Geophysical Journal International

Geophys. J. Int. (2022) **229,** 1736–1753
Advance Access publication 2022 January 18
GJI Seismology

https://doi.org/10.1093/gji/ggac018

Prospective evaluation of multiplicative hybrid earthquake forecasting models in California

J.A. Bayona [1], W.H. Savran,[2] D.A. Rhoades [3] and M.J. Werner [1]

[1] *School of Earth Sciences, University of Bristol, Queens Road, BS*81*QU, Bristol, UK. E-mail: jose.bayona@bristol.ac.uk*
[2] *Southern California Earthquake Center, University of Southern California, Los Angeles, CE* 90089-0742, *USA*
[3] *GNS Science,* 1 *Fairway Drive, Avalon* 5010 *PO Box* 30-368, *Lower Hutt* 5040, *New Zealand*

Accepted 2022 January 14. Received 2021 November 21; in original form 2021 July 2

SUMMARY

The Regional Earthquake Likelihood Models (RELM) experiment, conducted within the Collaboratory for the Study of Earthquake Predictability (CSEP), showed that the smoothed seismicity (HKJ) model by Helmstetter et al. was the most informative time-independent earthquake model in California during the 2006–2010 evaluation period. The diversity of competing forecast hypotheses and geophysical data sets used in RELM was suitable for combining multiple models that could provide more informative earthquake forecasts than HKJ. Thus, Rhoades et al. created multiplicative hybrid models that involve the HKJ model as a baseline and one or more conjugate models. In retrospective evaluations, some hybrid models showed significant information gains over the HKJ forecast. Here, we prospectively assess the predictive skills of 16 hybrids and 6 original RELM forecasts at a 0.05 significance level, using a suite of traditional and new CSEP tests that rely on a Poisson and a binary likelihood function. In addition, we include consistency test results at a Bonferroni-adjusted significance level of 0.025 to address the problem of multiple tests. Furthermore, we compare the performance of each forecast to that of HKJ. The evaluation data set contains 40 target events recorded within the CSEP California testing region from 2011 January 1 to 2020 December 31, including the 2016 Hawthorne earthquake swarm in southwestern Nevada and the 2019 Ridgecrest sequence. Consistency test results show that most forecasting models overestimate the number of earthquakes and struggle to explain the spatial distribution of epicenters, especially in the case of seismicity clusters. The binary likelihood function significantly reduces the sensitivity of spatial log-likelihood scores to clustering, however; most models still fail to adequately describe spatial earthquake patterns. Contrary to retrospective analyses, our prospective test results show that none of the models are significantly more informative than the HKJ benchmark forecast, which we interpret to be due to temporal instabilities in the fit that forms hybrids. These results suggest that smoothing high-resolution, small earthquake data remains a robust method for forecasting moderate-to-large earthquakes over a period of 5–15 yr in California.

Key words: Probabilistic forecasting; Earthquake interaction, forecasting, and prediction; Seismicity and tectonics; Statistical seismology.

1 INTRODUCTION

More than a third of the world's population lives in areas where earthquakes that cause at least slight damage are frequently expected (Marti et al. 2019). Therefore, the development and prospective evaluation of seismicity models is essential to improve estimates of seismic hazard in these locations, as earthquake-rate models are a key element of Probabilistic Seismic Hazard Assessment (PSHA). On this premise, the Working Group on Regional Earthquake Likelihood Models (RELM, Field 2007; Schorlemmer & Gerstenberger 2007; Schorlemmer et al. 2007) designed a community forecasting experiment, with associated models, data and tests to evaluate earthquake predictability in California. Participating RELM forecast models were based on a wide variety of geophysical data sets, including interseismic strain rates (e.g. SHEN, Shen et al. 2007), earthquake-catalogue data (e.g. KAGAN, Kagan et al. 2007) and geological fault slip rates (e.g. WARD-GEOL;

Ward, 2007). As first-order results, the RELM experiment, conducted within the Collaboratory for the Study of Earthquake Predictability (CSEP, Zechar et al. 2010; Michael & Werner 2018; Schorlemmer et al. 2018), showed that the HKJ smoothed-seismicity model developed by Helmstetter et al. (2007) was the most informative forecast during the 2006–2010 evaluation period (Schorlemmer et al. 2010; Zechar et al. 2013; Strader et al. 2017).

The diversity of competing forecasting models in Earth, atmosphere and climate sciences are well suited for ensemble modelling, which has the potential to gain greater predictive skills than their individual model components (e.g. Vere-Jones 1995; Gneiting and Raftery 2005; Gerstenberger et al. 2020). For Operational Earthquake Forecasting (OEF; Jordan et al. 2011), ensemble models additionally provide an objective method to support decision-making without the need to choose a single model *a priori* (Marzocchi et al. 2012, 2014). Although the procedure for optimally combining earthquake forecasts remains unclear, two main ensembling methods have been pursued recently: additive and multiplicative models. Additive models have occasionally shown forecasting abilities greater than their component models when the latter were uncorrelated (Rhoades & Gerstenberger 2009; Rhoades and Stirling 2012; Rhoades 2013; Taroni et al. 2014). In the case of RELM, Marzocchi et al. (2012) found no appreciable gain of probability-weighted forecasts over the best-performing individual model. In contrast, some multiplicative ensembles have been found to be statistically more informative than their constituent forecasting models in retrospective and pseudo-prospective evaluations (e.g. Shebalin et al. 2014, Bird et al. 2015; Rhoades et al. 2016). On a global scale, the outperformance of multiplicative blends over their individual parent components has also been reported after 1–6 yr of prospective testing (Bird, 2018; Strader et al. 2018; Bayona et al. 2021). In the context of RELM, Rhoades et al. (2014) created multiplicative hybrids that involve the HKJ forecast as baseline and one or more conjugate models. Specifically, the authors fitted two parameters for each conjugate model and an overall normalizing constant to optimize the performance of each hybrid model retrospectively. The authors computed corrected information gains per earthquake (IGPEc) using the corrected Akaike Information Criterion (AICc; Hurvich and Tsai, 1989) statistic, which penalizes the information score for the number of fitted parameters. Some hybrid models showed significant information gains over the individual HKJ forecast component, despite the penalty. In particular, seismicity models that make use of geodetic data were identified as effective conjugate approaches with HKJ, which relies on earthquake information.

The robustness of the fit to form multiplicative hybrid earthquake models for California can only be objectively described after prospective evaluation. Although useful as a sanity-check, retrospective test results might not be indicative of model performance on independent data (Werner et al. 2010). Hence, in this study we prospectively assess 16 hybrids as well as 6 original RELM models. Particularly, we use a set of traditional forecast tests implemented in CSEP to evaluate the consistency between the observed and the expected number, spatial, and likelihood distributions of earthquakes, and to compare the performance of each forecast to that of HKJ. Traditional CSEP tests are based on a likelihood function that approximates earthquakes in individual cells or bins as independent and Poisson distributed (Schorlemmer & Gerstenberger 2007; Schorlemmer et al. 2010; Zechar et al. 2010). However, the Poisson distribution insufficiently captures the spatiotemporal variability of earthquakes, especially in the presence of clusters of seismicity (Werner and Sornette, 2008; Lombardi and Marzocchi, 2010; Nandan et al. 2019). Therefore, we additionally introduce new tests based on a binary probability function that reduces the sensitivity of CSEP evaluations to clustering.

2 MODELS

2.1 RELM models

The RELM working group designed a 5-yr class of forecast experiments to test the consistency of models/hypotheses describing the seismogenesis with the observations (Schorlemmmer & Gerstenberger 2007). Within this class, two earthquake forecasting subclasses were created to differentiate the applicability of each experiment: mainshock and mainshock+aftershock models. In total, eleven forecast models were submitted to the mainshock class, and six modified versions were included in the mainshock+aftershock class (i.e. models forecast all earthquakes, and data are not declustered, see Fig. 1). Here, we only discuss the RELM models that were used for constructing hybrids and those that compete in the mainshock+aftershock class for comparability. The full set of models and results were described by Field (2007), Schorlemmer et al. (2010) and Zechar et al. (2013). Four RELM models were obtained by smoothing past seismicity based on different hypotheses: the EBEL-C (Ebel et al. 2007) model averages the 5-yr rate of $M \geq 5$ earthquakes using non-declustered seismicity recorded within the CSEP-California testing region (i.e. Schorlemmer & Gerstenberger 2007) from 1932 to 2004. Similarly, the WARD-SEIS (Ward et al. 2007) model smooths the locations of past earthquakes reported in a historical catalogue from 1850, while the KAGAN (Kagan et al. 2007) model extends the smoothed-seismicity approach by also including smoothed ruptures of large earthquakes observed in the region since 1800. In addition, the HKJ model by Helmstetter et al. (2007) uses an adaptative smoothing of high-resolution $M \geq 2$ earthquake data recorded since 1981 in the whole of California.

Compared to these models, the ALM approach by Wiemer and Schorlemmer (2007) computes spatial variations in a and b values of the Gutenberg–Richter distribution in each forecast bin to estimate the frequency of earthquakes. Alternatively, the SHEN (Shen et al. 2007), WARD-GEOD81 and WARD-GEOD85 (Ward, 2007) models convert maximum horizontal interseismic strain rates into rates of earthquake activity by distributing estimates of geodetic moment rate with Gutenberg–Richter relationships, parametrized by corner magnitudes 8.02, 8.1 and 8.5, respectively. Holliday et al. (2007) developed their earthquake forecast based on pattern informatics (PI) by assuming that regions of strongly fluctuating activity will be the regions of future large earthquakes in the near future, and Bird and Liu (2007) created the BIRD

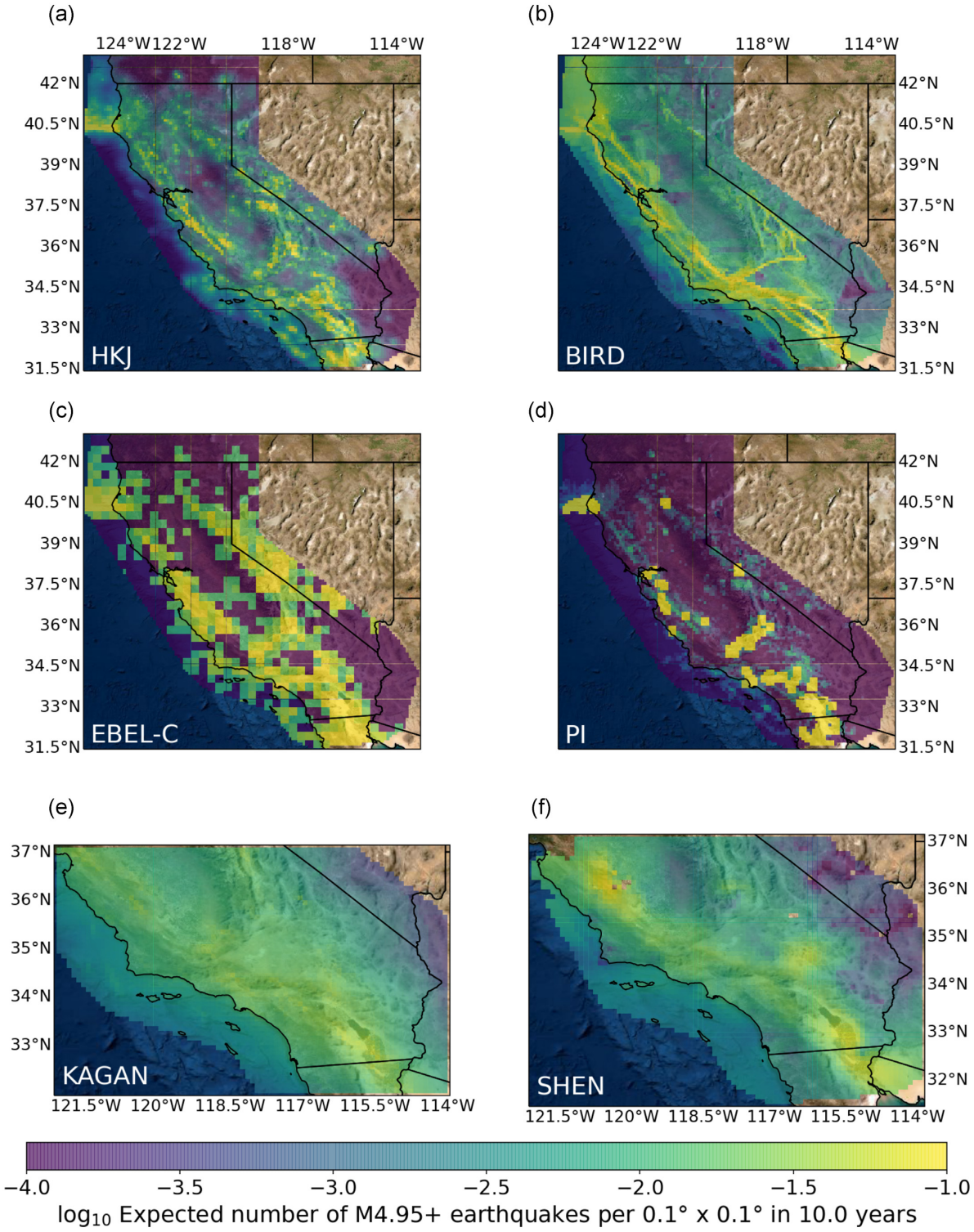

Figure 1. Original RELM models submitted to the mainshock+aftershock class of CSEP: (a) HKJ, (b) BIRD, (c) EBEL-C, (d) PI, (e) KAGAN and (f) SHEN. Expected number of $M \geq 4.95$ earthquakes are shown per $0.1^\circ \times 0.1^\circ$ unit area per decade.

forecast based on a kinematic model of surface velocities in California derived from tectonic, geodetic and geological data. The velocities were mapped into long-term rates of seismic moment and then into earthquake rates, using a global calibration of plate-boundary seismicity. Finally, Ward (2007) constructed the WARD-GEOL, WARD-SIM and WARD-COMBO forecasts by mapping fault slip rates into seismicity rates, simulating velocity-weaking friction on a fixed fault network for California, and averaging the WARD-SEIS, WARD-GEOD81/85 and WARD-GEOL models (see Table 1).

Table 1. Summary of the RELM models that were submitted to the mainshock+aftershock forecasting class of CSEP and those that were used to form hybrids.

Model name	Main features	Reference
EBEL-C	Smoothed-seismicity model based on the observed rate of $M \geq 5$ earthquakes from 1932 and 2004.	Ebel et al. (2007)
WARD-SEIS	Seismicity model based on the smoothing of past $M \geq 5.5$ earthquakes between 1850 and 2003.	Ward (2007)
KAGAN	Smoothed-seismicity model of the locations and finite rupture lenghts of $M \geq 5.0$ events observed between 1800 and 2006.	Kagan et al. (2007)
HKJ	Smoothed-seismicity model that uses $M \geq 2.0$ events observed since 1981, and an optimized adaptative smoothing kernel.	Helmstetter et al. (2007)
ALM	Assumes a Gutenberg–Richter distribution in each grid cell, based on declustered seismicity. Thus, variations in *a*- and *b*-values are used to predict the rates of larger events.	Wiemer and Schorlemmer (2007)
SHEN	Geodesy-based model that converts maximum shear strain rates into long-term rates of seismic moment, which are in turn converted to earthquake rates using a tapered Gutenberg-Richter relationship, with a corner magnitude of 8.02.	Shen et al. (2007)
WARD-GEOD81	Translates geodesy-based estimates of seismic moment into earthquake rates using a Gutenberg–Richter distribution, with upper magnitude 8.1.	Ward (2007)
WARD-GEOD85	Converts geodesy-based estimates of seismic moment into earthquake rates using a Gutenberg–Richter distribution, with upper magnitude 8.5.	Ward (2007)
PI	Assumes that fluctuations in seismicity rates may be related to the preparation phase of large events, and identifies those areas that have undergone strong fluctuating earthquake activity.	Holliday et al. (2007)
BIRD	Incorporates plate tectonic, geodetic and geological data into a kinematic model for California. The deformation, or moment rate, is then converted to rates of earthquake activity using a global calibration of similar plate boundary, crustal seismicity.	Bird and Liu (2007)
WARD-GEOL	Computes moment rates from specified fault slip rates, downdip extends and rigidity values, which are then converted to rates using a Gutenberg–Richter distribution, with an upper magnitude of 8.1.	Ward (2007)
WARD-SIM	Uses earthquake simulations that involve the balance between fault driving stress and fault frictional resistance. When the relation is unbalanced in favour of the fault stress, an earthquake occurs to re-establish the balance.	Ward (2007)
WARD-COMBO	Earthquake rates result from averaging the estimates of seismicity computed by the WARD-SEIS, WARD-GEOD81, WARD-GEOD85 and WARD-GEOL models.	Ward (2007)

2.2 Hybrid models

The five-year evaluation period of the RELM experiment provided 20 $M \geq 4.95$ target earthquakes in the mainshock class and 31 target events in the mainshock+aftershock class in the entire California testing region (Zechar et al. 2013). In southern California, the test period provided 11 earthquakes in the mainshock class and 22 in the mainshock+aftershock class. Based on this, Rhoades et al. (2014) used all the target earthquakes in the mainshock+aftershock class to have a satisfactory number of target events for an optimized multiplicative combination of forecast models. The authors proposed that the hybrid model earthquake rate λ_H in a spatial cell j and magnitude bin k can be estimated as:

$$\lambda_H(j,k) = \lambda_1(j,k)\,\exp\left(a + \sum_{i=2}^{q} f_i[\lambda_i(j,\cdot)]\right), \tag{1}$$

where $\lambda_1(j,k)$ is the expected number of earthquakes in each bin according to the baseline model during the 2006–2011 RELM testing period, $f_i[\lambda_i(j,\cdot)] = b_i(\log(1+\lambda_i))^{c_i}$ is an order-preserving function applied to the ith conjugate model, where $i = 2, 3,.., q$, and a, $b_i \geq 0$ and $c_i > 0$ are normalizing and shape parameters, empirically obtained by the maximum likelihood method. The modelers estimated corrected information gains per earthquake (IGPEc) as:

$$\mathrm{IGPEc} = \frac{\hat{N}_1 - \hat{N}_H}{N} - \frac{1}{2N}\left[2p + \frac{p+1}{N-p-1}\right] + \frac{1}{N}\sum_{n=1}^{N}[X_H(n) - X_1(n)], \tag{2}$$

where N is the number of target events, $\hat{N}_1$ and $\hat{N}_H$ denote the total number of earthquakes expected by the baseline and hybrid models, p is the number of fitted parameters (equal to three) and $X_H(n) = \ln \lambda_H(j_n, k_n)$ and $X_1(n) = \ln \lambda_1(j_n, k_n)$ represent the log-likelihood scores obtained by the hybrid and HKJ model in the bin with the nth target earthquake, respectively. Eq. (2) is equivalent to eq. (17) of Rhoades et al. (2011) that describes the information gain score per earthquake (IGPE), except for the second term that penalizes the IGPE for the number of fitted parameters and the number of target events (Hurvich and Tsai 1989). According to retrospective test results for the 2006–2010 period, earthquake-rate models for the whole of California, such as BIRD and PI gave gains close to 0.25 over HKJ as conjugates. In addition, the SHEN conjugate model in southern California gave a gain of more than 0.5, and several others obtained gains of about 0.2. In general, most hybrid earthquake forecasting models were found to gain predictive skill over HKJ, however; only HKJ-SHEN, HKJ-PI, HKJ-KAGAN, HKJ-WARD81 and HKJ-WARD85 were shown to be significantly more informative than the reference model. Thus, Rhoades et al. (2014) submitted 16 hybrid forecasts to the mainshock+aftershock forecast group of CSEP for prospective evaluation. These models comprised 5 hybrids for the entire California testing region and 11 hybrids for southern California (when one or more component models covered only a subregion; see Fig. 2). In this study, we rely on the confidence bounds calculated by Rhoades et al. (2014) for the IGPEc and use the standard

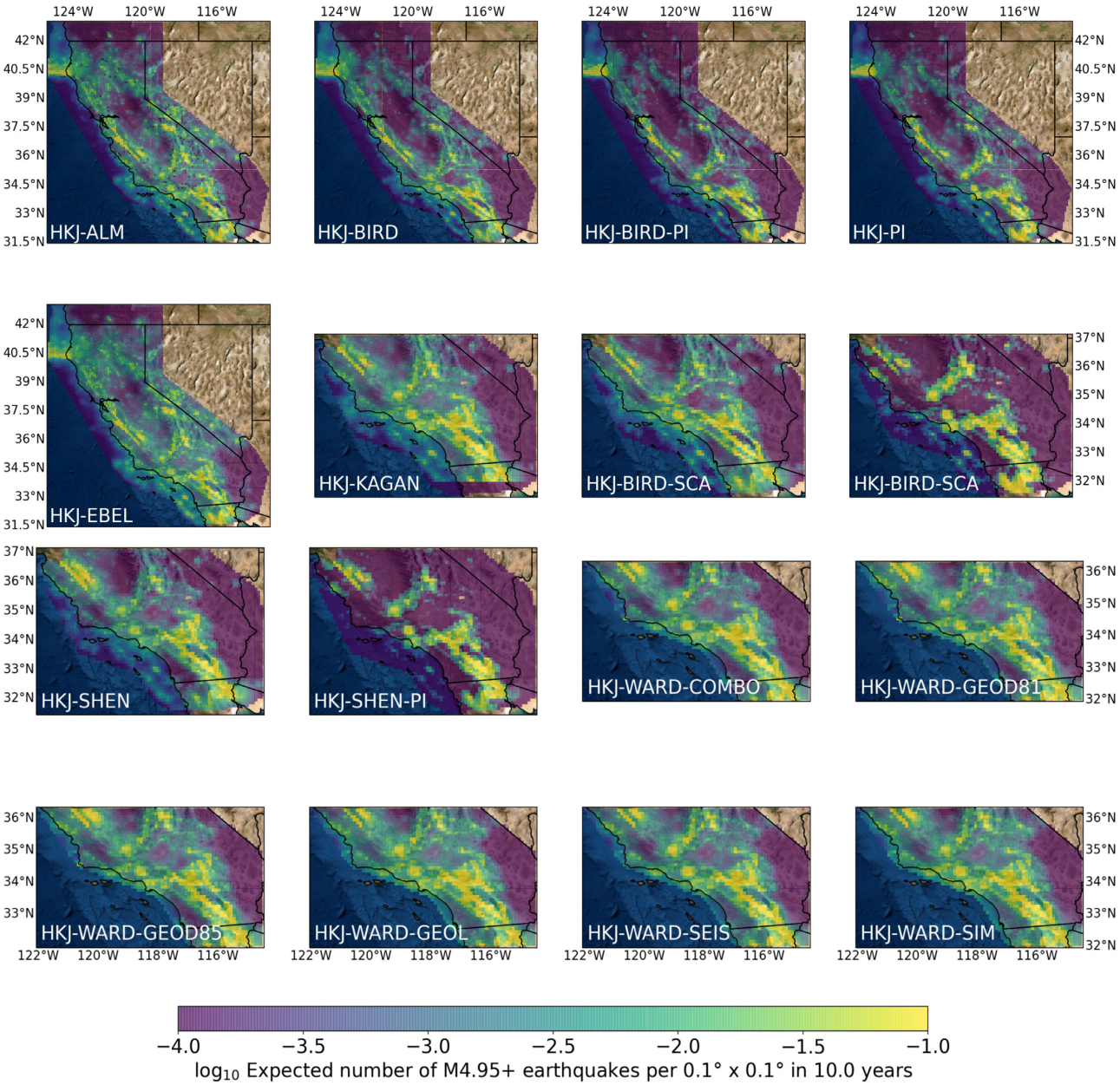

Figure 2. Hybrid earthquake models for California and southern California, constructed by Rhoades et al. (2014). These ensembles involve the HKJ model of Helmstetter et al. (2007) as baseline, and the ALM (Wiemer and Schorlemmer, 2007), BIRD (Bird and Liu, 2007), PI (Holliday et al. 2007), (corrected) EBEL-C (Ebel et al. 2007), KAGAN (Kagan et al. 2007), SHEN (Shen et al. 2007) and WARD-COMBO, WARD-GEOD81, WARD-GEOD85, WARD-GEOL, WARD-SEIS, WARD-SIM (Ward, 2007) conjugate models.

IGPE because the target events are independent of the fitted hybrids (see Section 4.1). However, we acknowledge that, although not obvious, the significance bounds for the IGPEc may not straightforwardly be the same as the ones derived for the IGPE.

3 DATA

We use $M \geq 4.95$ earthquakes reported in the Comprehensive Earthquake Catalog (ComCat) of the Advanced National Seismic System (ANSS; Guy et al. 2015) from 2011 January 1 to 2020 December 31 (see Table 2 and Fig. 3). Among others, this testing database includes the 2016 Hawthorne earthquake swarm (entries 16–18), as well as the M_W 6.4 foreshock (entry 22) and the M_W 7.1 main shock (entry 25) of the 2019 Ridgecrest sequence (Barnhart et al. 2019; Ross et al. 2019). We highlight these events, as they will be useful to assess the sensitivity of the results to clustering.

Table 2. Target earthquakes reported in the ANSS catalogue in California during the prospective evaluation period.

Event	Date	Latitude	Longitude	Magnitude
1	2011-02-18	32.047	−115.062	5.09
2	2012-02-13	41.143	−123.790	5.60
3	2012-07-21	40.412	−125.330	5.19
4	2012-08-26	33.017	−115.553	5.32
5	2012-08-26	33.019	−115.540	5.41
6	2012-10-21	36.310	−120.856	5.29
7	2013-02-13	38.022	−118.055	5.10
8	2013-05-24	40.191	−121.060	5.69
9	2014-03-10	40.829	−125.134	6.80
10	2014-03-29	33.933	−117.916	5.10
11	2014-08-24	38.215	−122.312	6.02
12	2015-01-28	40.318	−124.607	5.72
13	2016-06-10	33.432	−116.443	5.19
14	2016-08-10	39.329	−122.802	5.09
15	2016-12-14	38.822	−122.841	5.01
16	2016-12-28	38.376	−118.899	5.60
17	2016-12-28	38.390	−118.897	5.60
18	2016-12-28	38.378	−118.896	5.50
19	2017-07-29	40.782	−125.181	5.08
20	2018-04-05	33.838	−119.726	5.29
21	2019-06-23	40.274	−124.300	5.58
22	2019-07-04	35.705	−117.504	6.40
23	2019-07-05	35.760	−117.575	5.37
24	2019-07-06	35.725	−117.554	4.97
25	2019-07-06	35.901	−117.599	7.10
26	2019-07-06	35.901	−117.750	5.50
27	2019-07-06	35.904	−117.700	4.97
28	2019-07-06	35.910	−117.685	5.44
29	2019-07-06	40.392	−125.094	5.77
30	2020-03-09	40.348	−124.456	5.21
31	2020-03-18	38.053	−118.733	5.24
32	2020-04-11	38.169	−117.850	6.50
33	2020-05-15	38.181	−117.871	5.10
34	2020-05-15	38.201	−117.745	5.00
35	2020-05-20	38.231	−117.794	5.10
36	2020-05-22	35.615	−117.428	5.53
37	2020-06-24	36.447	−117.975	5.80
38	2020-06-30	38.154	−117.958	5.00
39	2020-11-13	38.169	−117.853	5.30
40	2020-12-01	38.164	−118.084	5.10

4 METHODS

4.1 Traditional CSEP tests

We use the number (N), spatial (S) and conditional likelihood (cL) tests implemented in the pyCSEP toolkit of CSEP (i.e. Savran et al., in preparation) to assess the consistency of the observed earthquakes with the forecasts (Schorlemmer and Gerstenberger 2007; Werner et al. 2010, 2011; Zechar et al. 2010). We do not present the results of the magnitude (M) test here, because all models use a Gutenberg–Richter frequency–magnitude relationship to distribute seismicity rates, rendering the results uninformative. In addition, we do not report the results of the likelihood L-test, due to its high sensitivity to the number of expected earthquakes (Werner et al. 2010, 2011; Zechar et al. 2013). Traditional CSEP tests are based on a probability function that approximates earthquakes as independent and Poisson-distributed:

$$\mathrm{P}(\omega|\lambda) = \frac{\lambda^{\omega}}{\omega!}\exp(-\lambda). \tag{3}$$

Eq. (3) describes the probability P of observing ω events given an expected rate/number of earthquakes λ. For the N-test, the total forecast rate is also Poissonian, with an expectation of $N_{\mathrm{fore}} = \sum\lambda(j, k)$, where the sum is over all space–magnitude bins (j, k). Thus, we evaluate if the overall number of observed earthquakes $N_{\mathrm{obs}} = \sum\omega(j, k)$ falls within the 95 per cent (and 97.5 per cent; see the problem of multiple tests section below) predictive range of this overall forecast rate distribution. Moreover, we compute the natural logarithm of the Poisson probability mass function, the log-likelihood (hereafter POLL), to measure how well the forecast can explain the observations in each

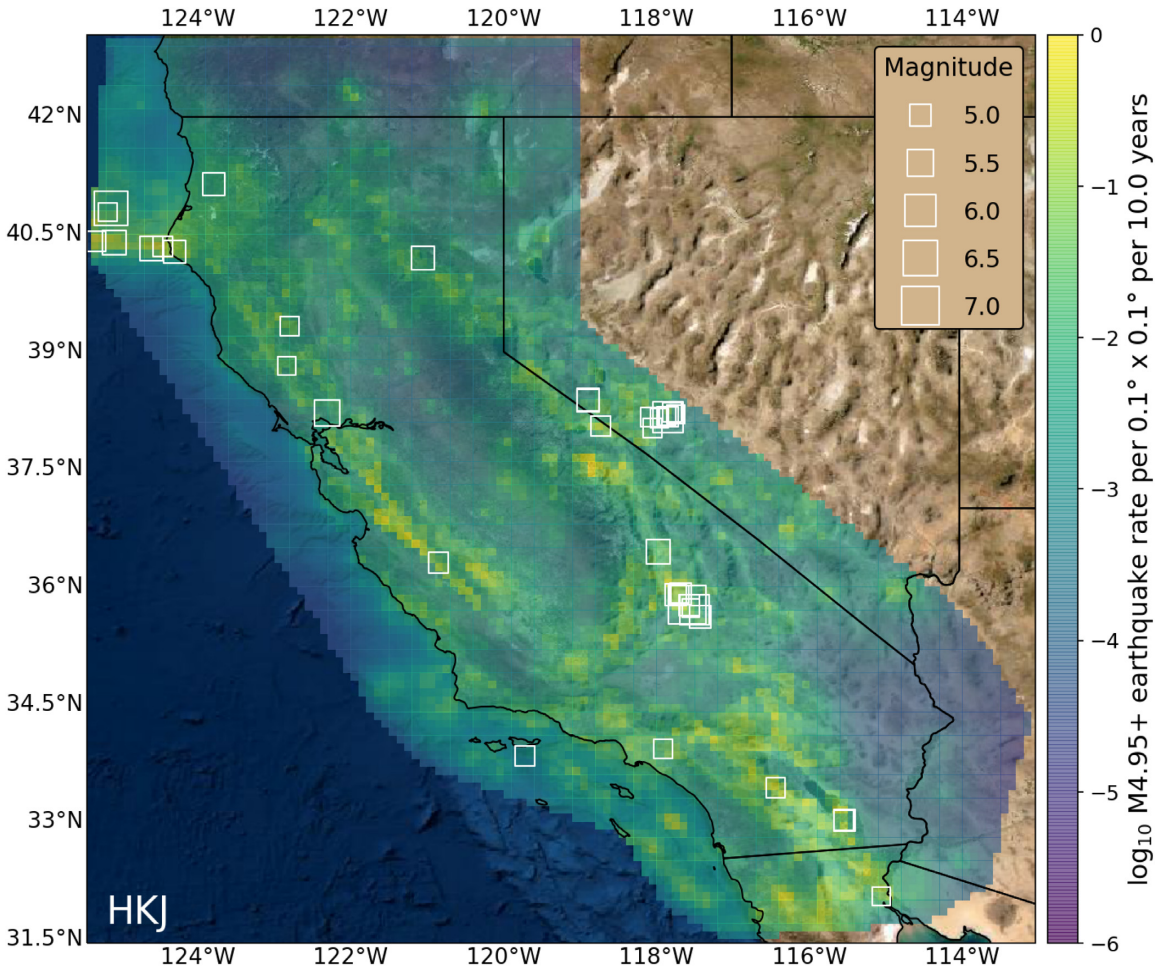

Figure 3. Hypocentral locations of $M \geq 4.95$ earthquakes observed in the CSEP-California testing region during the 2011–2020 prospective evaluation period of this study. We include a forecast map, showing the estimates of earthquake rate densities per decade provided by the baseline HKJ model.

cell/bin:

$$\mathrm{POLL} = \ln \mathrm{P}(\omega|\lambda) = -\lambda + \omega \ln\lambda - \ln(\omega!). \tag{4}$$

For the *cL*-test, we sum POLL scores over all space-magnitude bins (j, k) to obtain the observed joint log-likelihood based on the Poisson distribution (jPOLL) as:

$$\mathrm{jPOLL} = \sum_{j=1}^{l} \sum_{k=1}^{m} [-\lambda(j,k) + \omega(j,k) \ln\lambda(j,k) - \ln(\omega(j,k)!)]. \tag{5}$$

To account for forecast uncertainties, we simulate catalogs that are consistent with the number of observed earthquakes. First, we construct a discrete distribution normalizing and summing the rates in each space–magnitude bin, and assign in each bin the cumulative normalized sum. For a given simulation, the number of target events to simulate is fixed to N_{obs}. Then, for each simulated catalog, a random number from the uniform distribution over [0,1) is selected, and each earthquake is placed in the bin with the appropriate cumulative normalized sum value. We repeat the simulation process 10 000 times, obtaining a set of simulated catalogs, derived from a cell-wise homogeneous Poisson process conditional on the number of observed events, and computing for each of them a joint log-likelihood score. We thereupon compare whether the observed joint log-likelihood score falls within the 95 per cent and 97.5 per cent predictive ranges of the distribution of joint log-likelihood scores of the simulated data.

We also compute POLL scores for the spatial dimension of the forecasts to describe how consistent the observations are with the range of simulations. In this case, we sum over magnitude bins k and normalize the forecast sum, so that it matches the number of observed earthquakes (Zechar et al. 2010). Compared to the *cL*-test, the values of the normalized discrete distribution are based on the spatial distribution of the forecast, exclusively. For each simulated catalog, we compute a joint log-likelihood score and, after 10 000 simulations, we obtain a distribution of log-likelihood scores. We then evaluate if the observed joint log-likelihood score falls in the 95 per cent and 97.5 per cent predictive ranges of the one-sided distribution.

Finally, we compare the performance of hybrid and original RELM models with that of HKJ using the paired *T*-test developed by Rhoades et al. (2011). We select the HKJ forecast as the benchmark model to directly compare our prospective test results with the retrospective outcomes reported by Rhoades et al. (2014). This comparative test is based on the information gain per earthquake obtained by the HKJ benchmark model 1 over any conjugate or hybrid model *Z*, defined as:

$$\text{IGPE} = \frac{\hat{N}_1 - \hat{N}_Z}{N} + \frac{1}{N}\sum_{n=1}^{N}[X_Z(n) - X_1(n)]. \tag{6}$$

Using eq. (18) of Rhoades et al. (2011), we estimate the sample variance of $(X_Z(n) - X_1(n))$ as:

$$s^2 = \frac{1}{N-1} + \sum_{n=1}^{N}(X_Z(n) - X_1(n))^2 - \frac{1}{N^2 - N}\left[\sum_{n=1}^{N}(X_Z(n) - X_1(n))\right]^2. \tag{7}$$

Thus, the information gain intervals are computed as IGPE $\pm\, ts/\sqrt{N}$, where t is the appropriate quantile of the student t distribution with t_{N-1} degrees of freedom.

4.2 New CSEP tests

4.2.1 Negative-binomial N-test

The Poisson distribution insufficiently captures the variability of earthquake activity, as it could be underdispersed with respect to the true distribution of seismicity (Werner and Sornette 2008; Lombardi and Marzocchi 2010). In contrast, the negative binomial distribution (NBD) has been shown to better describe non-declustered activity, because it has a greater variance than the Poisson distribution to account for spatiotemporal earthquake clustering (Vere-Jones 1970; Jackson and Kagan 1999; Kagan 2010; Werner et al. 2010, 2011). Accordingly, we fit an NBD to the rates of $M \geq 4.95$ earthquakes in each forecast's testing region to evaluate number consistencies between the models and the observations. The probability mass function of an NBD is defined by:

$$p(\omega \mid \tau, \nu) = \frac{\Gamma(\tau + \omega)}{\Gamma(\tau)\ \omega!}\nu^{\omega}(1-\nu)^{\omega}, \tag{8}$$

where $\omega = 0, 1, 2,...$ is the number of events, $\tau > 0$ and $0 \leq \nu \leq 1$ are parameters, and Γ is the Gamma function. The mean μ and variance σ^2 of the NBD are given by

$$\mu = \tau\frac{1-\nu}{\nu};\ \sigma^2 = \tau\frac{1-\nu}{\nu^2}. \tag{9}$$

Following the approach by Werner et al. (2010), we use the expected number of earthquakes by each forecast as the mean and estimate the variance from historical observations reported in the ANSS catalog from 1932 until the end of 2010, in non-overlapping 10 yr periods within each forecast's unmasked test region. In this manner, we obtain a variance $\sigma_C^2 \approx 314.21$ for the whole of California, as well as $\sigma_{sCS}^2 \approx 185.70$ and $\sigma_{sCW}^2 \approx 164.83$ for the southern California SHEN and WARD testing regions, respectively.

4.2.2 Binary cL- and S-tests

As discussed above, the likelihood function used in CSEP evaluations cannot fully account for the likely dependence of multiple earthquakes that are closely clustered in space and time. Therefore, we introduce a binary likelihood function that reduces the sensitivity of CSEP tests to clustering by calculating the probability of earthquake activity in a forecast bin, rather than the likelihood of observing $\omega = 1, 2, 3,..,$ earthquakes. Based on eq. (3), the probability of observing $\omega = 0$ events given an expected number/rate λ is $p_0 = \exp(-\lambda)$, while the probability of observing more than zero events, or any activity, is $p_1 = 1 - p_0 = 1 - \exp(-\lambda)$. Thus, we calculate the log-likelihood score based on the binary (or Bernoulli) distribution (BILL) as:

$$\text{BILL} = X_i\ \ln\ p_1 + (1 - X_i)\ \ln\ p_0 = X_i\ \ln\ (1 - \exp(-\lambda)) + (1 - X_i)\ \ln\ (\exp(-\lambda)), \tag{10}$$

where the first term represents the contribution to the score if a cell/bin i contains one or more events, that is, $X_i = 1$, and the second term is the contribution if that cell/bin contains no earthquakes, that is, $X_i = 0$. According to eqs (4) and (10), substantial differences between POLL and BILL scores are expected in bins where $\omega \geq 2$ (note that POLL = BILL if $\omega = 0$, and POLL $\approx$ BILL if $\omega = 1$ and $\lambda \to 0$).

Similar to the traditional *cL*-test, we sum BILLs over all space-magnitude bins (j, k) to compute the observed joint log-likelihood based on the binary distribution (jBILL) as:

$$\text{jBILL} = \sum_{j=1}^{l}\sum_{k=1}^{m}[X(j,k)\ \ln\ (1 - \exp(-\lambda(j,k))) + (1 - X(j,k))\ \ln\ (\exp(-\lambda(j,k)))] \tag{11}$$

For the binary *S*-test, we first normalize the forecast rate $\lambda(j)$ to the number of active cells, and then we calculate the joint log-likelihood score as:

$$\text{jBILL} = \sum_{j=1}^{l} [X(j)\ \ln\ (1 - \exp(-\lambda(j))) + (1 - X(j))\ \ln\ (\exp(-\lambda(j)))] \tag{12}$$

The simulation procedure to account for forecast uncertainties in these binary tests is similar to that implemented for Poisson *S*- and *cL*-tests except that we simulate catalogs that are consistent with the number of active cells/bins.

4.2.3 Binary T-test

The Poisson distribution computes the probability of observing $\omega = 1, 2, 3,...$ earthquakes in a bin, given the expected value λ. Accordingly, the paired *T*-test of Rhoades et al. (2011) uses the earthquake information gain score as a comparative measure to evaluate model performances. In contrast, the binary likelihood function estimates the likelihood of an observation ω to be either zero or non-zero in a cell/bin, so that a fair measure to compare performance between models is the information gain score per active bin (IGPA):

$$\text{IGPA} = \frac{\hat{N}_1 - \hat{N}_Z}{M} + \frac{1}{M}\sum_{n=1}^{M} [X_Z(n) - X_1(n)]. \tag{13}$$

Note that eq. (13) is similar to eq. (6), except for the number of observed earthquakes *N*, which has been replaced here by the number of bins *M* containing activity. Finally, we compute confidence intervals for IGPAs using a similar approach to eq. (7), for which we also substitute the observed number of earthquakes *N* for the number of active bins *M*:

$$s^2 = \frac{1}{M-1} + \sum_{n=1}^{M}(X_Z(n) - X_1(n))^2 - \frac{1}{M^2 - M}\left[\sum_{n=1}^{M}(X_Z(n) - X_1(n))\right]^2. \tag{14}$$

4.3 The problem of multiple tests

CSEP and other statistical tests are based on rejecting the null hypothesis if the likelihood of the observed data under the null hypothesis is low (Schorlemmer et al. 2007; Werner et al. 2010; Zechar et al. 2013). Performing multiple tests among samples of the same distribution increases the probability of observing at least one rare event and, consequently, increases the probability of incorrectly rejecting a null hypothesis and obtaining an apparently statistically significant result (Kato, 2019). Hence, a correction of the significance level α is required to control the overall type I error rate, that is, the false-positive rate or the 'false-inconsistency' rate. Despite some drawbacks (Perneger 1998; Armstrong 2014), a Bonferroni-adjusted significance level α/T has been commonly used to circumvent the multiple tests problem, where *T* is the number of independent tests (e.g. Cheverud 2001; Cass and Tromans 2008; Pettit et al. 2020).

At CSEP, we do not formally 'reject' a model at a 0.05 significance level when it fails at least one of the tests, but rather we use the quantile scores and test results as indicators of potential discrepancies between data and models that might be of scientific interest. Therefore, we report *p*-values calculated for each original RELM and multiplicative hybrid earthquake model (see Supporting Information of this manuscript and Figs 5, 7 and 9) and primarily discuss consistency test results at a significance level $\alpha = 0.05$. In addition, we include consistency test results at a Bonferroni-adjusted significance level $\alpha_B = 0.05/2 = 0.025$ to account for the overall type I error rate. To justify this correction, we use the HKJ model as a data generator to simulate random numbers of events, magnitudes and locations from its Poisson distributions a thousand times. Then, we calculate quantile scores for each consistency test and estimate the correlation coefficients between them. Thus, we find that the *S*- and *cL*-tests are highly correlated with each other, whilst both are fairly independent from the *N*-test (see Fig. 4).

5 RESULTS

5.1 *N*-test results

Prospective *N*-test results show that most of the forecasts overestimate earthquake activity in California during the 10-yr evaluation period, causing all forecasting models to fail the Poisson *N*-test while only a few pass the negative-binomial *N*-test (see Fig. 5). Interestingly, the discrepancies between forecasts and observations are greater in southern California, as only the KAGAN and SHEN models pass the NBD *N*-test at a 0.05 significance level, and only HKJ-BIRD-SCA and HKJ-PI-SCA pass the test at a Bonferroni-corrected significance level of 0.025. *N*-test results for the BIRD forecast are particularly interesting, as the model aims to reflect the long-term rates that are derived from tectonic moment release. Nonetheless, prospective test results indicate that not as many earthquakes occurred along the San Andreas plate boundary system during the prospective evaluation period as expected by the model.

N-test results are due to the fact that the evaluation period contained a relatively low number of $M \geq 4.95$ earthquakes, compared to other 10-yr periods in the target regions (see Fig. 6). During the 1990s, for instance, the ANSS catalog reported 83 events in the California-CSEP

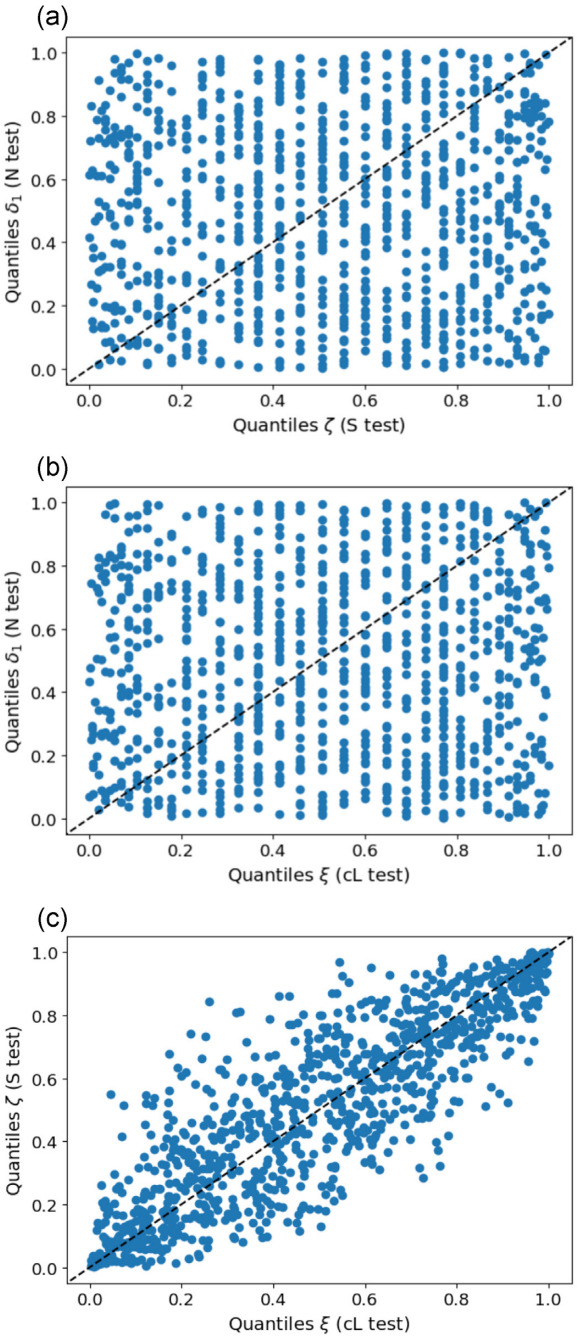

Figure 4. (a) N- and S-quantile scores obtained from simulated earthquake catalogs. We use the *N*-test δ_1 statistics to compare it against the quantile score ζ of the *S*-test. The correlation coefficient between these scores is $R_{\mathrm{NS}} = 0.01$. For comparison, we include a diagonal dashed line representing a perfect one-to-one correlation between the scores. (b) Scatter plot showing *N*-test quantile scores δ_1 against *cL*-test quantile scores ξ. The correlation coefficient between these statistics is $R_{\mathrm{NcL}} = 0.03$, which exhibits independence between the tests. (c) *S*-test quantile scores ζ against *cL*-test quantile scores ξ obtained by simulated catalogues. The correlation coefficient between these quantile scores is $R_{\mathrm{ScL}} = 0.86$.

testing region, including the 1992 M_{W} 6.1, 7.4, 6.2 Joshua Tree–Landers–Big Bear sequence (Hauksson et al. 1993), and the 1999 M 7.1 Hector Mine quake (Freed and Lin, 2001). During the 2001–2011 period, the catalog reported 53 target events, of which 31 were used to calibrate hybrid models for California and 22, including the 2010 M_{W} 7.2 El Mayor–Cucapah earthquake (Hauksson et al. 2011), were used to fit hybrids for southern California. Consequently, hybrid models in southern California expect a rate of 44 earthquakes per decade, which is significantly larger than the 13 and 16 events observed over the past 10 yr in the SHEN and WARD testing regions.

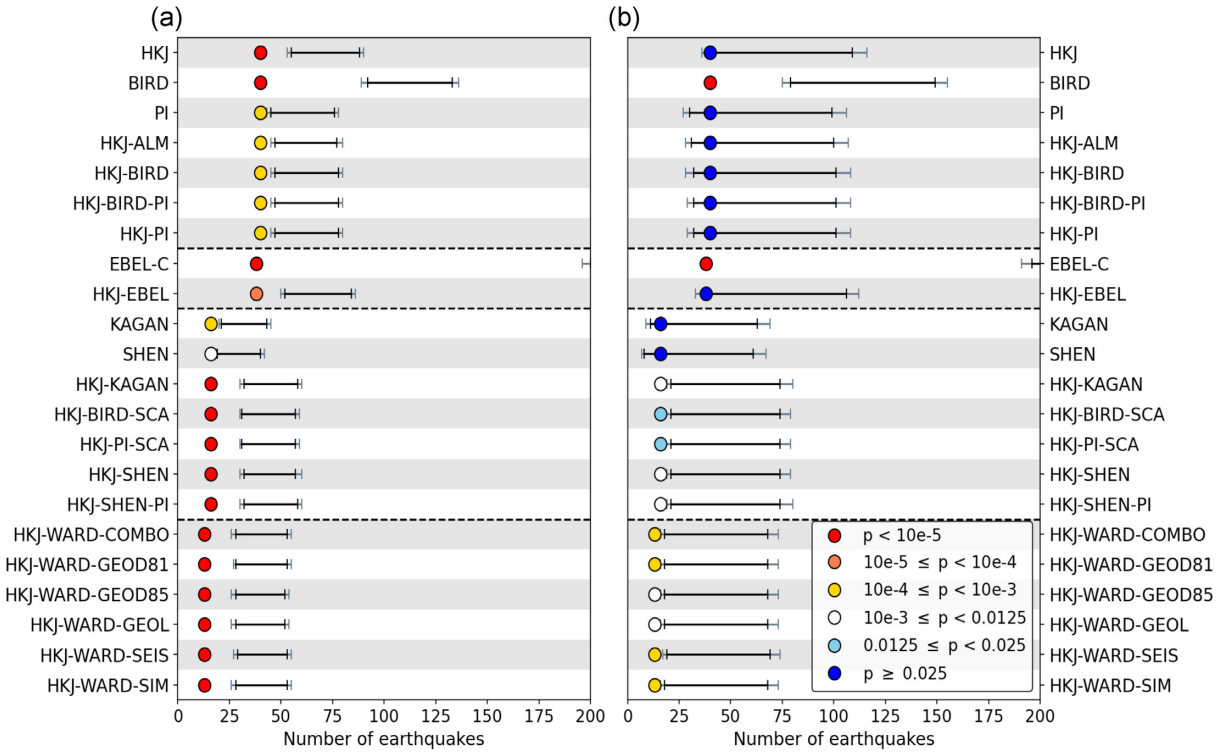

Figure 5. Results of the prospective (a) Poisson and (b) NBD *N*-tests during the 2011–2020 evaluation period. The circles represent the number of observed earthquakes and the colours denote the calculated *p*-values for earthquake forecasting models, obtained from the equation $p\text{-value} = 2 \cdot \min(P(X) \leq x, P(X) > x)$ proposed by Meletti et al. (2021). Blue colours indicate consistencies between forecasts and the observations and red-orange colours indicate the opposite. Solid black and gray bars depict the 95 per cent and 97.5 per cent predictive intervals of the model forecast likelihood distributions. Horizontal dashed lines separate different testing regions. The EBEL-C predictive range is too large to be shown.

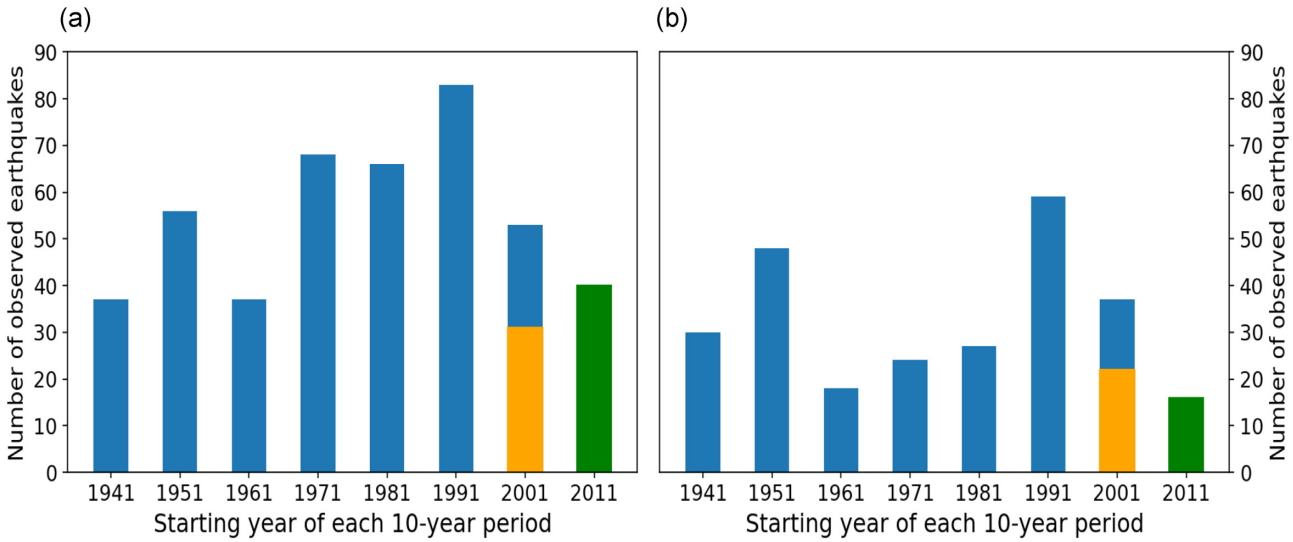

Figure 6. Number of $M \geq 4.95$ earthquakes reported within (a) the CSEP California and (b) the SHEN southern California testing regions during non-overlapping 10-yr periods by the ANSS catalog. We highlight in orange the number of earthquakes that were used to fit hybrid models and, in green, the prospective target events.

5.2 *S*-test results

S-test results reveal that most models struggle to explain the spatial distribution of epicenters, as only the KAGAN forecast passes the Poisson *S*-test, and only EBEL-C, KAGAN, SHEN, HKJ-WARD-GEOL, HKJ-WARD-SEIS and HKJ-WARD-SIM pass the binary *S*-test (see Fig. 7). By using a Bonferroni-adjusted significance level of 0.025, the EBEL-C and SHEN forecasts also pass the Poisson *S*-test, whilst the HKJ-EBEL, HKJ-WARD-GEOD81 and HKJ-WARD-GEOD85 hybrid models exhibit similar results when undergoing the binary *S*-test. In general, we detect that the observed joint log-likelihood score obtained by the baseline HKJ forecast is systematically better than the

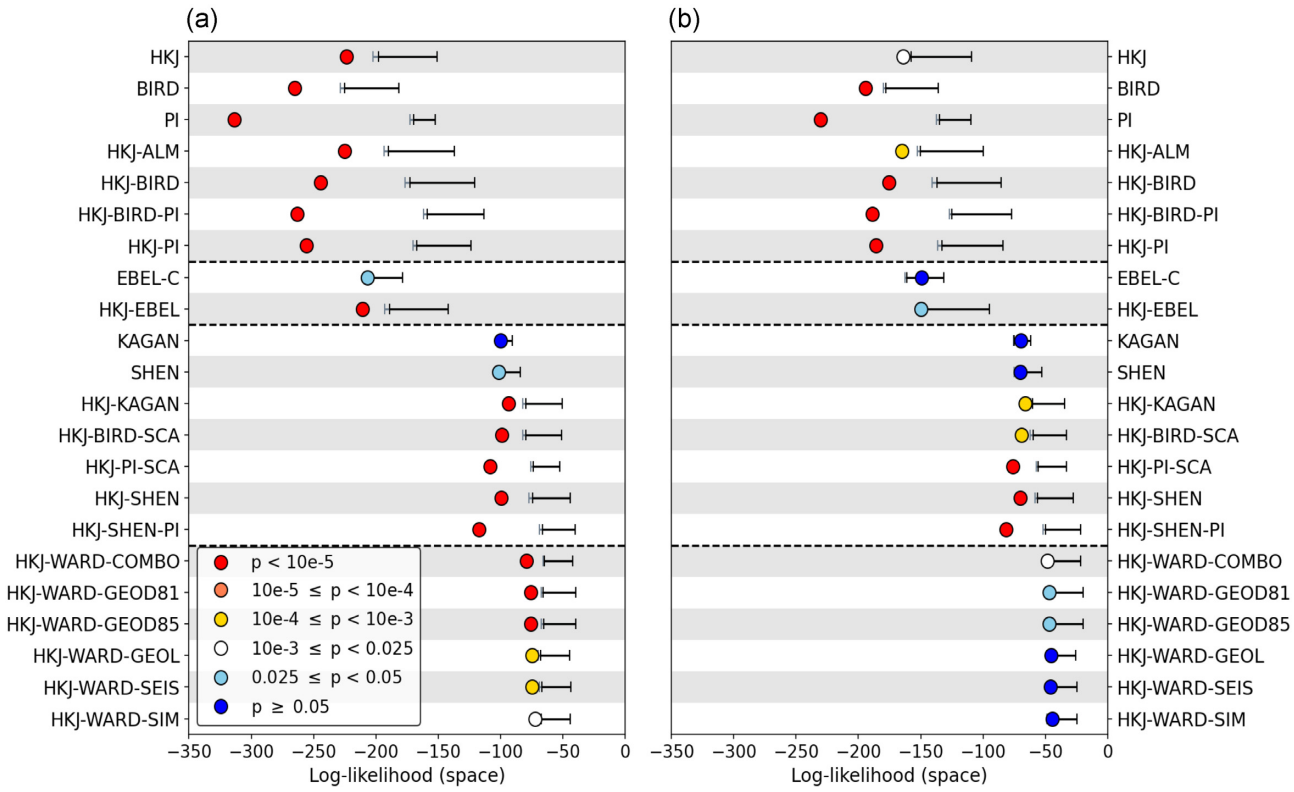

Figure 7. Results of the one-sided *S*-test for forecasts in California based on (a) a Poisson and (b) a binary likelihood function. Symbols depict observed joint (spatial) log-likelihood scores and the colours denote *p*-values calculated for each forecasting model. Blue colours indicate consistencies between forecasts and the observations, while red-orange colours indicate the opposite. Solid horizontal black and grey bars represent the 95 per cent and 97.5 per cent predictive intervals of the model forecast probability distributions.

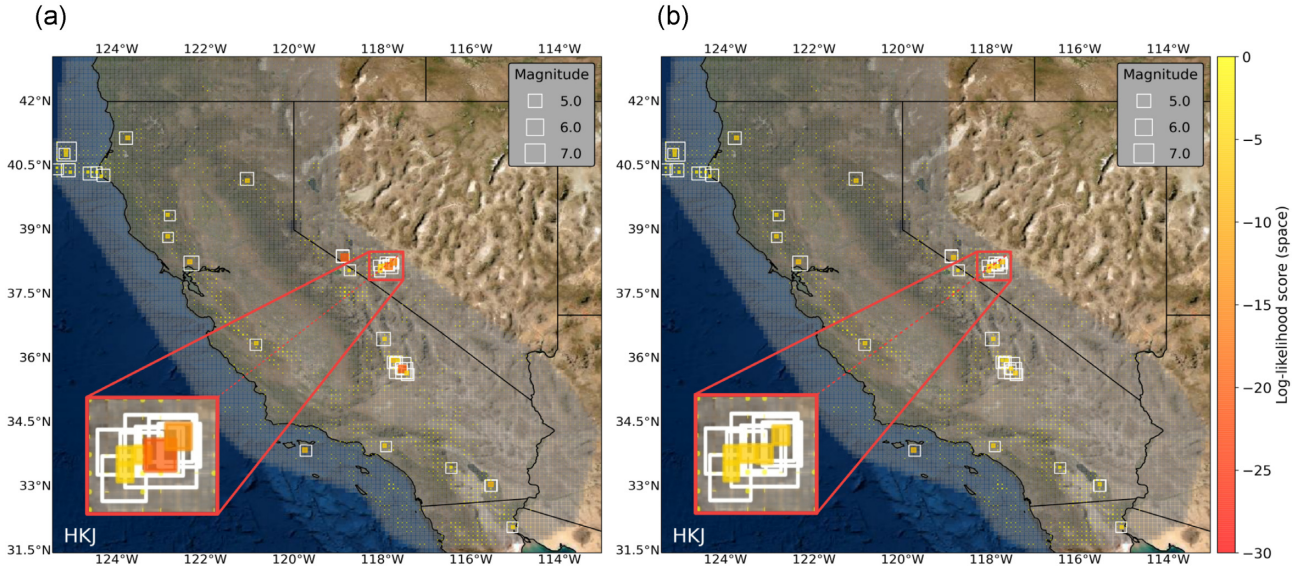

Figure 8. Spatial distribution of log-likelihood scores obtained in each spatial cell by the HKJ forecast, using (a) a Poisson and (b) a binary likelihood function. The Poisson-based *S*-test penalizes the model for the unlikely occurrence of the 2016 Hawthorne earthquake swarm (zoomed-in) and the 2019 Ridgecrest sequence in a few spatial cells more severely than the *S*-test, which relies on a binary probability function. White squares denote the epicentral locations of the $M \geq 4.95$ target earthquakes.

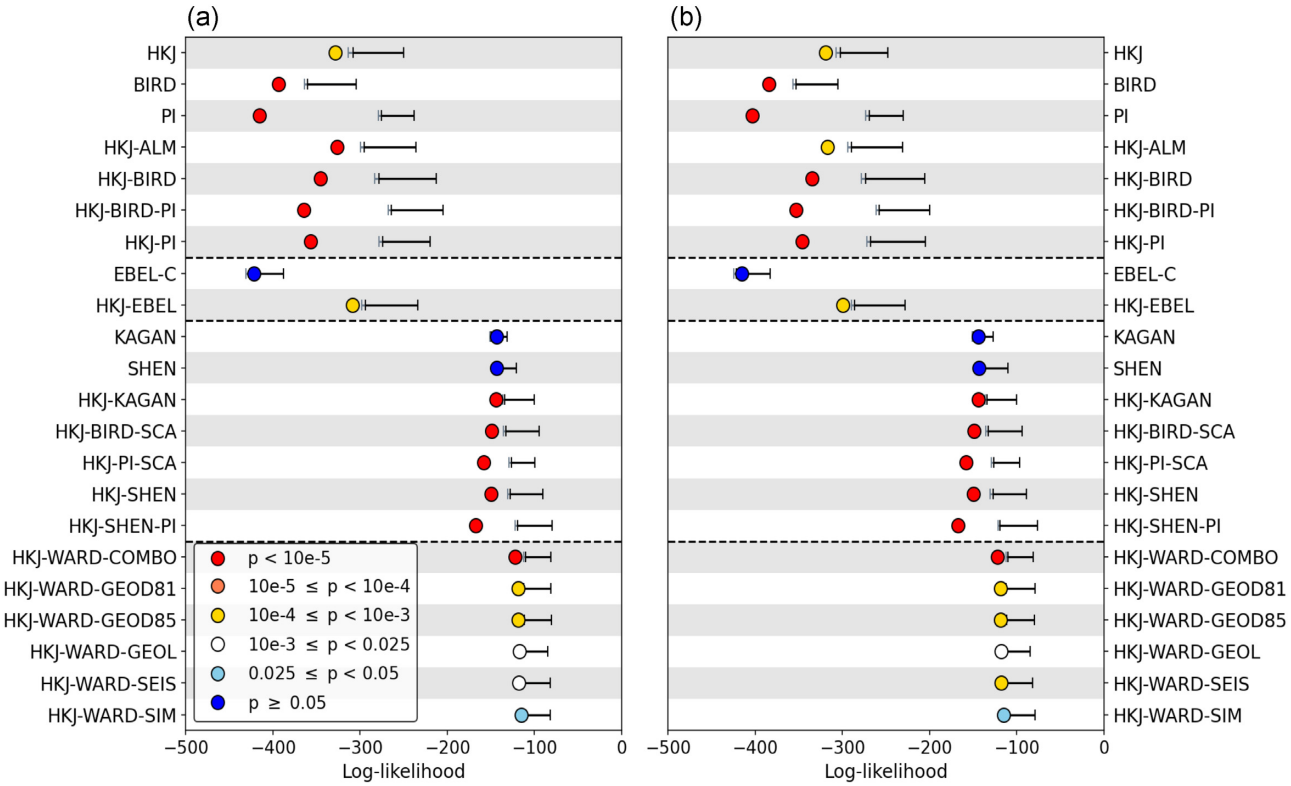

Figure 9. *cL*-test results for forecasts in California based on (a) a Poisson and (b) a binary probability function. The circles denote observed joint (spatial-magnitude) log-likelihood scores and colours represent the *p*-values calculated for the forecasting models. Blue colours indicate consistencies between the forecasts and the observations and red-orange colours denote the opposite. Solid horizontal black and grey bars represent the 95 per cent and 97.5 per cent predictive ranges of the model forecast likelihood distributions. Discrepancies between Poisson and binary *cL*-test results for forecasting models in the whole of California are negligible, because the evaluation period contains 40 earthquakes that fell into 39 space–magnitude bins. For models in southern California, the differences between Poisson and binary *cL*-test results are even smaller, as the number of observed earthquakes equals the number of active bins.

joint log-likelihood scores calculated for its hybrid model (e.g. HKJ-BIRD, HKJ-PI) and the conjugate model (e.g. BIRD, PI). In addition, we observe that the discrepancies between the expected and observed spatial distribution of earthquakes are more prominent for forecasts involving the PI model.

Poisson *S*-tests results are strongly affected by the occurrence of the 2016 Hawthorne earthquake swarm and the 2019 Ridgecrest sequence in a few cells. In Fig. 8, we highlight that HKJ obtains a POLL score of about −20 in the spatial cell where the Hawthorne 'triplet' was observed, which is significantly worse than the BILL score of approximately −6 that we compute for the model in the same spatial cell. Thus, BILL scores are less dependent on the range of likelihood that stems from clustering. Nonetheless, binary *S*-test results show that the observed distribution of earthquakes during the evaluation period is inconsistent with the distribution predicted by HKJ, which indicates that the target earthquakes nucleated in regions of low probability for the model.

5.3 *cL*-test results

Prospective results of the *cL*-test confirm that the occurrence of earthquakes during the 2011–2020 evaluation period is unlikely for most models, including all hybrids and the HKJ, BIRD and PI RELM forecasts (see Fig. 9). In the entire California testing region, only the EBEL-C model adequately describes the likelihood distribution of observed seismicity, while KAGAN and SHEN are the only models that pass the *cL*-tests in southern California, at a 0.05 significance level. When considering a Bonferroni-corrected significance level of 0.025, only the EBEL-C, KAGAN, SHEN and HKJ-WARD-SIM forecasts pass the Poisson and binary *cL*-tests. In consistency with *S*-test results, we find that models involving the PI forecast are particularly limited in their abilities to adequately describe the observed likelihood distribution of earthquakes. These observations are contrary to the PI hypothesis that, in the near future (5–10 yr), large earthquakes are likely to occur in regions where strongly fluctuating seismicity has been observed.

Discrepancies between Poisson and binary *cL*-test results are negligible, because the evaluation period contains 40 earthquakes that fall into 39 space–magnitude bins, that is, there is only one spatial cell in which two earthquakes within the same magnitude bin occurred (see entries 16 and 17 in Table 2). As previously discussed, significant differences between POLL and BILL scores can only be obtained in cells/bins containing two or more events. Thus, *cL*-test results are largely insensitive to the binary likelihood function, because only one space–magnitude bin contains more than one event during the target period.

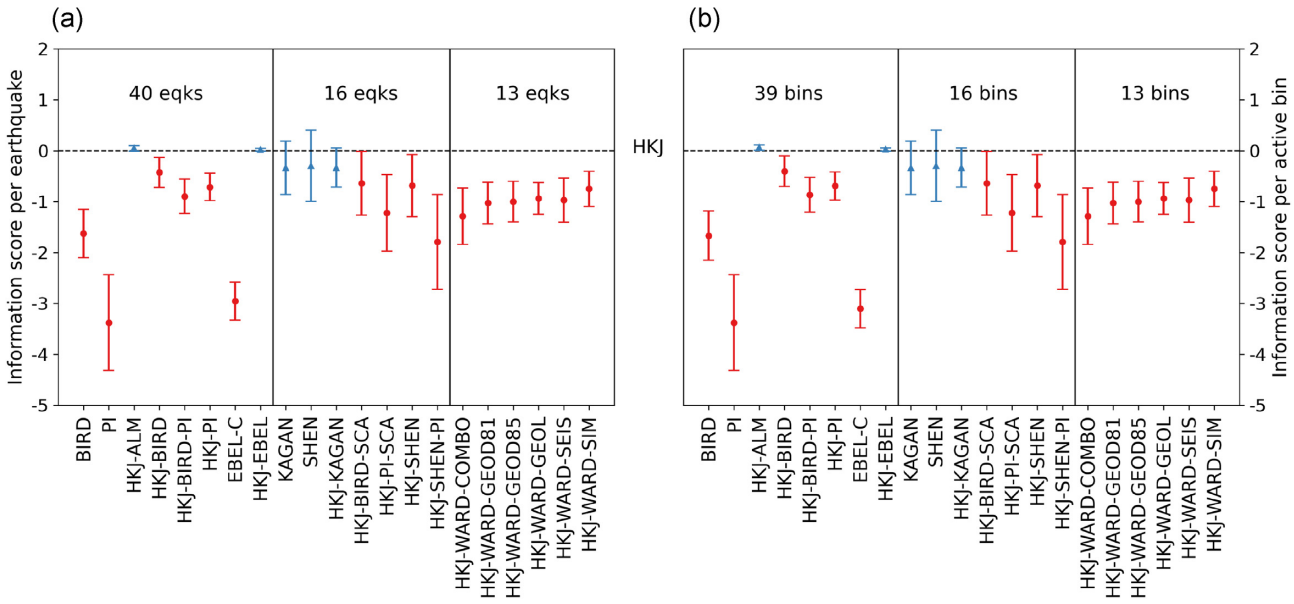

Figure 10. Comparison of information gains *T*-test results. Information gains per (a) earthquake and (b) active bin are presented as triangles and circles, depending on their relative values to that of the HKJ model: blue triangles denote that the model is statistically as informative as HKJ and red circles indicate that a model is significantly less informative than HKJ.

5.4 *T*-test results

Comparative *T*-test results show that the HKJ-ALM, HKJ and HKJ-EBEL are the most informative earthquake models for the whole of California during the prospective evaluation period (see Fig. 10). The IGPEs obtained by the two hybrids over the HKJ benchmark model, however, are not statistically significant at a 0.05 significance level. In southern California, KAGAN, SHEN and HKJ-KAGAN are statistically as informative as the HKJ model, and none of the hybrids involving the forecasts developed by Ward (2007) outperformes HKJ.

Similar to the *cL*-test results, *T*-test results based on information gain scores per active bin are largely insensitive to the binary likelihood function, as differences between POLL and BILL scores are rather insignificant. Discrepancies between *T*-test results for models in southern California are even lower, because the numbers of earthquakes and active bins are identical. These similarities are due to the relatively small number of earthquakes provided by the evaluation period, which implies a low probability of observing more than one event in the same space–magnitude bin.

6 DISCUSSION

Consistency test results show that most forecasting models inadequately describe seismicity patterns in California during the 2011–2020 target period. Regarding the number distribution of earthquakes, all models overestimate seismicity rates, thus failing the Poisson *N*-test, and only a few pass NBD *N*-test at significance levels of 0.05 and 0.025 (see Fig. 5). As discussed above, we interpret these results to be primarily due to a period of relative quiescence compared to other 10-yr observation periods in California (see Fig. 6). Similar to the number distribution of observed earthquakes, the spatial distribution of epicenters is also unlikely for most models, as reflected in prospective *S*-test results (see Fig. 7). In the case of HKJ, these results are derived from 7652 empty cells, 24 single-quake cells, 3 two-quake cells, 2 three-quake cells and 1 four-quake cell that cumulatively contribute 17 per cent, 44 per cent, 13 per cent, 17 per cent and 9 per cent to the jPOLL score, respectively (see Fig. 11). Thus, the Poisson *S*-test results are significantly influenced by the unlikely occurrence of the 2016 Hawthorne earthquake swarm and the 2019 Ridgecrest sequence.

The high dependence of Poisson-based consistency test results on the occurrence of a few clustered events has been widely discussed in other forecast evaluations (e.g. Werner et al. 2010; Taroni et al. 2018; Rhoades et al. 2018). Therefore, we introduce in this study a binary likelihood function to reduce the sensitivity of POLL scores to clustering. This function can be useful in assessing long-term, time-invariant seismicity models, given the lack of consensus on an optimal declustering scheme and potential misclassifications of foreshock, mainshock and aftershock activity (Bird et al. 2015; Strader et al. 2018). We detect substantial differences between Poisson and binary log-likelihood scores only in the spatial dimension of the forecasts (see Figs 7 and 8). For the HKJ model, for instance, we observe that single-, two-, three- and four-quake cells contribute overall 23 per cent, 59 per cent, 8 per cent, 7 per cent and 3 per cent to the jBILL score, respectively (see Fig. 11). Compared to POLL estimates, these scores depend less on the low likelihood scores that stem from clustered seismicity, however, HKJ is still limited in its ability to forecast the location of earthquakes, because the target events occurred in regions of low probability.

Poisson and binary *cL*-test results confirm that the probability of occurrence of earthquakes during the evaluation period is low for most forecasting models (see Fig. 9). We can rule out that these results are derived from limitations of traditional CSEP tests to accurately

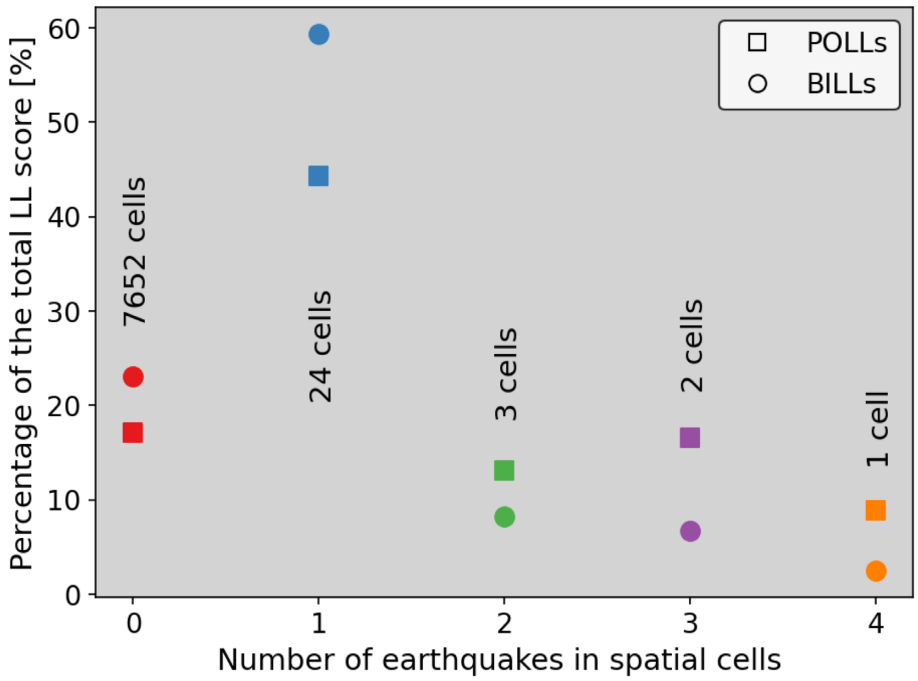

Figure 11. Percentage of the observed joint log-likelihood score obtained by the HKJ model that is due to the occurrence of zero (red), one (blue), two (green), three (purple) and four (orange) events in single spatial cells. We represent POLLs and BILLs with squares and circles, respectively.

capture seismicity clustering, as differences in space–magnitude bins between Poisson and binary *cL*-tests are negligible. These similarities are due to the fact that only one spatial cell contains more than one event within the same magnitude bin during the 2011–2020 target period. Nonetheless, we expect substantial variations between Poisson and binary *cL*-test results in subsequent prospective analyses with larger spatial cells, lower magnitude thresholds or longer target periods. Thus, we consider that, in general, the binary likelihood function could be useful in relaxing the Poisson assumption in future CSEP evaluations.

Comparative *T*-test results for the six RELM models are in agreement with the results reported by Schorlemmer et al. (2010), Zechar et al. (2013) and Strader et al. (2017), as we find that HKJ is the most informative seismicity model during the 2011–2020 evaluation period (see Fig. 10). Our results provide further evidence that the locations of small M2+, past earthquakes can contain more predictive skills of moderate-to-large earthquakes over a 5- to 15-yr period than many other forecast approaches that depend on fault, geodetic and tectonic data (Schorlemmer et al. 2018). Models that explicitly incorporate known faults or interseismic-strain rates are thought to provide better long-term forecasts than models lacking such information (Bird and Liu 2007; Field et al. 2009; Bayona Viveros et al. 2019). Therefore, we speculate that the relative performance of fault-, geodesy- and tectonic-based models like WARD-GEOL, WARD-GEOD81 and BIRD with respect to HKJ could improve in further prospective evaluations after the occurrence of large earthquakes along the major known faults in California.

Prospective *T*-test results for hybrid earthquake approaches are contrary to the retrospective results by Rhoades et al. (2014), who found that some forecasts like BIRD, PI and SHEN are effective conjugates with HKJ to gain predictive skill. In the retrospective evaluation, these hybrids obtained information gain scores of approximately 0.25, 0.25 and 0.5 over HKJ, although only the HKJ-SHEN model proved to be significantly more informative than the reference model. In contrast, our prospective results show that the hybrids involving the BIRD, PI and SHEN models obtain information scores relative to HKJ of about −0.42, −0.71 and −0.68, respectively (see Table S1 of the Supporting Information of this manuscript). We interpret these results to be due to four possible reasons: (1) based on the target events during the 5-yr RELM period, Rhoades et al. (2014) calculated optimal blending parameters and IGPEc's with associated confidence intervals that could vary significantly over time. Therefore, we infer that the discrepancies between retrospective and prospective *T*-test results might be due to temporal instabilities of the fit to form hybrid models that are derived from the relatively small number of target events used for their calibration. (2) These differences may also be due to the absence of large on-fault earthquakes during the target period, favouring the HKJ model over other earthquake approaches that use tectonic plate velocities, fault slip data or interseismic strain rates. (3) According to retrospective analyses during the 2006–2010 RELM period, the BIRD, PI, EBEL-C, KAGAN and SHEN models obtained IGPE scores of −0.70, −0.31, −1.64, −1.04 and −0.53 over HKJ, respectively. In contrast, these models obtain IGPEs of about −1.62, −3.38, −2.95, −0.33 and −0.29 over HKJ during the 2011–2020 prospective evaluation period (see Table S1 of the Supporting Information of this manuscript). Thus, we observe that the forecasting power of the models for the whole of California, that is, BIRD, PI and EBEL-C, decreased during the last decade with respect to HKJ, whilst the KAGAN and SHEN models for southern California showed a better performance over time (although still poorer than that of HKJ). Hence, discrepancies between prospective and retrospective test results for multiplicative hybrids could additionally be explained in terms of the loss of predictive ability of their conjugate models over time. (4) The spatial *S*-test results give a further clue to understanding the relatively poor performance of the multiplicative hybrids in the 2011–2020 period. As previously discussed, the multiplicative hybrid

models all had the same baseline model (HKJ) and were all fitted to the target earthquakes during the first five years of the RELM experiment. Through the fitting process, all assumed the frequency–magnitude distribution of the baseline model and the total expected rate of target earthquakes was normalized to match the observed rate in the first five years. This rate was only slightly different from that of the baseline model, which passed the *N*-test well in the first five years in all model classes. Therefore, the main difference between the multiplicative hybrids and the HKJ baseline model is in their spatial distributions of expected earthquakes. In the 2006–2010 period, nearly all the original models passed the *S*-test comfortably (Zechar et al. 2013). Although the other models were less informative overall than HKJ, it is not surprising that, as conjugate models with HKJ as a baseline, some of them could be used to form multiplicative hybrids that would have outperformed the baseline model over that five-year period. In contrast, in the 2011–2020 period most models—including the original models HKJ, BIRD and PI—convincingly failed the *S*-test (Fig. 7). The original models that passed the *S*-test—EBEL-C, KAGAN and SHEN—had log likelihoods towards the low end of the passing range. In multiplicative hybrids, cells with low spatial rates in a conjugate model tend to have low multipliers (less than one), so that the spatial rates for these cells tend to be even lower in the hybrid than in the baseline model. Target earthquakes that occur in such cells contribute to the hybrid performing worse than the baseline model. Therefore, it is not surprising that the multiplicative hybrids, including those that strongly outperformed HKJ over the 2006–2010 period, did not perform as well as HKJ over the 2011–2020 period.

Despite these prospective results, we recommend further investigation into ensemble earthquake forecasting by testing additive combinations involving correlated forecasting models and multiplicative mixtures comprising independent model components. On a global scale, a multiplicative log-linear combination of geodetic strain rates and earthquake-catalogue information has been used to construct the hybrid Global Earthquake Activity Rate (GEAR1; Bird et al. 2015) model. The optimized blend of the parent model components of GEAR1 was empirically determined by maximizing the I_1 (success) information score of Kagan (2009) using 1085 $M \geq 5.95$ earthquakes. After the occurrence of 651 independent target events, the continued outperformance of the hybrid over its individual constituents has been documented by Bird (2018), Strader et al. (2018) and Bayona et al. (2021). Although the stability of this technique seems to be given by the number of events used in its calibration, we believe that this method might be worth comparing with that of Rhoades et al. (2014) to identify the blends that can provide the greatest possible information gains in California.

7 CONCLUSION

In this study, we prospectively assessed 6 original RELM forecasts and 16 multiplicative hybrid models that involve the smoothed-seismicity HKJ model as a baseline and one or two other models as conjugates. We used a set of traditional and new CSEP tests that depend on a Poisson and a binary likelihood function, respectively. According to our consistency test results, most forecasts overestimate seismicity rates and fail to forecast the spatial distribution of earthquakes, especially the location of clusters of seismicity, such as the 2016 Hawthorne earthquake swarm and the 2019 Ridgecrest sequence. The binary likelihood function significantly reduces the sensitivity of spatial log-likelihood scores to the occurrence of these events, however; most models still inadequately describe spatial seismicity patterns during the target period.

Our comparative test results show that the HKJ model is the most informative earthquake forecast during the 2011–2020 evaluation period. These results are consistent with the outcomes of the RELM experiment reported by Schorlemmer et al. (2010), Zechar et al. (2013) and Strader et al. (2017), but contrary to the findings obtained by Rhoades et al. (2014) for hybrid models in retrospective analyses. Discrepancies between prospective and retrospective test results may be primarily due to temporal instabilities in the fit to construct multiplicative hybrids. Thus, we provide evidence that smoothing high-resolution, small earthquake data remains a robust method for describing moderate-to-large earthquake activity over a period of more than 10 yr in California.

ACKNOWLEDGMENTS

The authors sincerely thank the editor Margarita Segou, reviewer Matteo Taroni and one anonymous reviewer for their constructive comments to improve this manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 821115, Real-time earthquake rIsk reduction for a reSilient Europe (RISE), http://www.rise-eu.org). Additionally, this research was supported by the Southern California Earthquake Center (contribution no. 11011). SCEC is funded by NSF Cooperative agreement EAR-1600087 and USGS Cooperative agreement G17AC00047.

8 DATA AVAILABILITY

The ComCat of the ANSS is available at https://earthquake.usgs.gov/earthquakes/search/ (last accessed January 2021). The earthquake forecast files can be found on Zenodo at https://doi.org/10.5281/zenodo.5141567, and the reproducibility software package for this manuscript can be found on GitHub at https://github.com/bayonato89/Reproducibility-hybrids.

REFERENCES

Armstrong, R.A., 2014. When to use the Bonferroni correction, *Ophthalmic Physiol. Opt.,* **34**(5), 502–508.

Barnhart, W.D., Hayes, G.P. & Gold, R.D., 2019. The July 2019 Ridgecrest, California, earthquake sequence: kinematics of slip and stressing in cross-fault ruptures, *Geophys. Res. Lett.,* **46**(21), 11859–11867.

Bayona, J.A., Savran, W., Strader, A., Hainzl, S., Cotton, F. & Schorlemmer, D., 2021. Two global ensemble seismicity models obtained from the combination of interseismic strain measurements and earthquake-catalogue information, *J. geophys. Int.,* **224**(3), 1945–1955.

Bayona Viveros, J.A., von Specht, S., Strader, A., Hainzl, S., Cotton, F. & Schorlemmer, D., 2019. A regionalized seismicity model for subduction zones based on geodetic strain rates, geomechanical parameters, and earthquake-catalog data, *Bull. seism. Soc. Am.,* **109**(5), 2036–2049.

Bird, P., 2018. Ranking some global forecasts with the Kagan information score, *Seismological Research Letters,* **89**(4), 1272–1276.

Bird, P., Jackson, D.D., Kagan, Y.Y., Kreemer, C. & Stein, R.S., 2015. GEAR1: a Global Earthquake Activity Rate Model constructed from geodetic strain rates and smoothed seismicity, *Bull. seism. Soc. Am.,* **105**(5), 2538–2554.

Bird, P. & Liu, Z., 2007. Seismic hazard inferred from tectonics: California, *Bull. seism. Soc. Am.,* **78**(1), 37–48.

Cass, K. & Tromans, C., 2008. A biometric investigation of ocular components in amblyopia, *Ophthalmic Physiol. Opt.,* **28**(5), 429–440.

Cheverud, J.M., 2001. A simple correction for multiple comparisons in interval mapping genome scans, *Heredity,* **87**(1), 52–58.

Ebel, J.E., Chambers, D.W., Kafka, A.L. & Baglivo, J.A., 2007. Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, *Ophthalmic Physiol. Opt.,* **78**(1), 57–65.

Field, E.H. *et al.*, 2009. Uniform California earthquake rupture forecast, version 2 (UCERF 2), *Bull. seism. Soc. Am.,* **99**(4), 2053–2107.

Field, E.H., 2007. Overview of the working group for the development of regional earthquake likelihood models (RELM), *Seismological Research Letters,* **78**(1), 7–16.

Freed, A.M. & Lin, J., 2001. Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer, *Nature,* **411**(6834), 180–183.

Gerstenberger, M.C. *et al.*, 2020. Probabilistic seismic hazard analysis at regional and national scales: state of the art and future challenges, *Rev. Geophys.,* **58**(2), e2019RG000653.

Gneiting, T. & Raftery, A.E., 2005. Weather forecasting with ensemble methods, *Science,* **310**(5746), 248–249.

Guy, M.R. *et al.*, 2015. *National Earthquake Information Center systems overview and integration,* Reston, VA: US Department of the Interior, US Geological Survey.

Hauksson, E., Jones, L.M., Hutton, K. & Eberhart-Phillip, D., 1993. The 1992 Landers earthquake sequence: seismological observations, *J. geophys. Res.: Solid Earth,* **98**(B11), 19835–19858.

Hauksson, E., Stock, J., Hutton, K., Yang, W., Vidal-Villegas, J.A. & Kanamori, H., 2011. The 2010 Mw 7.2 El Mayor-Cucapah earthquake sequence, Baja California, Mexico and Southernmost California, USA: active seismotectonics along the Mexican Pacific Margin, *Pure appl. Geophys.,* **168**(8), 1255–1277.

Helmstetter, A., Kagan, Y.Y. & Jackson, D.D., 2007. High-resolution time-independent grid-based forecast for M≥ 5 earthquakes in California, *Seismological Research Letters,* **78**(1), 78–86.

Holliday, J.R., Chen, C.C., Tiampo, K.F., Rundle, J.B., Turcotte, D.L. & Donnellan, A., 2007. A RELM earthquake forecast based on pattern informatics, *Seismological Research Letters,* **78**(1), 87–93.

Hurvich, C.M. & Tsai, C.L., 1989. Regression and time series model selection in small samples, *Biometrika,* **76**(2), 297–307.

Jackson, D.D. & Kagan, Y.Y., 1999. Testable earthquake forecasts for 1999, *Seismological Research Letters,* **70**(4), 393–403.

Jordan, T.H. *et al.*, 2011. Operational earthquake forecasting. State of knowledge and guidelines for utilization, *Ann. Geophys.,* **54**(4), doi:10.4401/ag-5350.

Kagan, Y.Y., 2009. Testing long-term earthquake forecasts: likelihood methods and error diagrams, *J. geophys. Int.,* **177**(2), 532–542.

Kagan, Y.Y., 2010. Statistical distributions of earthquake numbers: consequence of branching process, *J. geophys. Int.,* **180**(3), 1313–1328.

Kagan, Y.Y., Jackson, D.D. & Rong, Y., 2007. A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity, *Seismological Research Letters,* **78**(1), 94–98.

Kato, M., 2019. On the apparently inappropriate use of multiple hypothesis testing in earthquake prediction studies, *Seismological Research Letters,* **90**(3), 1330–1334.

Lombardi, A.M. & Marzocchi, W., 2010. The assumption of Poisson seismic-rate variability in CSEP/RELM experiments, *Bull. seism. Soc. Am.,* **100**(5A), 2293–2300.

Marti, M., Stauffacher, M. & Wiemer, S., 2019. Difficulties in explaining complex issues with maps: evaluating seismic hazard communication-the Swiss case, *Natural Hazards Earth Syst. Sci.,* **19**(12), 2677–2700.

Marzocchi, W., Lombardi, A.M. & Casarotti, E., 2014. The establishment of an operational earthquake forecasting system in Italy, *Seismological Research Letters,* **85**(5), 961–969.

Marzocchi, W., Zechar, J.D. & Jordan, T.H., 2012. Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. seism. Soc. Am.,* **102**(6), 2574–2584.

Meletti, C. *et al.*, 2021. The new Italian seismic hazard model (MPS19), *Ann. Geophys*.

Michael, A.J. & Werner, M.J., 2018. Preface to the focus section on the Collaboratory for the Study of Earthquake Predictability (CSEP): new results and future directions, *Seismological Research Letters,* **89**(4), 1226–1228.

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**(4), 1650–1659.

Perneger, T.V., 1998. What's wrong with Bonferroni adjustments, *Bmj,* **316**(7139), 1236–1238.

Pettit, N.N., MacKenzie, E.L., Ridgway, J.P., Pursell, K., Ash, D., Patel, B. & Pho, M.T., 2020. Obesity is associated with increased risk for mortality among hospitalized patients with COVID-119, *Obesity,* **28**(10), 1806–1810.

Rhoades, D. A. & Stirling, M. W., 2012. An earthquake likelihood model based on proximity to mapped faults and cataloged earthquakes, *Bull. seism. Soc. Am.,* **102**(4), 1593–1599.

Rhoades, D.A., 2013. Mixture models for improved earthquake forecasting with short-to-medium time horizons, *Bull. seism. Soc. Am.,* **103**(4), 2203–2215.

Rhoades, D.A., Christophersen, A., Gerstenberger, M.C., Liukis, M., Silva, F., Marzocchi, W., Werner, M.J. & Jordan, T.H., 2018. Highlights from the first ten years of the New Zealand earthquake forecast testing center, *Seismological Research Letters,* **89**(4), 1229–1237.

Rhoades, D.A. & Gerstenberger, M.C., 2009. Mixture models for improved short-term earthquake forecasting, *Bull. seism. Soc. Am.,* **99**(2A), 636–646.

Rhoades, D.A., Gerstenberger, M.C., Christophersen, A., Zechar, J.D., Schorlemmer, D., Werner, M.J. & Jordan, T.H., 2014. Regional earthquake likelihood models II: information gains of multiplicative hybrids, *Bull. seism. Soc. Am.,* **104**(6), 3072–3083.

Rhoades, D.A., Liukis, M., Christophersen, A. & Gerstenberger, M.C., 2016. Retrospective tests of hybrid operational earthquake forecasting models for Canterbury, *J. geophys. Int.,* **204**(1), 440–456.

Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D. & Imoto, M., 2011. Efficient testing of earthquake forecasting models, *Acta Geophys.,* **59**(4), 728–747.

Ross, Z.E. *et al.*, 2019. Hierarchical interlocked orthogonal faulting in the 2019 Ridgecrest earthquake sequence, *Science,* **366**(6463), 346–351.

Schorlemmer, D. *et al.*, 2018. The collaboratory for the study of earthquake predictability: achievements and priorities, *Seismological Research Letters,* **89**(4), 1305–1313.

Schorlemmer, D. & Gerstenberger, M.C., 2007. RELM testing center, *Seismological Research Letters,* **78**(1), 30–36.

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D. & Rhoades, D.A., 2007. Earthquake likelihood model testing, *Seismological Research Letters,* **78**(1), 17–29.

Schorlemmer, D., Zechar, J.D., Werner, M.J., Field, E.H., Jackson, D.D., Jordan, T.H. & RELM Working, Group, 2010. First results of the regional earthquake likelihood models experiment, in *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II,* pp. 1305–1313, Springer, Basel.

Shebalin, P.N., Narteau, C., Zechar, J.D. & Holschneider, M., 2014. Combining earthquake forecasts using differential probability gains, *Earth Planets Space,* **66**(1), 1–14.

Shen, Z.K., Jackson, D.D. & Kagan, Y.Y., 2007. Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M5 earthquakes in southern California, *Ophthalmic Physiol. Opt.,* **78**(1), 116–120.

Strader, A., Schneider, M. & Schorlemmer, D., 2017. Prospective and retrospective evaluation of five-year earthquake forecast models for California, *J. geophys. Int.,* **211**(1), 239–251.

Strader, A., Werner, M., Bayona, J., Maechling, P., Silva, F., Liukis, M. & Schorlemmer, D., 2018. Prospective evaluation of global earthquake forecast models: 2 yrs of observations provide preliminary support for merging smoothed seismicity with geodetic strain rates, *Seismological Research Letters,* **89**(4), 1262–1271.

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, *J. geophys. Int.,* **196**(1), 422–431.

Taroni, M., Zechar, J.D. & Marzocchi, W., 2014. Assessing annual global M 6+ seismicity forecasts, *Seismological Research Letters,* **89**(4), 1251–1261.

Vere-Jones, D., 1970. Stochastic models for earthquake occurrence, *J. R. Stat. Soc.: Ser. B (Methodological),* **32**(1), 1–45.

Vere-Jones, D., 1995. Forecasting earthquakes and earthquake risk, *Int. J. Forecast.,* **11**(4), 503–538.

Ward, S.N., 2007. Methods for evaluating earthquake potential and likelihood in and around California, *Seismological Research Letters,* **78**(1), 121–133.

Werner, M.J., Helmstetter, A., Jackson, D.D. & Kagan, Y.Y., 2011. High-resolution long-term and short-term earthquake forecasts for California, *Bull. seism. Soc. Am.,* **101**(4), 1630–1648.

Werner, M.J. & Sornette, D., 2008. Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments, *J. geophys. Res.: Solid Earth,* **113**(B8).

Werner, M.J., Zechar, J.D., Marzocchi, W. & Wiemer, S., 2010. Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts, *Ann. Geophys.,* **53**(3), 11–30.

Wiemer, S. & Schorlemmer, D., 2007. ALM: An asperity-based likelihood model for California, *Seismological Research Letters,* **78**(1), 134–140.

Zechar, J.D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P.J. & Jordan, T.H., 2010. The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurr. Comput.: Pract. Exper.,* **22**(2), 1836–1847.

Zechar, J.D., Schorlemmer, D., Werner, M.J., Gerstenberger, M.C., Rhoades, D.A. & Jordan, T.H., 2013. Regional earthquake likelihood models I: first-order results, *Bull. seism. Soc. Am.,* **103**(2A), 787–798.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Table S1. Poisson-based *N*-, *S*-, *cL*- and *T*-test statistics for RELM and multiplicative hybrid earthquake forecasting models for California. The *N*-test metrics δ_1 and δ_2 describe the probabilities of observing at least and at most the number of actual earthquakes, respectively. At a 0.5 significance level, the forecast underestimates the number of target events if $\delta_1 < 0.025$; the forecast overestimates observed seismicity if $\delta_2 > 0.025$ and the forecast is consistent with the observations if $\delta_1 > 0.025$ and $\delta_2 > 0.025$. The space ζ and space–magnitude statistics ξ provide the percentage of POLL scores, derived from simulated catalogues, that are lower than the observed log-score. At a 0.05 significance level, the forecasted spatial and space–magnitude distributions are consistent with observed seismicity if these statistics are greater than 0.05. Thus, we highlight in bold the forecasts that properly describe the observations. In addition, we show IGPEs and confidence intervals (in brackets) over HKJ, obtained by each model during the prospective evaluation period.

Table S2. Non-Poissonian *N*-, *S*-, *cL*- and *T*-test statistics for RELM and multiplicative hybrid earthquake forecasting models for California. The NBD *N*-test metrics $\delta_{1\text{NBD}}$ and $\delta_{2\text{NBD}}$ describe the probabilities of observing at least and at most the observed number earthquakes, respectively. Moreover, the space ζ and space–magnitude statistics ξ provide the percentage of BILL scores, derived from simulated catalogues, that are lower than the observed log-score. At a 0.05 significance level, the criteria for model rejection and consistency are the same as for δ_1, δ_2, ζ and ξ in Table 1. Accordingly, we highlight in bold the forecasts that are consistent with the observed data. In addition, we report IGPAs and confidence intervals (in brackets) over HKJ, obtained by each model during the target testing period.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.