \in \mathcal{T}_n$ is given by $[(i-1)/2^n, i/2^n] \times [(j-1)/2^n, j/2^n]$ for $i, j \in \{1, \dots, 2^n\}$. The number of cells is thus $k_n = 2^{2n}$ and we choose $n \in \{1, \dots, 6\}$ for the simulations. As forecasts we rely on the intensity functions $f_0, \dots, f_5$ introduced in Section 4 which we transform into grid-based reports $f_{l,ij}^{(n)}$ by integrating f_l over the grid cell $B_{ij}^{(n)}$. These reports are then compared to the number of points per cells via $S_{\text{cell}}^{\mathcal{T}_n}$. We study the convergence of the rejection probabilities of DM tests based on $S_{\text{cell}}^{\mathcal{T}_n}$ for $N = 100$ i.i.d. samples of Φ and increasing n. The corresponding fractions converge to the values in Table S2, as illustrated in Figure S3 for the comparisons of f_0 to $f_1, \dots, f_5$. These simulations suggest that for forecasts which are far from the underlying truth $n = 2$, i.e. 16 grid cells, is already enough to obtain DM results based on $S_{\text{cell}}^{\mathcal{T}_n}$ which are in good agreement with the results based on S_2 (Table S2). For intensity functions closer to the truth, such as f_1, $n = 3$, i.e. 64 grid cells, seems necessary to obtain a good approximation.

S3.2 Product density

This subsection presents simulation experiments for the product density (Section S2.3). We simulate stationary and isotropic point processes with three different second order structures corresponding to inhibition, clustering, and no interaction. We draw $N = 30$

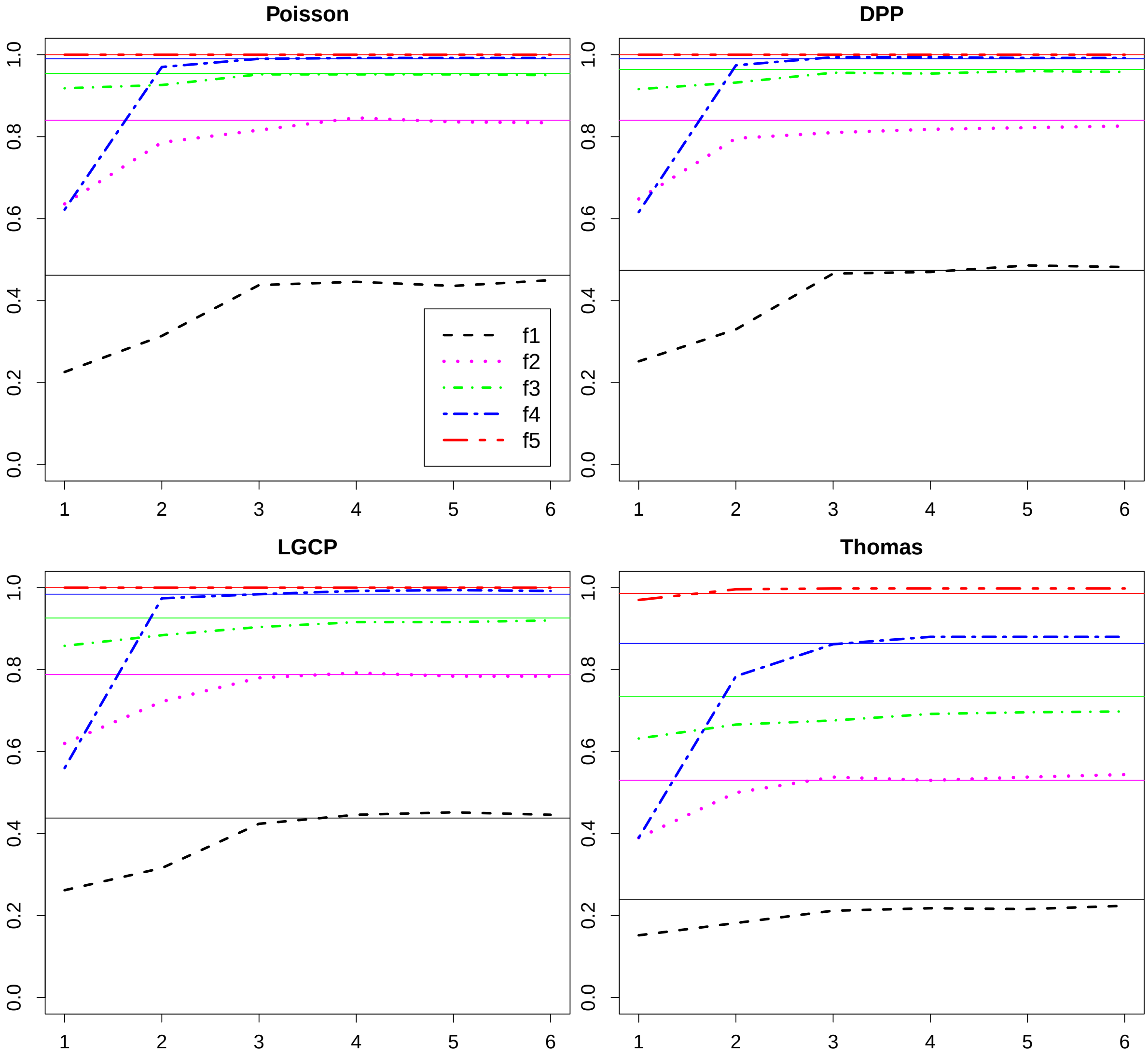

Figure S3: Fraction of replicates where f_0 was preferred over $f_1, \ldots, f_5$ by a standard DM test with level $\alpha = 0.05$ based on the scoring function $S_{\text{cell}}^{\mathcal{T}_n}$, with n varying along the horizontal axis, sample size $N = 100$, and $M = 500$ replicates. The solid lines represent the fractions resulting from the use of S_2 (see (6)), as given in Table S2. The legend in the upper left plot applies to all other plots, too.

Table S2: Fraction of replicates where the "row forecast" was preferred over the "column forecast" by a standard DM test with level $\alpha = 0.05$ based on the scoring function S_2 (6) and $M = 500$ replicates

Poisson

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.46	0.84	0.95	0.99	1.00
f_1	0.00		0.48	0.87	0.84	1.00
f_2	0.00	0.00		0.24	0.70	0.98
f_3	0.00	0.00	0.01		0.29	0.94
f_4	0.00	0.00	0.00	0.00		0.96
f_5	0.00	0.00	0.00	0.00	0.00	

DPP

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.47	0.84	0.96	0.99	1.00
f_1	0.00		0.43	0.93	0.86	1.00
f_2	0.00	0.00		0.22	0.69	0.98
f_3	0.00	0.00	0.01		0.28	0.93
f_4	0.00	0.00	0.00	0.01		0.97
f_5	0.00	0.00	0.00	0.00	0.00	

LGCP

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.44	0.79	0.93	0.98	1.00
f_1	0.00		0.39	0.84	0.83	0.99
f_2	0.00	0.00		0.27	0.68	0.96
f_3	0.00	0.00	0.00		0.23	0.92
f_4	0.00	0.00	0.00	0.01		0.97
f_5	0.00	0.00	0.00	0.00	0.00	

Thomas

	f_0	f_1	f_2	f_3	f_4	f_5
f_0		0.24	0.53	0.73	0.86	0.99
f_1	0.00		0.26	0.58	0.53	0.92
f_2	0.00	0.01		0.16	0.42	0.80
f_3	0.00	0.00	0.01		0.15	0.69
f_4	0.00	0.00	0.00	0.01		0.79
f_5	0.00	0.00	0.00	0.00	0.00	

i.i.d. samples φ_i from Φ and compare the mean scores for different forecasts, in the same way as in Section 4. The scoring function S is defined in Example S4 and the scaling factor $c = 10^{-5}$ is chosen such that the log and squared terms are of the same order of magnitude. We repeat the simulations $M = 500$ times to assess the variation in mean scores.

Details on the point process models We simulate three different stationary and isotropic data-generating processes Φ on the window $[0,1]^2$ with intensity $\lambda = 25$. The models are specified as follows:

1. A LGCP which is determined by a stationary and isotropic Gaussian process with mean $\mu \in \mathbb{R}$ and covariance function C_0, see e.g. Illian et al. (2008). Its second order product density $\varrho^{(2)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is given by $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$, where

$$\varrho_0^{(2)}(r) = \exp\left(2\mu + C_0(0) + C_0(r)\right).$$

   We choose C_0 as the Gaussian covariance function (7) with variance $\sigma^2 = \log 2$ and scale $s = 5/100$ and set $\mu = \log(\lambda) - \sigma^2/2$.

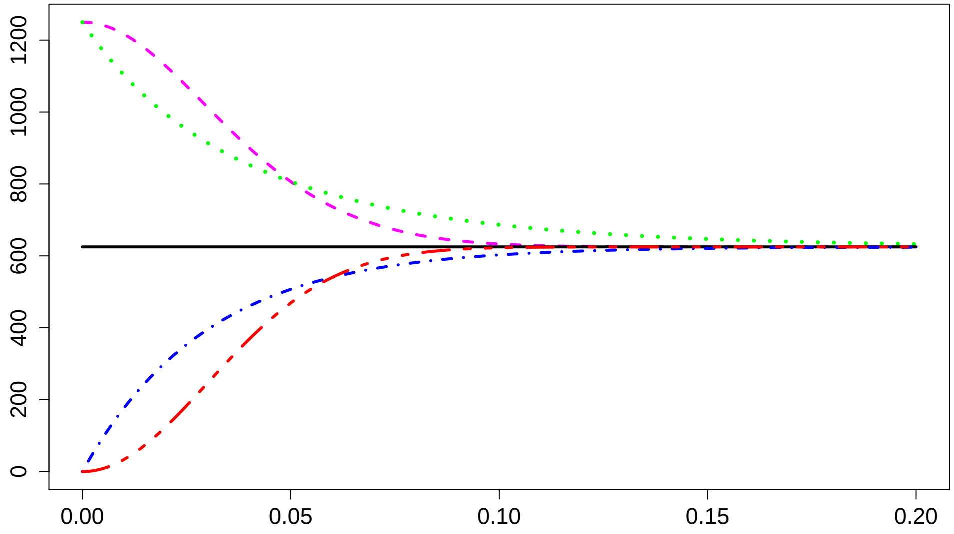

Figure S4: Plot of the five different choices for $\varrho_0^{(2)} : [0,\infty) \to [0,\infty)$ on which the product density forecasts in Section S3.2 are based. The first two (f_1 and f_2) represent clustering, the last two (f_4 and f_5) inhibition. The constant f_3 implies no interaction.

2. A homogeneous Poisson point process.

3. A DPP defined via the Gaussian covariance function (7), see e.g. Hough et al. (2006) and Lavancier et al. (2015). Its second order product density is given by $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$, where

$$\varrho_0^{(2)}(r) = C_0(0)^2 - C_0(r)^2,$$

and C_0 is the Gaussian covariance (7) with variance $\sigma^2 = \lambda^2$ and scale $s = 0.06$.

Forecast comparison The three simulation experiments compare five different product density forecasts, which are based on stationary and isotropic point processes, see Example S4. Hence, the forecasts take the form $\varrho^{(2)}(x_1, x_2) = \varrho_0^{(2)}(\|x_1 - x_2\|)$, with the function $\varrho_0^{(2)}$ given by

$$f_1(r) = \exp\left[2\mu + \sigma^2\left\{1 + \exp(-400r^2)\right\}\right]$$
$$f_2(r) = \exp\left[2\mu + \sigma^2\left\{1 + \exp(-20r)\right\}\right]$$
$$f_3(r) = \lambda^2$$

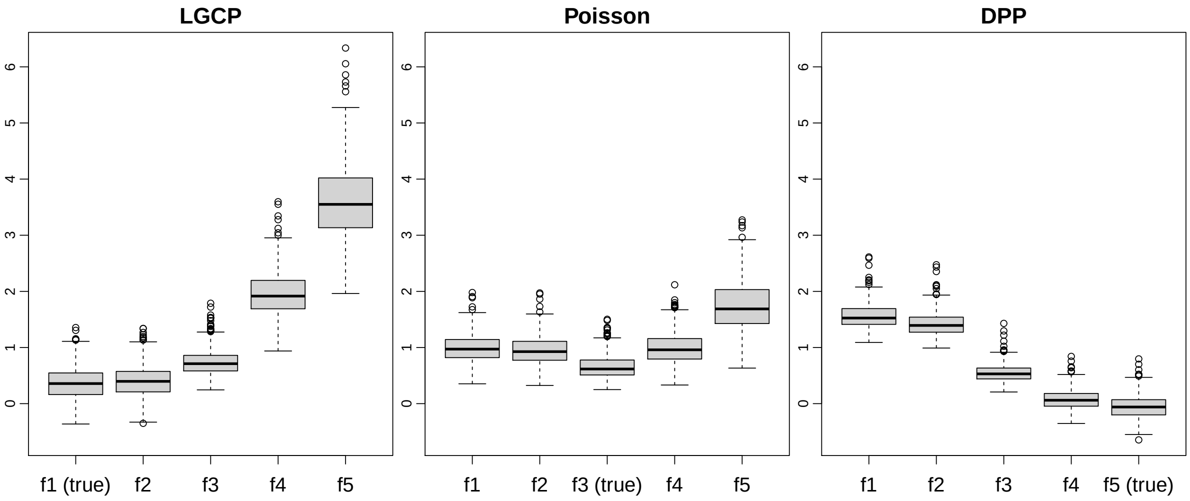

Figure S5: Boxplots of mean scores $\bar{s}_j$ for different product density forecasts, where Φ is a log-Gaussian Cox process (left), a homogeneous Poisson process (centre), or a Gaussian determinantal point process (right). Means are based on $N = 30$ realizations, boxplots on $M = 500$ replicates.

$$f_4(r) = \lambda^2 \{1 - \exp(-2r/s)\}$$
$$f_5(r) = \lambda^2 \left\{1 - \exp(-2(r/s)^2)\right\},$$

where $\mu = \log(\lambda) - \sigma^2/2$, $\sigma^2 = \log(2)$, $s = 0.06$, and $\lambda = 25$. See Figure S4 for a graphical comparison of the different functions. The forecasts f_1 and f_2 represent clustering, since they arise as product densities of LGCPs with Gaussian or exponential covariance function (see (7) and (8)). The constant function f_3 corresponds to a homogeneous Poisson process. The forecasts f_4 and f_5 arise as product densities of DPPs with Gaussian or exponential covariance function and thus represent inhibition. Our parameter choices ensure that the point process models corresponding to $f_1, \dots, f_5$ all have intensity equal to λ, so forecast misspecifications only occur in the product density.

In the first experiment the true Φ is a LGCP with a Gaussian covariance function such that its product density corresponds to f_1. In the second experiment Φ is a homogeneous Poisson process with intensity λ, such that f_3 becomes the optimal forecast in this situation. Lastly, we let Φ be a DPP with Gaussian covariance function and parameters such that f_5 is optimal. We thus perform one experiment for each of the three phenomena clustering, no interaction, and inhibition.

The simulated mean scores are displayed in Figure S5 for all three experiments. The optimal forecast consistently achieves the lowest mean score. In the case of clustering (left subfigure) the LGCP related forecasts f_1 and f_2 perform roughly similar, while the

Table S3: Fraction of times the “row forecast” was preferred over the “column forecast” by a standard DM test with level $\alpha = 0.05$ in the product density experiments (Section S3.2), based on $M = 500$ repetitions

LGCP

	f_1	f_2	f_3	f_4	f_5
f_1		0.18	0.63	0.99	1.00
f_2	0.00		0.57	0.99	1.00
f_3	0.00	0.00		1.00	1.00
f_4	0.00	0.00	0.00		1.00
f_5	0.00	0.00	0.00	0.00	

DPP

	f_1	f_2	f_3	f_4	f_5
f_1		0.00	0.00	0.00	0.00
f_2	0.90		0.00	0.00	0.00
f_3	1.00	1.00		0.00	0.00
f_4	1.00	1.00	1.00		0.00
f_5	1.00	1.00	0.96	0.57	

Poisson

	f_1	f_2	f_3	f_4	f_5
f_1		0.01	0.00	0.04	0.43
f_2	0.17		0.00	0.06	0.50
f_3	0.77	0.68		0.64	0.96
f_4	0.07	0.05	0.00		1.00
f_5	0.00	0.00	0.00	0.00	

misspecified no interaction and inhibition forecasts f_3, f_4 and f_5 lead to considerably higher mean scores. A similar, but mirrored behaviour is apparent in the inhibition experiment (right subfigure): The forecast f_4, which gets the nature of point interactions right, attains low mean scores, even though it is not optimal. The mean scores of the Poisson forecast f_3 are always in between the “extremes”. The DM test probabilities of the three experiments are given in Table S3 and support these observations. Additionally, the DM results illustrate that the clustering forecasts f_1 and f_2 are preferred more often over the inhibition forecast f_5 in the case of Poisson data (centre table).

S4 Additional details for the case study

This material extends Section 5.2. Figure S6 reproduces Figure 3 but with the quadratic score S_{quad} rather than the Poisson score S_{pois}. In contrast to Figure 3 we see that there are periods without events where the LG model rather than the FMC model attains the lowest scores.

Figures S7 and S8 use the same methods as in Figure 4 to compare the LM model to the LG and the SMA model. The regions of superior or inferior forecast performance of the LM model remain generally the same across the three comparisons. The right plots of these figures compare the forecasts after spatial aggregation, for which we give details now.

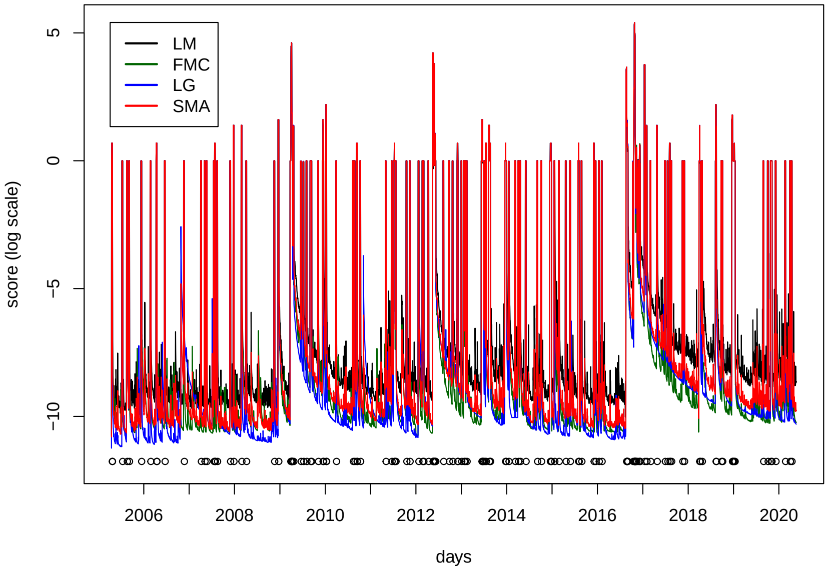

Figure S6: Daily scores $s_{j,t}$ from (10) based on S_{quad} for the four forecasting models from 2005 to 2020, logarithmic scale. The circles indicate the days of M4+ earthquakes and the tickmarks on the horizontal axis mark the first day of each year.

S4.1 Spatial aggregation

We follow the notation of Section 5.2, except that we introduce a coordinate notation for the testing region (Figure 2). For each grid cell B_i we now write $B_{k,l}$ where k is the horizontal and l the vertical coordinate. A cell with a higher value of k is further east and a cell with a higher value of l is further north. Similarly, let $x^{(j)}_{k,l,t}$ be the forecast of model j corresponding to cell $B_{k,l}$ on day t. For combinations of k and l that fall outside the testing region we use the convention $x^{(j)}_{k,l,t} = 0$ and $B_{k,l} = \emptyset$.

Let $\delta \in \mathbb{N}_0$ be a given level of aggregation. We define the locally aggregated forecast and the locally aggregated grid cell at coordinate (k, l) and aggregation level δ via

$$\bar{x}^{(j)}_{k,l,t} := \sum_{\mu=-\delta}^{\delta} \sum_{\nu=-\delta}^{\delta} x^{(j)}_{k+\mu,l+\nu,t} \qquad \text{and} \qquad \bar{B}_{k,l} := \bigcup_{\mu=-\delta}^{\delta} \bigcup_{\nu=-\delta}^{\delta} B_{k+\mu,l+\nu}$$

respectively. In the interior of the testing region, this is an aggregation of the forecasts over a square neighbourhood with edge length $2\delta + 1$ centred at (k, l). At the boundary of the testing region the aggregation neighbourhoods will be smaller, however, as there are almost no events in this area, this does not affect the plots. Due to the linearity of expectations, the values $\bar{x}^{(j)}_{k,l,t}$ are again valid mean forecasts that can be compared via consistent scoring functions, e.g. the Poisson score (9). The right plots of Figures 4,

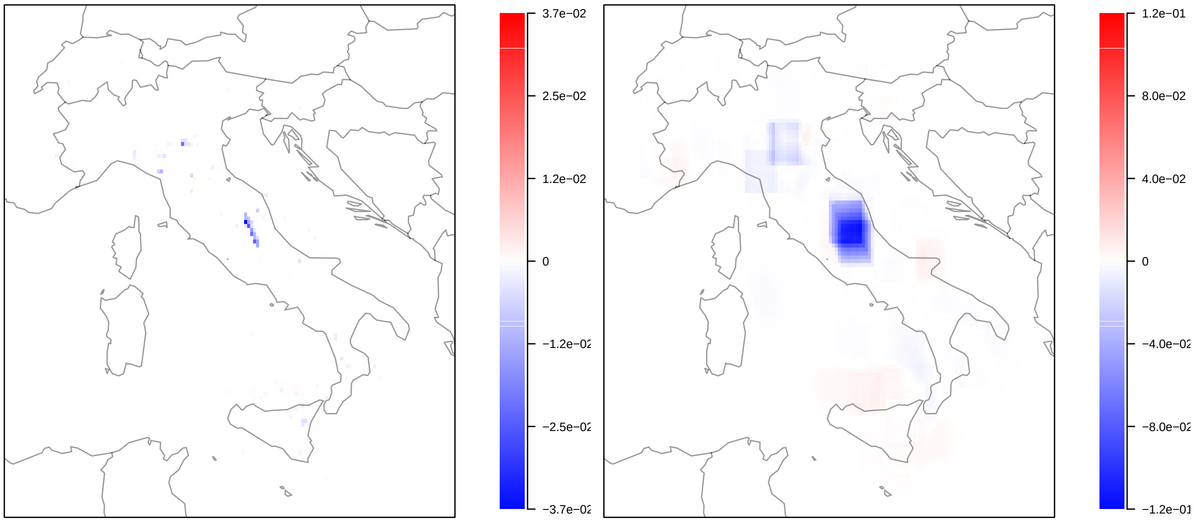

Figure S7: Mean score difference based on S_{pois} (11) between the LM and the LG model, without (left) and with (right) aggregation. Negative values (blue) indicate that the LM model has superior forecast performance, and positive values (red) vice versa.

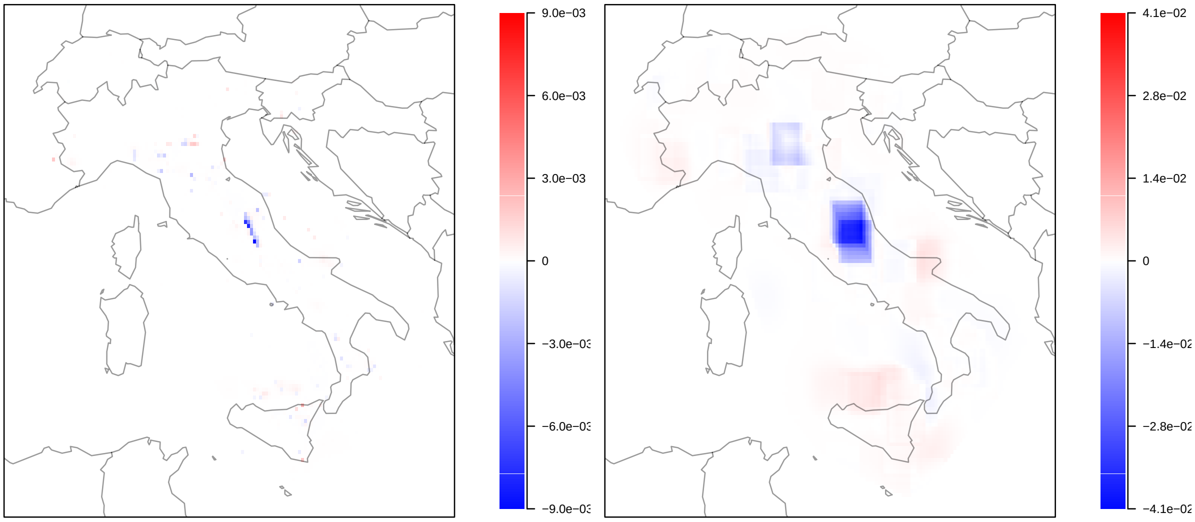

Figure S8: Mean score difference based on S_{pois} (11) between the LM and the SMA model, without (left) and with (right) aggregation. Negative values (blue) indicate that the LM model has superior forecast performance, and positive values (red) vice versa.

S7, and S8 show this comparison via the mean score difference of the locally aggregated forecasts

$$\bar{\Delta}_{k,l}^{(j,j')} := \frac{1}{5514} \sum_{t=1}^{5514} \left(S_{\mathrm{pois}}(\bar{x}_{k,l,t}^{(j)}, \varphi_t(\bar{B}_{k,l})) - S_{\mathrm{pois}}(\bar{x}_{k,l,t}^{(j')}, \varphi_t(\bar{B}_{k,l})) \right),$$

where $\delta = 5$. For $\delta = 0$ there is no aggregation, so $\bar{\Delta}_{k,l}^{(j,j')}$ simplifies to $\Delta_i^{(j,j')}$, the (non-aggregated) mean score difference (11). For δ large enough there is essentially only one big grid cell and one forecasted number remaining. The corresponding plot would show only one colour, indicating the forecast performance of the models with respect to the total number of events in the testing region.

S4.2 Sample size considerations

Point process forecasting is often challenged by a lack of data, and particularly a lack of data to properly test newly proposed prediction models. In this light, a critical question is how much data is required to reach valid conclusions on superior predictive ability. As discussed, a commonly used tool is the Diebold–Mariano (DM) test, 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, and a well known, crude rule of thumb (Lehr, 1992; van Belle, 2008) states that for a one-sample, two-tailed t-test with level 0.05, a sample size $n = 8s^2/d^2$ yields an approximate power of 0.80, where s^2 is the variance of the score differentials, and d is the difference to be detected. Phrased differently, if the variance s^2 and the sample size n are given, a difference $d_n = (8s^2/n)^{1/2}$ is detectable, subject to the above specifications of the size and the power of the t-test.

In Tables S4 and S5 we return to Table 1 in the main paper, where we compare the predictive performance of the LM, FMC, LG, and SMA models, respectively. We show the mean score differential and its variance, and find the detectable difference d_{5514} at the given sample size of $n = 5514$ daily forecasts of earthquake activity over the subsequent seven-day period, for the Poisson score and the quadratic score, respectively. Figures S9 and S10 show the sample autocorrelation function for the score differentials. Not surprisingly, there is considerable dependency at lags up to about seven to nine days ahead, due to the overlap in the seven-day outlook, though autocorrelations are small to negligible at higher lags. As standard power calculations assume independent samples, a more appropriate quantification of a detectable difference is based on a sample size of $[5514/7] = 787$. A further alternative is to use an estimate of the effective sample size (Thiébaux and Zwiers, 1984), which reduces the regular sample size according to the autocorrelation of the series, in line with the handling of dependencies in DM tests.

Table S4: Mean m and variance s^2 of the score differential, and detectable difference d_n for sample size $n = 787$ and $n = 5514$ according to the rule of thumb by Lehr (1992), under the Poisson score and for the models from Table 1 in the main paper.

Poisson score	LG−LM	LG−SMA	LG−FMC	FMC−LM	FMC−SMA	SMA−LM
Mean m	0.307	0.285	0.221	0.086	0.064	0.022
Variance s^2	11.936	6.438	4.885	2.542	0.695	0.983
d_{5514}	0.132	0.097	0.084	0.061	0.032	0.038
d_{787}	0.348	0.256	0.223	0.161	0.084	0.100

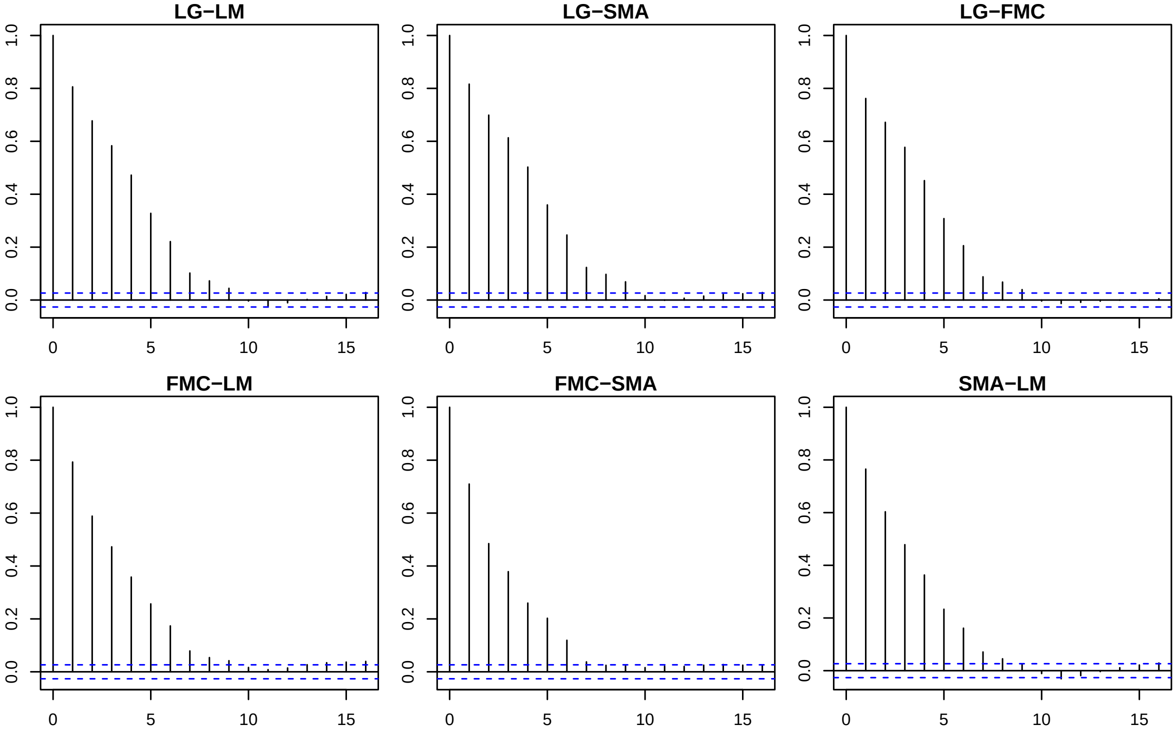

Figure S9: Sample autocorrelation function of the Poisson score differentials for the forecasts from Table 1 in the main paper, with lag in days

Table S5: Same as Table S4, but under the quadratic score. All entries are to be divided by a factor of 100.

Quadratic score	LG−LM	FMC−LM	SMA−LM	LG−SMA	FMC−SMA	LG−FMC
Mean m	0.563	0.505	0.293	0.270	0.211	0.058
Variance s^2	1.690	1.303	0.605	0.295	0.159	0.152
d_{5514}	0.495	0.435	0.296	0.207	0.152	0.149
d_{787}	1.311	1.151	0.784	0.548	0.402	0.393

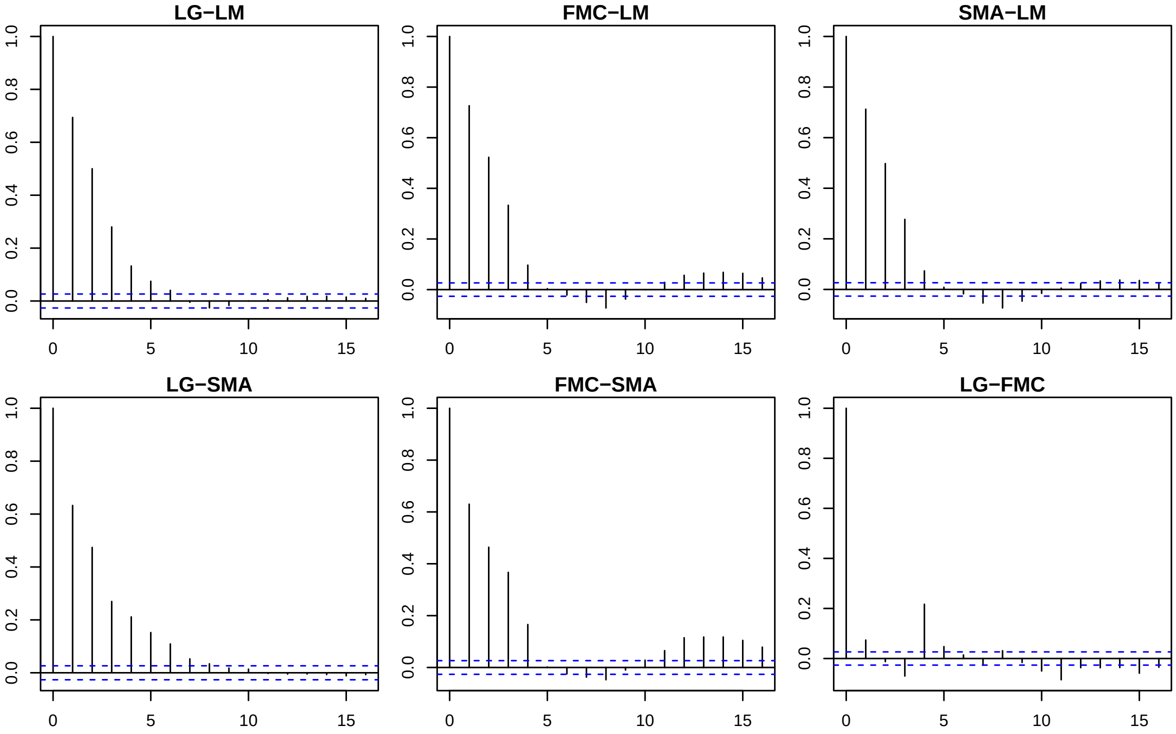

Figure S10: Same as Figure S9, but under the quadratic score

Interestingly, under both the Poisson and the quadratic score, and for each of the six binary model comparisons, the actual mean score differential m tends to be nested in between the (overly) optimistic estimate d_{5514} and the (arguably) realistic estimate d_{787} for a detectable difference, which indicates that the comparative evaluation might reasonably be considered to be based on sufficient data. Evidently, this current analysis is crude and preliminary, using default specifications from the biostatistical literature for size and power, and we encourage follow-up studies.

References

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. *Proceedings of the Cambridge Philosophical Society*, 48, 600–607. URL https://doi.org/10.1017/s0305004100076386.

Barthelmé, S., Trukenbrod, H., Engbert, R. and Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, 13, 1–34. URL https://doi.org/10.1167/13.12.1.

Chiu, S. N., Stoyan, D., Kendall, W. S. and Mecke, J. (2013). *Stochastic Geometry and Its Applications*. 3rd edition. John Wiley & Sons, Chichester. URL https://doi.org/10.1002/9781118658222.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I.* 2nd edition. Springer-Verlag, New York.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263. URL https://doi.org/10.1198/073500102753410444.

Ehm, W. and Gneiting, T. (2012). Local proper scoring rules of order two. *Annals of Statistics*, 40, 609–637. URL https://doi.org/10.1214/12-AOS973.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, vol. 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. URL https://doi.org/10.1007/978-3-642-33483-2.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D. and Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, 111, 564–584. URL https://doi.org/10.1080/01621459.2015.1135802.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74, 1545–1578. URL http://dx.doi.org/10.1111/j.1468-0262.2006.00718.x.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746–762. URL https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378. URL https://doi.org/10.1198/016214506000001437.

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

Hough, J. B., Krishnapur, M., Peres, Y. and Virág, B. (2006). Determinantal processes and independence. *Probability Surveys*, 3, 206–229. URL https://doi.org/10.1214/154957806000000078.

Hwang, E. and Shin, D. W. (2012). Random central limit theorems for linear processes with weakly dependent innovations. *Journal of the Korean Statistical Society*, 41, 313–322. URL https://doi.org/10.1016/j.jkss.2011.10.004.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 695–709. URL https://jmlr.org/papers/v6/hyvarinen05a.html.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Ltd., Chichester.

Lavancier, F., Møller, J. and Rubak, E. (2015). Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77, 853–877. URL https://doi.org/10.1111/rssb.12096.

Lee, S. (1997). Random central limit theorem for the linear process generated by a strong mixing process. *Statistics & Probability Letters*, 35, 189–196. URL https://doi.org/10.1016/S0167-7152(97)00013-8.

Lehr, R. (1992). Sixteen s-squared over d-squared: A relation for crude sample size estimates. *Statistics in Medicine*, 11, 1099–1102. URL https://doi.org/10.1002/sim.4780110811.

Mikosch, T. (2009). *Non-Life Insurance Mathematics. An Introduction with the Poisson Process*. 2nd edition. Universitext, Springer-Verlag, Berlin. URL https://doi.org/10.1007/978-3-540-88233-6.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106, 100–108. URL https://doi.org/10.1198/jasa.2011.ap09546.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11, 1833–1874. URL https://doi.org/10.1214/17-AOAS1041.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50, 379–402. URL https://doi.org/10.1023/A:1003403601725.

Ogata, Y. and Tanemura, M. (1984). Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society Series B: Methodological*, 46, 496–518. URL https://doi.org/10.1111/j.2517-6161.1984.tb01322.x.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *Annals of Statistics*, 40, 561–592. URL https://doi.org/10.1214/12-AOS971.

Schoenberg, F. P., Hoffmann, M. and Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, 71, 1271–1287. URL https://doi.org/10.1007/s10463-018-0690-9.

Shang, Y. (2012). A central limit theorem for randomly indexed m-dependent random variables. *Filomat*, 26, 713–717. URL https://doi.org/10.2298/FIL1204713S.

Stoyan, D. and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, 15, 61–78. URL https://doi.org/10.1214/ss/1009212674.

Thiébaux, H. J. and Zwiers, F. W. (1984). The interpretation and estimation of effective sample size. *Journal of Applied Meteorology and Climatology*, 23, 800–811. URL https://doi.org/10.1175/1520-0450(1984)023<0800:TIAEOE>2.0.CO;2

van Belle, G. (2008). *Statistical Rules of Thumb*. 2nd edition. John Wiley & Sons, Chichester. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470377963.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97, 369–380. URL https://doi.org/10.1198/016214502760046925.