

Select earthquake forecasting models demonstrate consistency with prospective decadal observations in California

Received: 20 January 2026

Accepted: 22 July 2026

Published online: 06 August 2026

Cite this article as: Bayona, J.A., Serafini, F., Silva, F. *et al.* Select earthquake forecasting models demonstrate consistency with prospective decadal observations in California. *Nat Commun* (2026). <https://doi.org/10.1038/s41467-026-76243-7>

José A. Bayona, Francesco Serafini, Fábio Silva, Pablo Iturrieta, William H. Savran, Marcus Herrmann, Warner Marzocchi, Philip J. Maechling & Maximilian J. Werner

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Select earthquake forecasting models demonstrate consistency with prospective decadal observations in California

José A. Bayona^{1, *}, Francesco Serafini^{1, +}, Fábio Silva², Pablo Iturrieta³, William H. Savran⁴, Marcus Herrmann⁵, Warner Marzocchi⁵, Philip J. Maechling², and Maximilian J. Werner¹

¹School of Earth Sciences, University of Bristol, United Kingdom.

²Statewide California Earthquake Center, United States of America.

³Helmholtz Centre for Geosciences (GFZ Potsdam), Germany.

⁴Nevada Seismological Laboratory, University of Nevada Reno, United States of America.

⁵University of Naples Federico II, Italy.

*Corresponding author: José A. Bayona (Email: jose.bayona@bristol.ac.uk)

+Now at: Global Earthquake Model (GEM) Foundation

ABSTRACT

Earthquake forecasting systems are now operationalized by agencies in several countries, providing situational awareness to at-risk communities. To ensure that forecasts follow the best science available, the underlying models require rigorous testing against prospective data and benchmarking. Between August 2007 and August 2018, 27 forecasting models, including the types used for operational earthquake forecasting, were prospectively operated as part of a multi-institutional project by the Collaboratory for the Study of Earthquake Predictability, generating 51,625 next-day $M_w \geq 3.95$ seismicity forecasts for California. Here, we comprehensively evaluate their predictive skills against 597 $M_w \geq 3.95$ earthquakes using community-vetted statistical methods. We show that approximately 95%, two-thirds, and 95% of the models produced forecasts consistent with the number, locations, and magnitudes of observed earthquakes, respectively. Individually, select versions of well-established, operational-type models demonstrated sustained consistency with prospective observations. We provide detailed recommendations, software tools, and benchmarking data to support the development of operational earthquake forecasting systems worldwide.

Introduction

Seismologists are often asked 'When will the next big one occur?', an answer to which requires providing precise information (and uncertainties) about the timing, location, and magnitude of future earthquakes¹. Numerous peer-reviewed studies suggest that diagnostic seismic precursor signals could be routinely observable worldwide and that, consequently, large earthquakes might be predictable². However, these studies rely heavily on retrospective observations, meaning their conclusions can only be considered robust enough after independent prospective evaluations³. Thus, the lack of successful deterministic predictive models of earthquake occurrence has led seismologists, including members of the Collaboratory for the Study of Earthquake

Predictability (CSEP)⁴, to devote scientific efforts to forecasting earthquakes probabilistically^{5,6}.

CSEP is a global community of scientists dedicated to advancing earthquake predictability research through the prospective evaluation of seismicity forecasts in fully transparent and reproducible experiments, in which strict definitions of the modeling parameters and evaluation datasets are agreed upon in advance to avoid common biases. Based on these criteria, models providing probabilistic estimates of seismicity in temporal, spatial, and magnitude domains are evaluated against future, yet-to-be-collected data of earthquakes⁴. One of CSEP's major community accomplishments is the development and automation of 27 time-varying, spatially inhomogeneous seismicity models, which between August 1, 2007, and August 30, 2018, generated a fully prospective database of 51,625 next-day $M_w \geq 3.95$ earthquake forecasts for California⁷. Each of these forecasts was issued at midnight (UTC), covering a 24-hour time horizon and using earthquakes observed before each start time as input (see Fig. 1, Supplementary Figure 2, Supplementary Data / Tables 1 and 2, and the companion data publication⁷ to this article for more details).

This extensive set of increasingly complex earthquake forecasting models was gradually put into operation (see Supplementary Figure 1) and included: i) different flavors of the well-established Epidemic-Type Aftershock Sequence (ETAS)^{5,8} and Short-Term Earthquake Probability (STEP)⁹ models, which combine empirical statistical descriptions of past clustered seismicity, including the Omori–Utsu time-decay law¹⁰ and the Gutenberg–Richter frequency-magnitude relation¹¹, to forecast background and triggered seismicity; ii) versions of "alternative" models to ETAS and STEP, such as the Proximity to Past Earthquakes (PPE)¹² and Every Earthquake a Precursor According to Scale (EEPAS)¹³ models (e.g., PPE-BACKGROUND-M2 and DR-EEPAS, respectively); iii) nonparametric models or models that do not require a predetermined relationship between their parameters (e.g., GSF-ANISOTROPIC¹⁴ and HW-KERNEL-M2¹⁵) and; iv) ensemble models or combinations of individual models (e.g., ETAS-HW-KERNEL-M2 and ONEDAY-BAYESIAN).

Currently, ETAS and STEP are the main types of clustering models used by agencies such as the National Institute of Geophysics and Volcanology (INGV) in Italy, Earth Sciences New Zealand, and the United States Geological Survey (USGS) for Operational Earthquake Forecasting (OEF) systems, which collect and disseminate authoritative information on the time dependence of seismic hazards in near-real time¹⁶. OEF models often forecast over different time windows (e.g., weeks, months, years), which can be useful for various users but makes their evaluation and direct comparison with other models difficult¹⁷. Therefore, a comprehensive evaluation of this curated, prospective database of next-day earthquake forecasts is of paramount importance for addressing both practical and fundamental scientific questions, such as: How well can the underlying models describe the spatiotemporal evolution of earthquake sequences?; Are they capable of adequately forecasting background and triggered seismicity over extended periods?; What level of predictability should OEF users expect from these models?

During the 11 years in which the models were operating prospectively, the Comprehensive Earthquake Catalog (ComCat) of the Advanced National Seismic System (ANSS)¹⁸ reported 597 $M_w \geq 3.95$ earthquakes with hypocentral depths ≤ 30 km in the CSEP-California test region¹⁹, including the 2010 M_w 7.2 El Mayor–Cucapah earthquake sequence²⁰, the 2012 $M_w \geq 3.95$ Brawley seismic swarm²¹, and the 2014 M_w 6.0 South Napa earthquake²² (see Fig. 2). In this work, we prospectively evaluate, daily and cumulatively, the ability of these models to forecast the number, epicentral locations, and magnitude distribution of observed earthquakes, using statistical tests developed by CSEP and reviewed by the broader scientific community^{19,23} (see the Methods section). Furthermore, we quantitatively compare their long-term performance with that of a time-invariant smoothed seismicity model, hereafter referred to as HKJ-ADAPTIVE-M2²⁴, which over 15 years has outperformed other time-invariant

models that forecast moderate-to-large earthquakes in California²⁵ and can therefore help quantify how much predictive skill is gained by time-varying models. We focus on a model group of 21 $M_w \geq 3.95$ seismicity models because abundant data are available for testing and because OEF systems forecast earthquakes at this magnitude threshold. Nonetheless, prospective test results for a subset of 6 $M_w \geq 4.95$ earthquake forecasting models can also be found in the supplementary material for this article.

We show that, overall, approximately 95%, two-thirds, and 95% of the models produced next-day forecasts consistent with the number, locations, and magnitudes of observed earthquakes, respectively. At the individual model level, select versions of ETAS and STEP demonstrated statistical agreement with prospective observations, supporting the use of this type of triggering models in public OEF.

Results

How good were the models at forecasting the cumulative number of observed earthquakes?

Quantitative assessment of the degree of (mis)match between the number of expected and observed earthquakes over extended periods can help elucidate whether the background and triggered seismicity components of models are correctly calibrated. Hence, we assess number (in)consistencies between forecasts and observations during the different periods in which the underlying models were operational. Prospective results of the CSEP number test²⁶ show that only ETAS-PPE-M2.95 and STEPJAVA forecasts were statistically consistent with the observed data (see the proximity between the circles [observations] and the black solid bars [models' predictive intervals] in Fig. 3c). This (in)consistency test, however, relies heavily on the assumption that the number of earthquakes can be modeled using a Poisson distribution, which depends on a single parameter and is not particularly skewed, resulting in predictive intervals that do not fully capture the spatiotemporal variability of earthquake rates due to their strong clustering nature²⁷.

Acknowledging this limitation, we also perform the CSEP Negative Binomial Distribution number test, whose underlying probability distribution exhibits a larger variance than the Poisson distribution and can therefore be used to better account for earthquake-rate variability²⁸. In this case, we show that, with the exception of KJ-SMOOTHING-V0, all models statistically matched the observations (see how the circles fall within the models' 95% confidence intervals, shown as gray dashed bars, in Fig. 3c). These results represent a milestone achievement in earthquake predictability research, considering in particular that i) the models' parameters that account for earthquake productivity were not updated throughout the evaluation period and that ii) all next-day forecasts were automatically generated on CSEP servers over extended periods without human intervention.

For practitioners working with (symmetric) percentage differences between forecasts and observations, we further show that 9 (almost half) of the 21 time-varying $M_w \geq 3.95$ seismicity models provided earthquake counts that differed by 20% or less from the total number of observed earthquakes. Among these models are versions of STEP and ETAS, such as STEPJAVA, ETAS-SMOOTHING-V0, and STEP, which over periods of approximately 8, 11, and 6 years forecasted nearly 254, 678, and 398 $M_w \geq 3.95$ earthquakes, respectively, differing by 3%, 17% and 17% from the 246, 579, and 332 events observed during such periods, respectively (Fig. 3c). We remind the reader that the models were confronted against different numbers of earthquakes ω , because each one has unique operationalization start and end dates (Supplementary Figure 1). In this sense, (in)consistency test results based on more prospective data are more indicative of their actual predictive skill.

How good were the models at forecasting the daily number of events observed during specific earthquake scenarios?

Evaluating the models' ability to forecast earthquakes on a daily basis provides valuable information on how closely they can track the temporal evolution of earthquake sequences, seismic swarms, and background seismicity, and is particularly useful after large events, as triggered seismicity can negatively impact emergency protocols and inflict further damage to already weakened infrastructure²⁹. We therefore analyze (in)consistencies between observed and expected daily counts during a) the M_w 6.5 Ferndale earthquake sequence, whose largest event struck off the coast of Northern California on January 10, 2010³⁰; b) the M_w 7.2 El Mayor–Cucapah earthquake on April 4, 2010, which is the largest earthquake to hit California since the 1992 M_w 7.3 Landers earthquake²⁰; c) the Brawley seismic swarm of August 26, 2012, which has been claimed to be induced by fluid pumping into a nearby geothermal field²¹; d) the M_w 6.8 earthquake that occurred in the Mendocino area of Northern California on March 10, 2014; e) the M_w 6.0 South Napa earthquake of August 24, 2014, which, contrary to expectations based on the Bath's law³¹, did not trigger any $M_w \geq 3.95$ earthquakes^{22,32}; and f) a randomly-selected period of relative seismic calm in California, spanning from January 1 to 20, 2018.

All forecasts statistically underestimated the number of $M_w \geq 3.95$ earthquakes observed during the days when the 2010 M_w 6.5 Ferndale, 2010 M_w 7.2 El Mayor–Cucapah, 2012 $M_w \geq 3.95$ Brawley, and 2014 M_w 6.8 Mendocino earthquakes occurred (see the red circles in Fig. 4a-d). These number inconsistencies reflect the major limitations of these types of seismicity models in that their forecasts of large earthquakes (and their imminent aftershocks) will generally be of low probability during those specific days, unless the seismicity is already considerably high. In the particular case of the El Mayor–Cucapah earthquake (Fig. 4b), we note that such underestimations persisted during the first 2 days of the sequence (considering the day of the main event as day 0), until ETAS-SMOOTHING-V0 and STEP were able to adequately capture the observed daily count. Conversely, during the two weeks following the 2010 M_w 6.5 Ferndale earthquake, the 2012 Brawley seismic swarm, and the two 2014 M_w 6.0 and M_w 6.8 Northern California earthquakes, we find that over 80% of the models were statistically consistent with daily observations (see Figs. 4a,c,d-e).

The results for the 2012 Brawley seismic swarm (Fig. 4c) are particularly interesting because the models were not explicitly designed to forecast swarm-type seismicity, which is characterized by the lack of a single dominant earthquake and a temporal evolution that cannot be described by any simple law comparable to Omori's law³³. This swarm lasted one month, but produced the 9 largest ($M_w \geq 3.95$) earthquakes on the first day alone, including two $M_w \geq 5.3$ strike-slip events and two $M_w \geq 4.3$ normal-faulting earthquakes²¹. The models, treating this swarm as a regular earthquake sequence, anticipated that the daily count would decay from a maximum of 5 to a maximum of 1 $M_w \geq 3.95$ earthquakes during the first 15 days of the sequence, which brought them into statistical agreement with observations. Finally, during the two randomly-selected weeks of relative seismic calm in California, over 80% of the available models adequately forecasted the daily background rate, providing predictive intervals that included a number of $M_w \geq 3.95$ earthquakes per day between zero and one (Fig. 4f).

How good were the models at forecasting the spatial distribution of observed earthquakes?

Properly forecasting where earthquakes are most likely to occur can help vulnerable communities implement more targeted mitigation plans, as earthquakes, especially after major events, cluster not only temporally but also spatially. Based on this premise, we perform CSEP binary spatial tests²⁵ to assess, for each model and for each of the days between August 2007 and August 2018 on which $M_w \geq 3.95$ earthquakes were recorded in California, the log-likelihoods (referred to here as jbill

values) of observing either no seismicity or at least one target earthquake in each of the forecasts' spatial grid cells (see the Methods section). This test is summarized by the observed spatial quantile ζ , which is the fraction of simulated catalogs of target activated cells with spatial log-likelihoods smaller than the j bill observed on a seismically active day. If $\zeta > 0.05$, then the spatial distributions of observed and expected earthquakes can be considered statistically indistinguishable.

By sorting for each model all the observed spatial quantiles ζ in ascending order (pink curves in Fig. 5) and comparing these distributions with spatial quantiles drawn from uniform distributions (gray dashed diagonals), we show that 15 ($\sim 71\%$) of the 21 time-varying $M_w \geq 3.95$ seismicity models generated next-day earthquake forecasts statistically consistent with the spatial distribution of observed epicenters (see the pink areas [A] between the curves and the diagonals; the smaller the area, the greater the agreement between forecasts and observations). Among the most consistent models are different flavors of ETAS and STEP, including ETAS-ADAPTIVE-M3 ($A = 0.03$), ETAS-PPE-M2 ($A = 0.04$) and STEPJAVA ($A = 0.04$), the HW-KERNEL-M2 and HW-KERNEL-M3 nonparametric models ($A = 0.04$ in both cases), and the ensemble models resulting from combining HW-KERNEL-M3 with the ETAS-ADAPTIVE-M* sibling models, namely ETAS-HW-KERNEL-M2 ($A = 0.06$) and ETAS-ETAS-HW-KERNEL-M3 ($A = 0.04$).

The shape and position of the pink curves relative to the gray diagonals in Fig. 5 provide additional information on the level of spatial precision that the models can offer. In particular, if the pink curves lie below the diagonals, this indicates that models generated spatially overly smoothed forecasts compared to observations, whereas, if the pink curves lie above the diagonals, this denotes that models provided spatial forecasts that are too localized or inaccurate (see the Methods section). To illustrate this further, and to specifically evaluate spatial forecasts during major periods of activity, Fig. 6 shows $M_w \geq 3.95$ seismicity forecasts generated by models available during the first week of the 2010 M_w 6.5 Ferndale, 2010 M_w 7.2 El Mayor–Cucapah, 2012 $M_w \geq 3.95$ Brawley, and 2014 M_w 6.8 Mendocino earthquake sequences.

During these sequences, ETAS-SMOOTHING-V0 obtained relatively high spatial quantiles ζ of 0.91, 0.99, 0.67, and 0.81, respectively, mainly because the model produced overly smoothed spatial forecasts, assigning similar earthquake probabilities to wide forecast regions, while the observed seismicity was spatially more concentrated. Conversely, during the Ferndale, El Mayor–Cucapah, and Brawley sequences, STEP obtained relatively low spatial quantiles (0.05, 0.00, and 0.16, respectively), primarily because the model exhibited spatial overconfidence in forecasting that earthquakes would occur only in a few yellow spatial cells, assigning much lower probabilities elsewhere (note that the colorbar scale in Fig. 6 is logarithmic). In contrast, during the 2014 Mendocino sequence, models such as ETAS-ADAPTIVE-M2, HW-KERNEL-M2 and ETAS-HW-KERNEL-M2 obtained spatial quantiles of 0.38, 0.80, and 0.53, respectively, presumably because they adopted nonparametric components and adaptive smoothing kernels whose bandwidths vary with the level of observed seismicity, thus achieving a comparatively more effective balance between spatial precision and accuracy.

How good were the models at forecasting the magnitudes of observed earthquakes?

This question is relevant from a societal and scientific point of view due to the growing interest in expected ground shaking following a large earthquake. To answer it, we quantitatively compare the incremental frequency-magnitude distribution of observed $M_w \geq 3.95$ earthquakes with that predicted by models during the periods in which they were available. Prospective results of the CSEP magnitude test²⁶ indicate that all models provided magnitude distributions statistically consistent with observations (see the $\phi > 0.05$ values, which are the p -values of the test, in Fig. 7). This magnitude test, however, is limited in its ability to identify whether two data samples come from the same distribution, so we complement this analysis with

nonparametric Kolmogorov–Smirnov tests (see the Methods section). We find that, with the exception of STEPJAVA, all time-varying models provided magnitude forecasts statistically consistent with observations, obtaining K_m values (p -values) greater than 0.05. Among the most consistent models are versions of ETAS, such as ETAS-SMOOTHING-V1 [$K_m = 0.99$], ETAS-ADAPTIVE-M3 [$K_m = 0.92$], and ETAS-PPE-M2 [$K_s = 0.89$], the ETAS-HW-KERNEL-M* ensemble models [$K_m = 0.80$ and $K_m = 0.69$, respectively], and the HW-KERNEL-M* nonparametric models [$K_m = 0.66$ and $K_m = 0.45$, respectively].

How did each next-day model compare to the benchmark model?

In addition to quantifying the level of (dis)agreement between forecasts and observations, we compare the predictive skill of their underlying models with that of the time-invariant HKJ-ADAPTIVE-M2 model, using a measure from information theory called cumulative binary information gain per activated spatial cell (IGPA²⁵; see the Methods section). This binary metric reduces the sensitivity of the results to the underlying test assumption that earthquakes in individual forecast cells are independent to each other and that their expected rates follow a Poisson distribution²⁶.

Because the forecast database was implemented gradually over time (Supplementary Figure 1), we perform comparative analyses across six non-overlapping periods, each containing a different subset of available models. This staggered availability could influence comparisons between models, particularly with older ones, meaning that comparisons based on more prospective data better reflect the relative performance of each model. Acknowledging this limitation, we report that, between August 2007 and January 2009, when only HKJ-ADAPTIVE-M2, ETAS-SMOOTHING-V0, and STEP were available, STEP statistically outperformed the two other existing models, obtaining an IGPA of approximately 1.25 over HKJ-ADAPTIVE-M2 (see Fig. 8a). Subsequently, between January 2009 and September 2010, when the 2010 M_w 6.5 Ferndale and M_w 7.2 El Mayor–Cucapah earthquakes occurred and the KJ-SMOOTHING-V0 model came into operation, STEP and ETAS-SMOOTHING-V0 were the most informative, as both obtained IGPAs of nearly 5.85 over the HKJ-ADAPTIVE-M2 model (Fig. 8b). Note that, while STEP was more skilled than ETAS-SMOOTHING-V0 at forecasting the January 2010 M_w 6.5 Ferndale earthquake and its triggered events, ETAS-SMOOTHING-V0 slightly outperformed STEP during the El Mayor–Cucapah sequence in April of that year.

Later, between September 2010 and July 2012, STEPJAVA, which is an updated version of STEP written in Java, defined on a region smaller than the entire CSEP-California test region, and which had just been included in the pool of predictive models, ranked first by obtaining IGPAs of nearly 2.12 over the ETAS-SMOOTHING-V0 model (see Fig. 8c). Conversely, KJ-SMOOTHING-V0 was the least informative model, obtaining a cumulative IGPA (or indeed loss) of about -0.66 over HKJ-ADAPTIVE-M2. Moving forward, between July and October 2012 (Fig. 8d), when the Brawley seismic swarm struck Southern California and the ETAS-SMOOTHING-V1 and ONEDAY(-SEQ)-BAYESIAN models joined the forecast experiment, STEP and ETAS-SMOOTHING-V1 obtained the largest IGPAs (approximately 0.81 and 0.58, respectively) over HKJ-ADAPTIVE-M2, while KJ-SMOOTHING-V0, in contrast, obtained the lowest gain (approximately -3.96). The IGPAs obtained by STEP and ETAS-SMOOTHING-V1 over the benchmark model are, however, statistically insignificant (see Supplementary Figure 9d), indicating that HKJ-ADAPTIVE-M2 can be considered as informative as the time-varying models during this period.

Between January 2013 and July 2016 (Fig. 8e), when the 2014 M_w 6.8 Mendocino and M_w 6.0 South Napa earthquakes occurred, STEP was discontinued, and 10 new models were submitted to CSEP for prospective evaluation, the ETAS-ADAPTIVE-M* and HW-KERNEL-M* models, as well as their ensembles, ETAS-HW-KERNEL-M2 and ETAS-HW-KERNEL-M3, were the most informative models, with the latter two obtaining IGPAs of about 4.16 and 4.05 over HKJ-ADAPTIVE-M2, respectively. The statistically significant outperformance of ETAS-SMOOTHING-V1 over ETAS-SMOOTHING-V0 during this 3.5-year period

(Supplementary Figure 9e) is particularly interesting, because these models only differ in their spatial smoothing parameters, demonstrating that the comparatively finer smoothing procedure implemented in ETAS-SMOOTHING-V1 resulted in better earthquake forecasting. During this period, we also note a clear distinction in models' performance following the M_w 6.8 Mendocino earthquake on March 10, 2014, with some of the newer models obtaining the largest IGPA's and some of the older ones tending to be the least informative (e.g., KJ-SMOOTHING-V0, ETAS-SMOOTHING-V0, and ONEDAY-BAYESIAN).

Finally, between October 2016 and August 2018 (Fig. 8f), when four additional models underwent prospective testing (albeit another five were discontinued, as seen in Supplementary Figure 1), we find that two of them, ETAS-PPE-M2.95 and GSF-ANISOTROPIC, along with ETAS-PPE-M2, were the most informative, obtaining IGPA's of about 1.67, 1.90, and 1.91 over HKJ-ADAPTIVE-M2, respectively. The KJ-SMOOTHING-V0 model was once again the least informative, obtaining an IGPA of approximately -0.42 over HKJ-ADAPTIVE-M2. Despite the limited number of forecast cells activated during this period [33], there were a total of 15 different $M_w \geq 3.95$ earthquake forecasting models available, making comparisons between them meaningful, particularly around their relative abilities to forecast background seismicity. As in the case of the 2013–2016 period (Fig. 8e), we find that the most recent models obtained the largest IGPA's, while older ones tended to do the opposite. Furthermore, we observe that, during the e–f periods, i.e., between January 2013 and August 2018, model versions calibrated on small earthquakes (e.g., ETAS-PPE-M2) obtained slightly larger IGPA's than their analogous models based on larger events (e.g., ETAS-PPE-M3; see Supplementary Figure 12), suggesting that small(er) earthquakes may add predictive skill²⁴.

Discussion

Lessons learned for the next generation of earthquake forecasting models

Prospective evaluations reveal that the 21 time-varying $M_w \geq 3.95$ seismicity models, as a whole, were most effective at forecasting the number and magnitude distributions of observed earthquakes, followed by their epicentral locations. Individually, select versions of ETAS and STEP, such as ETAS-PPE-M*, ETAS-HW-KERNEL-M*, and STEPJAVA, were among the most consistent with prospective observations. Regarding the number components, we show that, based on the CSEP Negative Binomial Distribution number test²⁸, 20 models statistically captured the cumulative number of observed earthquakes (see Fig. 3c). Some of the most consistent models, such as STEP(JAVA) and ETAS-SMOOTHING-V0, were calibrated on historical $M_w \geq 3.95$ earthquake catalogs dating back to 1932 and 1898, respectively, while other models, such as HW-KERNEL-M*, ETAS-HW-KERNEL-M*, and ETAS-ADAPTIVE-M*, used a more recent instrumental catalog containing $M_w \geq 2.0$ earthquakes (see Supplementary Table 2). In the particular cases of ETAS-PPE-M2(.95) and ETAS-PPE-M3, the models also used a time-varying (PPE) background component¹³ (see Supplementary Table 1). Taken together, these results suggest that integrating historical earthquake records with higher-resolution seismicity catalogs and assuming time-varying background components are key elements to consider when developing seismicity models consistent with the total number of observed earthquakes in California.

Over shorter timescales, we show that more than 80% of the available models properly forecasted the daily count of earthquakes observed during the 15 days following the 2010 M_w 6.5 Ferndale, 2014 M_w 6.8 Mendocino, and 2014 M_w 6.0 South Napa sequences (see Fig. 4a, d–e). However, during the first two days of the 2010 El Mayor–Cucapah earthquake sequence, ETAS-SMOOTHING-V0, STEP, and KJ-SMOOTHING-V0 statistically underestimated the number of observations (Fig. 4b). This could be because the models were calibrated with spatiotemporally uniform completeness magnitudes (CM column

in Supplementary Table 2), which, after major earthquakes, are known to increase dramatically and bias the expected productivity^{34,35}. More importantly, these results could be due to the fact that most models used mean aftershock productivity parameters (AP in Supplementary Table 2), calibrated on past earthquakes observed throughout the CSEP-California test region, which may not entirely represent the diversity of productivity patterns observed more locally across the region^{32,36}. Nevertheless, a counterargument is that STEP, despite including a model component that estimates sequence-specific earthquake productivity as time progresses and data accumulate⁹, also had difficulty forecasting triggered seismicity. Hence, we conclude that the sequence was indeed more productive than expected (see Supplementary Figure 4) and, therefore, suggest that future improvements to the number components of these models could be achieved by: i) considering spatially variable productivities³⁶; ii) accounting for short-term catalog incompleteness issues³⁴; iii) updating daily forecasts as soon as moderate-to-large earthquakes occur³⁷; and/or iv) updating productivity parameters more regularly. Furthermore, we recommend investigating the use of productivity parameters that depend on measurable properties, such as tectonic setting, Gutenberg–Richter b -values, hypocentral depths, seismogenic crustal thickness, afterslip, and heat flow, as suggested by different studies^{38–42}. Clearly, models incorporating these covariates will require prospective testing.

In the particular case of KJ-SMOOTHING-V0 (and its sibling model KJ-SMOOTHING-V1), we also advocate for technical reviews of the codes that generate their seismicity forecasts, as both models presented notable difficulties in statistically capturing the number of observed earthquakes. While KJ-SMOOTHING-V0 significantly underestimated the 510 $M_w \geq 3.95$ earthquakes observed over a period of almost 10 years (Fig. 3), KJ-SMOOTHING-V1 significantly overestimated the number of observations, providing a long-term rate of approximately $1.6E-3 M_w \geq 4.95$ earthquakes per day per km^2 (estimated from forecast files) that differed substantially from the rate of approximately $3.5E-7 M_w \geq 4.95$ earthquakes per day per km^2 reported for the model⁴³ (see Supplementary Figure 3). Revised versions of these models would undoubtedly honor the unparalleled contributions, legacy, and impact of Prof. Yan Kagan and Prof. Dave Jackson (KJ in KJ-SMOOTHING-V*) within and beyond the field of seismology⁴⁴.

In the context of background rates, we encourage further research (and prospective testing) on the optimal combination/use of PPE-BACKGROUND-M2, the time-dependent background component of ETAS-PPE-M2, as the model overestimated the long-term rate of observed seismicity by $\sim 44\%$ (Fig. 3c). This could be because PPE-BACKGROUND-M2 was calibrated on earthquakes recorded in periods of comparatively high seismic activity in California, including the 1989 M_w 6.9 Loma Prieta, 1992 M_w 6.1, 7.3, 6.2 Joshua Tree-Landers-Big Bear, and 1999 M_w 7.1 Hector Mine earthquakes¹⁸ (see Supplementary Table 2, CP column). We also highlight the remarkable performance of HKJ-ADAPTIVE-M2 during the 2012 Brawley seismic swarm, as this model, like the time-varying models, was not explicitly designed to account for variations in background rates that arise from features not contained in seismicity catalogs, such as aseismic processes⁴⁵. This swarm lasted a total of one month, although its peak activity occurred on August 22, producing 9 $M_w \geq 3.95$ shocks²¹. In response, over 80% of the models adequately forecasted the (lack of) observed earthquakes during the 15 days following these events. In fact, during the period between July and October 2012, HKJ-ADAPTIVE-M2 was found to be statistically as informative as most of the next-day models (Figs. 8d, and S9d), providing support for the use of these models, as a first approximation, to describe the temporal evolution of swarm-type seismicity. Undoubtedly, adding more specific swarm-type model components could improve OEF^{45,46}.

Moving into the spatial domain of the models, we find that 15 of the 21 evaluated models generated earthquake forecasts consistent with the spatial distribution of observed epicenters (see Fig. 5). This is a remarkable achievement considering that

none of the different spatial parameters/ methods was updated throughout the evaluation period. Broadly speaking, models that incorporated nonparametric components and "intermediate" smoothing procedures, such as adaptive smoothing, produced the most consistent spatial forecasts. E.g., the HW-KERNEL-M* models¹⁵, which employ a power-law kernel with variable bandwidth, along with the ETAS-ADAPTIVE-M* models, which assume a power-law kernel for $M < 5.5$ triggering earthquakes and an anisotropic one for $M \geq 5.5$ events, were among the most consistent with observations. In contrast, models such as ETAS-SMOOTHING-V0 and STEP, which predominantly use fixed isotropic smoothing functions, exhibited the poorest spatial performances, producing oversmoothed and overlocalized seismicity forecasts, respectively. The spatial limitations of ETAS-SMOOTHING-V0, nonetheless, appear to have been addressed by its successor model, ETAS-SMOOTHING-V1, which uses a comparatively smaller kernel bandwidth (see Fig. 1), resulting in improved spatial forecasting (see Figs. 8e and S9e).

The use of anisotropic smoothing kernels that assign higher earthquake probabilities to regions along the strike angle of main faults is another ingredient that appears to produce spatially consistent seismicity forecasts. The ETAS-ADAPTIVE-M* models mimic this by smoothing the locations of early aftershocks (observed within 3 days) to then forecast subsequent events. Similarly, the STEP(JAVA) models are formulated to switch to a spatially anisotropic kernel once sufficient triggered seismicity is observed⁹. However, these model components require more data than were available even after the 2010 M_w 7.2 El Mayor–Cucapah earthquake, which is the largest event in the evaluation database (see the largely isotropic earthquake forecasts generated by STEP in Fig. 6). Thus, we recommend investigating new methods to improve the spatial triggering kernel in STEP, which, in addition to early triggered seismicity, could also rely on early USGS products such as ShakeMaps⁴⁷ and finite source models⁴⁸, geodetic information⁴⁹ and/or preexisting fault maps⁴¹.

Regarding the magnitude component of the models, we show that all of them provided earthquake frequency-magnitude distributions consistent with observations, based on the CSEP magnitude test²⁶ (see Fig. 7). Similarly, by performing nonparametric Kolmogorov–Smirnov tests, we demonstrate that, with the exception of STEPJAVA, all models provided magnitude distributions statistically indistinguishable from the magnitude distribution of observed earthquakes. The results for STEPJAVA are most likely due to its predicted distribution deviating from a typical Gutenberg–Richter distribution towards higher values within the 6.0 – 7.0 magnitude range (similar to a characteristic-type distribution) and the fact that no $M_w \geq 6.0$ earthquakes occurred in the STEPJAVA region during the time the model was operational. In general, these test results suggest using tapered Gutenberg–Richter distributions with mean b -values = 1.0 and corner magnitudes = 8.0 to forecast earthquake magnitudes in California (see Supplementary Table 1 and the references therein).

In terms of the amount of information that each competing model can provide relative to the reference model, we note that model pairs calibrated on small(er) earthquakes outperformed their sibling models based on larger events (see, e.g., the performance of ETAS-ADAPTIVE-M2 [IGPA = 0.30] over ETAS-ADAPTIVE-M3 in Fig S12). Although the improvements are modest, these findings are encouraging, considering that the next generation of high-resolution seismicity catalogs, enhanced with machine learning, template matching, and other techniques, are expected to improve probabilistic earthquake forecasting, as they contain many more small earthquakes than current instrumental catalogs⁵⁰.

The fact that some of the most recent next-day models obtained significantly larger information gains over HKJ-ADAPTIVE-M2 than older models indicates that seismologists are already producing increasingly informative earthquake forecasts. For instance, while STEP(JAVA) were the most informative models between August 2007 and October 2012 (Fig. 8a–d), more recent flavors of ETAS, such as ETAS-ADAPTIVE-M* and ETAS-PPE-M*, along with newer nonparametric models, such as

HW-KERNEL-M* and GSF-ANISOTROPIC, took over between January 2013 and August 2018 (see Figs. 8e–f and S9). This provides compelling evidence that models, whether using higher resolution data, better spatial smoothing techniques, or more sophisticated nonparametric forecasting methods, are improving over time. In terms of data acquisition, newer models could have benefited from comparatively better constraints on earthquake locations and magnitude estimates, given the modernization and expansion of the ANSS seismic network since the early 2010s⁵¹.

Another promising indication of increasing predictive skill over time is that the newer ETAS-HW-KERNEL-M* ensembles were among the most informative models during the 2012–2018 evaluation period, outperforming, although not always statistically significantly, their parent forecast models (see Figs. 8d–f and S9). Looking ahead, we recommend combining next-day California models by assigning weights that maximize the skill of an ensemble as a whole rather than weighting models based on their individual skills⁵², as ensembles such as ONEDAY-(SEQ-)BAYESIAN, which use the latter Bayesian technique, were unable to outperform their individual model components (compare them against ETAS-SMOOTHING-V1 in Fig. 8d–e).

Indeed, time-invariant ensemble models have been shown to underperform their parent models in fully prospective evaluations in California²⁵. However, on a global scale, contrasting prospective test results have also been reported, with ensemble models such as the Global Earthquake Activity Rate (GEAR1)⁵³ model outperforming its two individual model components based separately on smoothed seismicity and interseismic strain data^{54,55}. In this context, we also recommend combining (and prospectively testing) next-day models for California that use a broader range of geophysical datasets, including fault-kinematic data, geodetic information, and historical/paleoseismological earthquake records⁵⁶. Furthermore, we encourage prospective comparisons between statistical models and models that incorporate alternative (physical) descriptions of earthquake nucleation, including those based on elasto-static stress transfer estimates and rate-and-state friction laws⁵⁷, multicycle earthquake simulations⁵⁸, and neural point processes⁵⁹.

Implications for operational earthquake forecasting

When discussing the implications of this model evaluation for OEF, it is worth reminding the reader that ETAS and STEP are the main types of clustering models used for this purpose in countries such as Italy, New Zealand, and the United States. The OEF system at INGV in Italy issues probabilistic aftershock forecasts based on an ensemble model that combines, in a Bayesian fashion, local versions of ETAS, STEP, and the Epidemic-Type Earthquake Sequence (ETES) model¹⁷. To our knowledge, this is the only OEF system in the world that, since 2013, has been producing probabilistic local magnitude $M_L \geq 4.0$ earthquake forecasts every day at midnight for the week following and immediately after the occurrence of $M_L \geq 3.5$ earthquakes³⁷. Prospective evaluations of daily and weekly forecasts generated by these models over the past decade have shown good overall agreement between observed and forecasted seismicity, with the exceptions of the 2012 M_w 6.1 Emilia and 2016–2017 Amatrice-Norcia M_w 6.6 earthquake sequences.

During the Emilia sequence, STEP had difficulty forecasting the spatial distribution of observed earthquakes, mainly because the model, as also noted in this work, provided spatial forecasts that were too localized compared to observations⁶⁰. In the case of the Amatrice–Norcia sequence, the OEF-Italy models underestimated the number of observed earthquakes, presumably due to short-term catalog incompleteness issues arising during the first days after the largest events³⁵. Nonetheless, in practice, this aspect was immediately mitigated by applying an empirical correction to the forecasts produced by these models⁶¹. Thus, based on these number and spatial test results and those presented here for California, we argue that future improvements to the OEF models in Italy could be achieved by: i) developing and prospectively testing new versions of STEP that use larger smoothing

distances and/or assign greater weights to their anisotropic triggering kernel components; ii) developing and validating new flavors of ETAS that assume adaptive spatial kernels and time-dependent background rates; and iii) updating model parameters more regularly, as these were last updated in April 2009³⁷.

The OEF system at Earth Sciences New Zealand has been making time-varying seismicity forecasts available to the public since 2010 following major events, such as the 2010 M_w 7.1 Darfield, 2011 M_w 6.2 Christchurch, and 2016 M_w 7.8 Kaikōura earthquakes⁶². The underlying OEF model is an ensemble composed of: i) short-term ETAS and STEP models (ranging from hours to weeks); ii) a medium-term EEPAS¹³ model (ranging from months to years) and; iii) long-term models like PPE¹² (spanning decades). Prospective evaluations of the forecasts generated by these models show that, during the 6 months following the 2010 Darfield earthquake, when the devastating M_w 6.2 Christchurch earthquake occurred, ETAS was statistically consistent with the total number of observations, while a next-day version of PPE significantly underestimated the observed earthquake count. Conversely, during the 3 months following the Kaikōura earthquake, STEP and ETAS significantly overestimated the observed earthquake rate, while PPE once again significantly underestimated the total number of observations⁶³.

ETAS overforecasting following the Kaikōura earthquake has been partially attributed to magnitude recalibration discrepancies between local data acquisition systems, which biased the model parameters¹⁷. However, based on our test results for California, we recommend considering that the use of a time-invariant background component in this version of ETAS may also explain these number inconsistencies. Regarding the PPE model, our test results differ from those reported for New Zealand, as the ETAS-BACKGROUND-M2 model, a version of PPE for California, overestimated the cumulative number of observed earthquakes by 44% (Fig. 3). Thus, we recommend further prospective evaluations of all the New Zealand OEF models using the pyCSEP toolkit^{19,23} and the newly developed floatCSEP platform⁶⁴.

Lastly, OEF in the United States is provided by the USGS and consists of an automated system for earthquakes in U.S. states and territories and a manual system for particularly complex domestic sequences and international earthquakes with estimated fatalities exceeding 100¹⁷. The automated system is based on the Reasenbergs and Jones model, which combines the Omori–Utsu law with the Gutenberg–Richter relation and treats all aftershocks as ‘offsprings’ of the same mainshock⁶. One limitation of this model is that it does not consider that aftershocks can also trigger other earthquakes, thereby neglecting the contribution of higher-order triggered seismicity to the total earthquake rate. Therefore, the USGS is currently developing a new version of ETAS based on regionalized earthquake parameters to underpin this automated system. The manual system, in contrast, relies on a global version of ETAS that accounts for higher-order aftershock activity, in addition to aleatory variability in earthquake productivity and time-dependent catalog incompleteness issues. This model, however, lacks a background rate component, which could lead to an underestimation of long-term earthquake hazard, particularly in regions of high seismic activity¹⁷.

In the particular case of California, the USGS also operates the UCERF3-ETAS model, which combines ETAS with a long-term model component that provides earthquake probabilities on faults based on elastic rebound statistics⁶⁵. UCERF3-ETAS can be run with or without modeled faults, but given its high computational cost, the model has only been used on demand following notable earthquakes in California, such as the 2019 M_w 7.1 Ridgecrest sequence⁶⁶. To support ongoing OEF developments in the U.S., we recommend: i) calibrating the new ETAS model for the automated system on increasingly small seismicity; ii) considering adaptive/nonparametric spatial methods and time-dependent background components; iii) quantitatively comparing the performance of the new ETAS model with that of an existing ETAS version whose parameters are regularly updated (for the sake of computational efficiency); iv) comparing the global ETAS model with other global forecast models⁵³; and v)

pseudoprospectively comparing the performance of UCERF3-ETAS with that of the next-day forecast models.

The prospective test results and recommendations discussed in this work are intended to be of practical use in improving future probabilistic earthquake forecasts and supporting OEF initiatives worldwide¹⁷. To contribute to these efforts, we accompany this manuscript with a data publication that describes in detail the underlying assumptions and features of all the forecasting models evaluated, as well as a comprehensive tutorial on how to download and access the database of over 50,000 next-day forecasts produced by these models⁷. This extensive database, available in a standardized format, could serve as a reference for developing and rigorously testing new earthquake forecasting models. E.g., users could quantitatively compare new forecasts with this database, develop and rank novel model combinations, and design new statistical metrics to evaluate specific aspects of the models. Furthermore, we provide a peer-reviewed reproducibility software package, containing code, documentation, and data, to transparently replicate the figures, calculations, and results discussed in this work (see the Code Availability section).

We acknowledge that the evaluated statistical models are to some extent limited in their ability to describe seismicity, since the space-magnitude bins that compose them provide only mean values of expected earthquake distributions, which might not fully account for forecast uncertainties²⁷. Therefore, we used multiple statistical metrics that minimize the resulting bias, finding that: i) 95% of the models statistically captured the cumulative number of observed prospective earthquakes; ii) over two-thirds of the models provided spatially consistent forecasts with observations; and iii) 95% of the models provided incremental frequency-magnitude distributions statistically consistent with the magnitude distribution of observed seismicity. These findings mark an significant milestone in earthquake predictability research, providing compelling support for the use of this type of seismicity models in public OEF. Thus, we envision that these prospective test results, benchmarking data, and open-source software tools, will help foster an earthquake culture where vulnerable communities are not necessarily concerned with 'When will the next big one occur?', but rather with 'How reliable will their earthquake forecasts be?'.

Methods

CSEP (in)consistency tests

Most next-day earthquake forecasts are defined within the CSEP-California testing region¹⁹ and are represented as a regular grid of 7,682 cells with a spatial resolution of 0.1°, each of which is subdivided into 0.1 magnitude intervals ranging from 3.95 (or 4.95 in some cases) to 8.05+. Each of the resulting 391,782 space-magnitude bins provides an average number of expected earthquakes that depends on parameters feeding each model and earthquakes observed up to 23:59:59 of the previous day. The parameters of most of these models remained unchanged throughout the evaluation period, however, their generated forecasts updated every 24 hours as data accumulated. We use the CSEP number test²⁶, implemented in the pyCSEP software toolkit¹⁹, to quantitatively evaluate the degree of (dis)agreement between the daily and cumulative number of observed and expected earthquakes. This test assumes that forecast cells are spatially independent of each other and that their expected daily rates follow a Poisson distribution²⁶. Thus, the test calculates the probability P of observing ω earthquakes, given the expected number (or rate) of earthquakes λ :

$$P(\omega|\lambda) = \frac{\lambda^\omega}{\omega!} \exp(-\lambda). \quad (1)$$

Given that the sum of Poisson variables is a Poisson variable with a rate equal to the sum of rates of individual variables, the number test considers that, for each model, the daily expected forecast rate, i.e., the sum over all space-magnitude bins, is a

Poisson process with expectation N_{fore} . The same applies to the overall expected forecast rate, which is simply the sum of the daily expected number of earthquakes over all days. Based on this, the test assesses whether the observed earthquake rate, N_{obs} , falls within the 95% confidence interval of the (two-sided) forecast distribution, using N_{fore} as the mean.

The underlying Poisson number distribution assumed in this (in)consistency test is known to not always be able to fully account for the observed spatiotemporal variability of earthquakes²⁷. Therefore, we complement this test with the CSEP Negative Binomial Distribution (NBD) number test²⁸, whose underlying probability distribution exhibits a larger variance than the Poisson distribution, thereby more effectively accounting for earthquake clustering.

The probability mass function of a NBD is defined as:

$$p(\omega | \tau, \nu) = \frac{\Gamma(\tau + \omega)}{\Gamma(\tau) \omega!} \nu^\omega (1 - \nu)^\tau, \quad (2)$$

with $\omega = 0, 1, 2, \dots$, being the number of observed earthquakes, $\tau > 0$ and $0 \leq \nu \leq 1$ the shape parameters and Γ the Gamma function. The mean μ and variance σ^2 of a NBD can be expressed as:

$$\mu = \tau \frac{1 - \nu}{\nu}; \quad \sigma^2 = \frac{1 - \nu}{\nu^2}. \quad (3)$$

Following previously established CSEP testing methods^{25,67}, we use the total number of earthquakes provided by the models during the time they were available as the mean and estimated the variance from the ANSS historical catalog¹⁸ in non-overlapping periods consistent with the time windows in which the models were in operation. The estimated NBD variances resulted in comparatively wider 95% confidence intervals.

For forecast users working with (symmetric) percentage discrepancies between forecasts and observations, Δ , we also calculate and report these in the main text as:

$$\Delta = \{ |N_{\text{fore}} - N_{\text{obs}}| / \max(N_{\text{fore}}, N_{\text{obs}}) \} * 100 \quad (4)$$

To assess (in)consistencies between the observed and forecasted spatial distributions of earthquakes, we first isolate the spatial component of the forecasts by summing, in each cell and for each model, the expected daily earthquake rate over all magnitude bins and normalizing that rate to match the number of 'activated' cells, i.e., cells containing at least one earthquake. Then, we use the CSEP binary spatial test, which, compared to the CSEP Poisson spatial test²⁶, is based on a Bernoulli likelihood function that calculates the probabilities of observing either 0 or ≥ 1 earthquakes in each forecast cell, under a model²⁵. More explicitly, the probability P_0 of observing $\omega = 0$ earthquakes, given an expected earthquake rate λ and assuming a Poisson distribution (Eq. 1), is $P_0 = \exp(-\lambda)$, while the probability of observing $\omega \geq 1$ events, P_1 , is $P_1 = 1 - P_0 = 1 - \exp(-\lambda)$. For each seismically active day, i.e., when at least one target earthquake is observed, we compute the natural logarithm of these probabilities, or the binary log-likelihood score (hereafter BILL), as:

$$\text{BILL} = c(i) \ln(1 - \exp(-\lambda(i))) + (1 - c(i)) \ln(\exp(-\lambda(i))), \quad (5)$$

where $c(i)$ is the indicator function $c(i) = 1$ if cell i is activated or $c(i) = 0$ otherwise, so that the first term represents the contribution to the log-likelihood if at least one target earthquake falls in a given forecast cell i with expected rate λ and the second term represents the contribution to the BILL if no target earthquakes are observed there, i.e., $c(i) = 0$. Then, we sum the calculated BILLs across all cells i to obtain the observed (spatial) joint log-likelihood score, jBILL, as:

$$\text{jBILL} = \sum_{i=1}^n c(i) \ln(1 - \exp(-\lambda(i))) + (1 - c(i)) \ln(\exp(-\lambda(i))), \quad (6)$$

The greater the log-score, the more consistent the forecast will be with the prospective data. To account for forecast uncertainties, we simulate 10,000 earthquake catalogs consistent with the number of activated cells on each day. The binary spatial test is then summarized by the quantile score ζ , which is the fraction of simulated catalogs of target activated cell with spatial log-likelihoods smaller than the observed jBILL. Assuming a one-sided quantile distribution, if $\zeta > 0.05$, then the spatial forecast can be considered statistically consistent with the spatial distribution of earthquakes⁶⁰.

Given that we evaluate the models over several seismically active days, we assume that, if the models were the data generators, then the distributions of observed and simulated spatial quantiles over multiple periods should be uniform. Consequently, deviations from a uniform distribution will reflect discrepancies between forecasts and observations⁶⁸. Particularly, if the observed quantile distributions (pink curves in Fig. 5) lie above the one-to-one relationship (i.e., the uniform distribution shown in gray), this indicates a high proportion of low quantiles compared with a uniform distribution, suggesting overly localized or inaccurate spatial forecasts. Conversely, if the general trend of each curve lies below the one-to-one line, this indicates a low proportion of low quantiles compared with a uniform distribution, suggesting overly dispersed spatial forecasts (see Fig. 5). We complement these quantile-vs-quantile analyses by calculating the area (A) between observed and uniform quantile distributions, via numerical integration (see the links to the two versions of the reproducibility software package in the Code Availability section). The smaller the area, the more consistent the forecasts and observations will be. Moreover, we perform nonparametric Kolmogorov–Smirnov tests to determine whether we can reject the null hypothesis that each observed spatial quantile distribution is drawn from a uniform quantile distribution. In this case, if the p -value of the test, K_s , is smaller than 0.05, we can conclude that a significant difference between both distributions exists.

We then use the CSEP magnitude test²⁶ to evaluate the ability of the models to forecast the overall magnitude distribution of observed earthquakes. To do so, for each day, we isolate the magnitude component of the forecasts by summing their forecast rates per magnitude bin over all spatial cells and normalizing the rate to the number of observed earthquakes. Then, we estimate the overall expected magnitude distribution by summing the expected daily magnitude distributions for all the days that each model was in operation. Thus, the magnitude test evaluates whether the observed magnitude distribution falls within the 95% predictive interval of the expected distribution, which is obtained from 10,000 simulated catalogs that are consistent with the cumulative number of observations. This test is summarized by the quantile score ϕ , which is the fraction of simulated earthquake catalogs with magnitude log-likelihoods smaller than the observed magnitude statistic. We again use a significance level of 0.05 to identify (dis)agreements between forecasts and observations.

We complement these magnitude test results by calculating 95% confidence intervals of predicted earthquake rates within each magnitude bin, assuming a Poisson distribution (gray shades in Fig. 7). If the observed incremental frequency-magnitude distributions (pink curves) fall within these intervals, forecasts and observations can be considered statistically indistinguishable.

Furthermore, we perform two-sample Kolmogorov–Smirnov tests to assess whether the expected and observed magnitude distributions of earthquakes come from the same underlying discrete distribution. To asymptotically calculate the p -value of this test, annotated as K_m in Fig. 7, we first normalize the observed and expected incremental magnitude distributions to the total number of observations N_{obs} , as:

$$P_o = \frac{N_{o_j}}{N_{\text{obs}}}; \quad P_p = \frac{N_{p_j}}{N_{\text{obs}}}, \quad (7)$$

where N_{o_j} and N_{p_j} are the number of observed and predicted earthquakes within the magnitude bin j , given the catalog and the model, respectively. Then, we calculate the empirical cumulative distribution functions of P_o and P_p , F_o and F_p , respectively, so we can compute the Kolmogorov–Smirnov statistic D as:

$$D = \max |F_o - F_p| \quad (8)$$

Based on D , we then calculate K_m as a series approximation given by:

$$K_m \approx 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 \cdot z^2), \quad (9)$$

where $z = D \cdot \sqrt{N_{\text{obs}}/2}$ and k takes values from 1 to 1000.

CSEP comparative tests

In addition to (in)consistency analyses between forecasts and observations, our model evaluation includes quantitative comparisons between models. We assess the performance of each next-day forecasting model relative to the HKJ-ADAPTIVE-M2 benchmark by calculating, daily and cumulatively, binary information gains per activated spatial forecast cell (IGPA), which are measures of the relative amount of information about future activated cells that two competing models can provide²⁵. To estimate these, we first compute the daily probabilities of observing either 0 or 1 (or more) target earthquakes in their actual space-magnitude (i, j) bins, under two competing models M_x (with $x = 1, 2$) as:

$$\text{BILL}_{M_x} = c(i, j) \ln(1 - \exp(-\lambda_{M_x}(i, j))) + (1 - c(i, j)) \ln(\exp(-\lambda_{M_x}(i, j))). \quad (10)$$

Similar to the binary spatial test, λ_{M_x} has previously been normalized to the number of activated space-magnitude bins. M_1 and M_2 denote the HKJ-ADAPTIVE-M2 benchmark and the competing next-day model, respectively. We then sum over all space-magnitude bins to obtain joint binary likelihood scores per day, jBILL_{M_1} and jBILL_{M_2} , which we then subtract as $\text{jBILL}_{M_2} - \text{jBILL}_{M_1}$ to obtain the IGPA of the competing model over the benchmark:

$$\text{IGPA} = \frac{1}{M_{\text{ac}}} \sum_{i=1}^{M_{\text{ac}}} \text{jBILL}_{M_2} - \text{jBILL}_{M_1}, \quad (11)$$

where M_{ac} is the total number of spatial cells activated over different evaluation periods (see Fig. 8). We then sum these gains over all days on which forecasts provided by the two competing models are available (note that forecasts are missing for some

days)⁷, regardless whether seismicity is observed or not, so that the resulting value is the cumulative binary information gain (or loss) of a competing model over HKJ-ADAPTIVE-M2. The more positive the IGPA, the better the relative performance of the competing model will be.

To assess whether IGPAs are statistically significant, we perform one-sided paired Student's *t*-tests⁶⁹. This comparative test evaluates whether we can reject, at a 0.05 significant level, the null hypothesis that, for each pair of models, the mean of the differences between their cumulative jBILLS is zero. To do this, we first calculate *t*-values, which are given by the ratio between the mean of the differences in jBILLS and the standard deviation of the mean (assuming a normal distribution). Thus, based on *t*-values and *n* - 1 degrees of freedom, with *n* being the number of analyzed days in each cumulative curve in Fig. 8, we finally calculate the *p*-values of the comparative test. *p*-values > 0.05 indicate that the competing model is statistically significantly more informative than the benchmark model.

Data Availability

All forecast files in HDF5 format⁷, along with fully documented code and a tutorial on how to download and access the data, are available on Zenodo under a Creative Commons open-access license at <https://zenodo.org/records/15076187>. Earthquake forecasts generated by individual models can also be downloaded directly from the CSEP website at <https://cseptest.org/grid-based-forecasts/>. The earthquake catalog used for this prospective evaluation is freely available at <https://earthquake.usgs.gov/earthquakes/search/> (last accessed June 2026), while the USGS Quaternary Fault database can be found at <https://www.usgs.gov/programs/earthquake-hazards/faults>.

Code Availability

Both the code and tutorial for accessing the forecast database, as well as the test results discussed in this study, rely heavily on the pyCSEP software toolkit^{19,23}, which is open access and features comprehensive documentation on GitHub at <https://github.com/SCECcode/pycsep>. The peer-reviewed version of the reproducibility software package accompanying this manuscript can be found on Ocean Code at <https://codeocean.com/capsule/3685334/tree>, while the non-peer-reviewed version is freely available on GitHub at https://github.com/bayonato89/reproducibility_next_day_california. If the user employs the peer-reviewed version of this software package, please cite it as: Bayona, J.A., Serafini, F., Silva, F., Iturrieta, P., Savran, W.H., Herrmann, M., Marzocchi, W., Maechling, P.J. and Werner, M.J., Select earthquake forecasting models demonstrate consistency with prospective decadal observations in California, Ocean Code, <https://doi.org/10.24433/CO.0698383.v1>, 2026.

References

1. Geller, R. Earthquake prediction: a critical review. *Geophys. J. Int.* **131**, 425–450, <https://doi.org/10.1111/j.1365-246X.1997.tb06588.x> (1997).
2. Blettery, Q. & Nocquet, J. The precursory phase of large earthquakes. *Science* **381**, 297–301, <https://doi.org/10.1126/science.adg2565> (2023).
3. Owen, R. Do earthquakes produce signals before they rupture? maybe, scientists say. *Temblor* <http://doi.org/10.32858/temblor.317> (2023).

4. Schorlemmer, D. *et al.* The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismol. Res. Lett.* **89**, 1305–1313, <https://doi.org/10.1785/0220180053> (2018).
5. Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. association* **83**, 9–27, <https://doi.org/10.1080/01621459.1988.10478560> (1988).
6. Reasenberg, P. & Jones, L. Earthquake hazard after a mainshock in California. *Science* **243**, 1173–1176, <https://doi.org/10.1126/science.243.4895.1173> (1989).
7. Serafini, F. *et al.* A benchmark database of ten years of prospective next-day earthquake forecasts in california from the collaboratory for the study of earthquake predictability. *Sci. Data* **12**, 1501, <https://doi.org/10.1038/s41597-025-05766-3> (2025).
8. Zhuang, J. Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, planets space* **63**, 207–216, <https://doi.org/10.5047/eps.2010.12.010> (2011).
9. Gerstenberger, M., Wiemer, S., Jones, L. & Reasenberg, P. Real-time forecasts of tomorrow’s earthquakes in California. *Nature* **435**, 328–331, <https://doi.org/10.1038/nature03622> (2005).
10. Omori, F. On after-shocks. *Seismol. journal Jpn.* **19**, 71–80 (1894).
11. Gutenberg, B. & Richter, C. F. Frequency of earthquakes in California. *Bull. Seismol. society Am.* **34**, 185–188 (1944).
12. Rhoades, D. & Evison, F. Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure applied geophysics* **161**, 47–72, <https://doi.org/10.1007/s00024-003-2434-9> (2004).
13. Rhoades, D. Mixture models for improved earthquake forecasting with short-to-medium time horizons. *Bull. Seismol. Soc. Am.* **103**, 2203–2215, <https://doi.org/10.1785/0120120233> (2013).
14. Gordon, J., Fox, E. & Schoenberg, F. A nonparametric Hawkes model for forecasting California seismicity. *Bull. Seismol. Soc. Am.* **111**, 2216–2234, <https://doi.org/10.1785/0120200349> (2021).
15. Helmstetter, A. & Werner, M. Adaptive smoothing of seismicity in time, space, and magnitude for time-dependent earthquake forecasts for California. *Bull. Seismol. Soc. Am.* **104**, 809–822, <https://doi.org/10.1785/0120130105> (2014).
16. Jordan, T. *et al.* Operational Earthquake Forecasting: State of knowledge and guidelines for implementation. *Annals Geophys.* <https://doi.org/10.4401/ag-5350> (2011).
17. Mizrahi, L. *et al.* Developing, testing, and communicating earthquake forecasts: Current practices and future directions. *Rev. Geophys.* **62**, <https://doi.org/10.1029/2023RG000823> (2024).
18. Guy, M. R. *et al.* National Earthquake Information Center systems overview and integration. Tech. Rep., US Geological Survey (2015).
19. Savran, W. *et al.* pyCSEP: a Python toolkit for earthquake forecast developers. *Seismol. Soc. Am.* **93**, 2858–2870, <https://doi.org/10.1785/0220220033> (2022).
20. Hauksson, E. *et al.* The 2010 mw 7.2 El Mayor-Yucapah earthquake sequence, Baja California, Mexico and southernmost California, USA: active seismotectonics along the Mexican pacific margin. *Pure Appl. Geophys.* **168**, 1255–1277, <https://doi.org/10.1007/s00024-010-0209-7> (2011).

21. Wei, S. *et al.* The 2012 Brawley swarm triggered by injection-induced aseismic slip. *Earth Planet. Sci. Lett.* **422**, 115–125, <https://doi.org/10.1016/j.epsl.2015.03.054> (2015).
22. Brocher, T. *et al.* The Mw6.0 24 August 2014 South Napa earthquake. *Seismol. Res. Lett.* **86**, 309–326, <https://doi.org/10.1785/0220150004> (2015).
23. Graham, K. *et al.* New features in the pyCSEP toolkit for earthquake forecast development and evaluation. *Seismol. Soc. Am.* **20**, 1–15, <https://doi.org/10.1785/0220240197> (2024).
24. Helmstetter, A., Kagan, Y. & Jackson, D. High-resolution time-independent grid-based forecast for $m \geq 5$ earthquakes in California. *Seismol. Res. Lett.* **78**, 78–86, <https://doi.org/10.1785/gssrl.78.1.78> (2007).
25. Bayona, J., Savran, W., Rhoades, D. & Werner, M. Prospective evaluation of multiplicative hybrid earthquake forecasting models in California. *Geophys. J. Int.* **229**, 1736–1753, <https://doi.org/10.1093/gji/ggac018> (2022).
26. Zechar, J. *et al.* The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. *Concurr. Comput. Pract. Exp.* **22**, 1836–1847, <https://doi.org/10.1002/cpe.1519> (2010).
27. Lombardi, A. & Marzocchi, W. The ETAS model for daily forecasting of italian seismicity in the csep experiment. *Annals Geophys.* (2010).
28. Werner, M. J., Zechar, W., J. D. and Marzocchi & Wiemer, S. Retrospective evaluation of the five-year and ten-year csep-italy earthquake forecasts. *arXiv preprint arXiv:1003.1092* <https://doi.org/10.48550/arXiv.1003.1092> (2010).
29. Hardebeck, J. *et al.* Aftershock forecasting. *Annu. Rev. Earth Planet. Sci.* **52**, <https://doi.org/10.1146/annurev-earth-040522-102129> (2024).
30. Storesund, R. *et al.* M6.5 earthquake offshore northern California january 9, 2010. *GEER Field Reconnaissance Summ.* (2010).
31. Båth, M. Lateral inhomogeneities of the upper mantle. *Tectonophysics* **2**, 483–514, [https://doi.org/10.1016/0040-1951\(65\)90003-X](https://doi.org/10.1016/0040-1951(65)90003-X) (1965).
32. Llenos, A. & Michael, A. Forecasting the (un) productivity of the 2014 m6.0 South Napa aftershock sequence. *Seismol. Res. Lett.* **88**, 1241–1251, <https://doi.org/10.1785/0220170050> (2017).
33. Hainzl, S. Seismicity patterns of earthquake swarms due to fluid intrusion and stress triggering. *Geophys. J. Int.* **159**, 1090–1096, <https://doi.org/10.1111/j.1365-246X.2004.02463.x> (2004).
34. Hainzl, S. Rate-dependent incompleteness of earthquake catalogs. *Seismol. Res. Lett.* **87**, 337–344, <https://doi.org/10.1785/0220150211> (2016).
35. Spassiani, I., Falcone, G., Murru, M. & Marzocchi, W. Operational earthquake forecasting in Italy: validation after 10 yr of operativity. *Geophys. J. Int.* **234**, 2501–2518, <https://doi.org/10.1093/gji/ggad256> (2023).
36. Hardebeck, J., Llenos, A., Michael, A., Page, M. & Van Der Elst, N. Updated California aftershock parameters. *Seismol. Res. Lett.* **90**, 262–270, <https://doi.org/10.1785/0220180240> (2019).
37. Marzocchi, W., Lombardi, A. & Casarotti, E. The establishment of an operational earthquake forecasting system in Italy. *Seismol. Res. Lett.* **85**, 961–969, <https://doi.org/10.1785/0220130219> (2014).

38. Page, M., van der Elst, N., Hardebeck, J., Felzer, K. & Michael, A. J. Three ingredients for improved global aftershock forecasts: Tectonic region, time-dependent catalog incompleteness, and intersequence variability. *Bull. Seismol. Soc. Am.* **106**, 2290–2301, <https://doi.org/10.1785/0120160073> (2016).
39. Dascher-Cousineau, K., Brodsky, E., Lay, T. & Goebel, T. What controls variations in aftershock productivity? *J. Geophys. Res. Solid Earth* **125**, e2019JB018111, <https://doi.org/10.1029/2019JB018111> (2020).
40. Churchill, R., Werner, M., Biggs, J. & Fagereng, Å. Relative afterslip moment does not correlate with aftershock productivity: Implications for the relationship between afterslip and aftershocks. *Geophys. Res. Lett.* **49**, e2022GL101165, <https://doi.org/10.1029/2022GL101165> (2022).
41. Trugman, D. & Ben-Zion, Y. Coherent spatial variations in the productivity of earthquake sequences in California and Nevada. *The Seism. Rec.* **3**, 322–331, <https://doi.org/10.1785/0320230039> (2023).
42. Manganiello, E., Herrmann, M. & Marzocchi, W. New physical implications from revisiting foreshock activity in southern California. *Geophys. Res. Lett.* **50**, e2022GL098737, [10.1029/2022GL098737](https://doi.org/10.1029/2022GL098737) (2023).
43. Kagan, Y. & Jackson, D. Short-and long-term earthquake forecasts for California and Nevada. *Pure Appl. Geophys.* **167**, 685–692, <https://doi.org/10.1007/s00024-010-0073-5> (2010).
44. Schoenberg, F. The magnificent partnership of yan kagan and dave jackson, <https://doi.org/10.1785/0220240188> (2024).
45. Llenos, A., McGuire, J. & Ogata, Y. Modeling seismic swarms triggered by aseismic transients. *Earth Planet. Sci. Lett.* **281**, 59–69, <https://doi.org/10.1016/j.epsl.2009.02.011> (2009).
46. McBride, S. K., Llenos, A. L., Page, M. T. & van der Elst, N. # earthquakeadvisory: Exploring discourse between government officials, news media, and social media during the 2016 bombay beach swarm. *Seismol. Res. Lett.* **91**, 438–451, <https://doi.org/10.1785/0220190082> (2020).
47. Wald, D. J., Quitoriano, V., Worden, C. B., Hopper, M. & Dewey, J. W. Usgs “did you feel it?” internet-based macroseismic intensity maps. *Annals geophysics* **54**, <https://doi.org/10.4401/ag-5354> (2011).
48. Goldberg, D. E., Koch, P., Melgar, D., Riquelme, S. & Yeck, W. L. Beyond the teleseism: Introducing regional seismic and geodetic data into routine usgs finite-fault modeling. *Seismol. Soc. Am.* **93**, 3308–3323 (2022).
49. Zeng, Y., Petersen, M. D. & Shen, Z.-K. Earthquake potential in california-nevada implied by correlation of strain rate and seismicity. *Geophys. Res. Lett.* **45**, 1778–1785 (2018).
50. Beroza, G., Segou, M. & Mostafa Mousavi, S. Machine learning and earthquake forecasting—next steps. *Nat. communications* **12**, 4761, <https://doi.org/10.1038/s41467-021-24952-6> (2021).
51. USGS. *Advanced National Seismic System—current Status, Development Opportunities, and Priorities for 2017-2027* (US Geological Survey, 2017).
52. Herrmann, M. & Marzocchi, W. Maximizing the forecasting skill of an ensemble model. *Geophys. J. Int.* **234**, 73–87, <https://doi.org/10.1093/gji/ggad020> (2023).
53. Bird, P., Jackson, D. D., Kagan, Y. Y., Kreemer, C. & Stein, R. S. Gear1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity. *Bull. Seismol. Soc. Am.* **105**, 2538–2554, <https://doi.org/10.1785/0120150058> (2015).

54. Strader, A. *et al.* Prospective evaluation of global earthquake forecast models: 2 yrs of observations provide preliminary support for merging smoothed seismicity with geodetic strain rates. *Seismol. Res. Lett.* **89**, 1262–1271, <https://doi.org/10.1785/0220180051> (2018).
55. Bayona, J. *et al.* Two global ensemble seismicity models obtained from the combination of interseismic strain measurements and earthquake-catalogue information. *Geophys. J. Int.* **224**, 1945–1955, <https://doi.org/10.1093/gji/ggaa554> (2021).
56. Iturrieta, P. *et al.* Evaluation of a decade-long prospective earthquake forecasting experiment in Italy. *Seismol. Res. Lett.* **95**, 3174–3191, <https://doi.org/10.1785/0220230247> (2024).
57. Mancini, S., Segou, M., Werner, M. J. & Cattania, C. Improving physics-based aftershock forecasts during the 2016–2017 central Italy earthquake cascade. *J. Geophys. Res. Solid Earth* **124**, 8626–8643, <https://doi.org/10.1029/2019JB017874> (2019).
58. Vazquez, L. A. & Jordan, T. H. Assimilation of deterministic multicycle earthquake simulations into probabilistic rupture forecasts. *Geophys. J. Int.* **241**, 1243–1261, <https://doi.org/10.1093/gji/ggaf101> (2025).
59. Stockman, S., Lawson, D. J. & Werner, M. J. Forecasting the 2016–2017 central apennines earthquake sequence with a neural point process. *Earth's Futur.* **11**, e2023EF003777, <https://doi.org/10.1029/2023EF003777> (2023).
60. Taroni, M. *et al.* Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy. *Seismol. Res. Lett.* **89**, 1251–1261, <https://doi.org/10.1785/0220180031> (2018).
61. Marzocchi, W., Taroni, M. & Falcone, G. Earthquake forecasting during the complex amatrice-norcia seismic sequence. *Sci. Adv.* **3**, e1701239, <https://doi.org/10.1126/sciadv.1701239> (2017).
62. Christophersen, A. *et al.* Progress and challenges in operational earthquake forecasting in New Zealand. In *New Zealand society for earthquake engineering annual technical conference*, 27–29 (2017).
63. Rhoades, D. *et al.* Highlights from the first ten years of the New Zealand earthquake forecast testing center. *Seismol. Res. Lett.* **89**, 1229–1237, <https://doi.org/10.1785/0220180032> (2018).
64. Iturrieta, P. *et al.* floatcsep: An application to deploy and conduct reproducible prospective earthquake forecasting experiments. *J. Open Source Softw.* **11**, 9408, <https://doi.org/10.21105/joss.09408> (2026).
65. Field, E. H. *et al.* A spatiotemporal clustering model for the third uniform California earthquake rupture forecast (ucrf3-etlas): Toward an operational earthquake forecast. *Bull. Seismol. Soc. Am.* **107**, 1049–1081, <https://doi.org/10.1785/0120160173> (2017).
66. Milner, K. R., Field, E. H., Savran, W. H., Page, M. T. & Jordan, T. H. Operational earthquake forecasting during the 2019 Ridgecrest, California, earthquake sequence with the ucrf3-etlas model. *Seismol. Res. Lett.* **91**, 1567–1578, <https://doi.org/10.1785/0220190294> (2020).
67. Werner, M., Helmstetter, A., Jackson, D. & Kagan, Y. High-resolution long-term and short-term earthquake forecasts for California. *Bull. Seismol. Soc. Am.* **101**, 1630–1648, <https://doi.org/10.1785/0120090340> (2011).
68. Savran, W. *et al.* Pseudoprospective evaluation of UCERF3-ETAS forecasts during the 2019 Ridgecrest sequence. *Bull. Seismol. Soc. Am.* **110**, 1799–1817, <https://doi.org/10.1785/0120200026> (2020).
69. Rhoades, D. A. *et al.* Efficient testing of earthquake forecasting models. *Acta Geophys.* **59**, 728–747, <https://doi.org/10.2478/s11600-011-0013-5> (2011).

70. Wells, D. L. & Coppersmith, K. J. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull. seismological Soc. Am.* **84**, 974–1002, <https://doi.org/10.1785/BSSA0840040974> (1994).

Acknowledgments

We sincerely thank David Rhoades, Frederic Schoenberg, Jiancang Zhuang, and Matthew Gerstenberger for providing detailed descriptions of their earthquake forecasting models.

Funding Statement

This research was supported by the Leverhulme Trust through its Early Career Fellowship program, the USGS (grant numbers G24AP00059 and G25AP00379), and the Statewide California Earthquake Center (SCEC; contribution No. 14154). SCEC is funded by NSF Cooperative Agreement EAR-2225216 and USGS Cooperative Agreement G24AC00072-00.

Author Contributions Statement

J.A.B. funded the project, performed the tests, created the figures, wrote the manuscript, and produced the reproducibility software packages. F.Se. created the HDF5 forecast files, designed and documented the code and tutorial to access the database, supervised the statistical analyses, and edited the manuscript. F.Si. contributed to managing the original forecast files from which the HDF5 files were created, and reviewed the manuscript. P.I. made significant contributions to pyCSEP, created Fig. 2, and edited the manuscript. W.H.S. led the development of pyCSEP and edited the manuscript. M.H. contributed significantly to pyCSEP and substantially edited the manuscript. W.M. discussed the results in terms of OEF implications and edited the manuscript. P.J.M. contributed to pyCSEP, facilitated the management of original forecast files, and reviewed the manuscript. M.J.W. conceived and funded the project, supervised the investigation, and edited the manuscript.

Competing Interests Statement

The authors declare no competing interests.

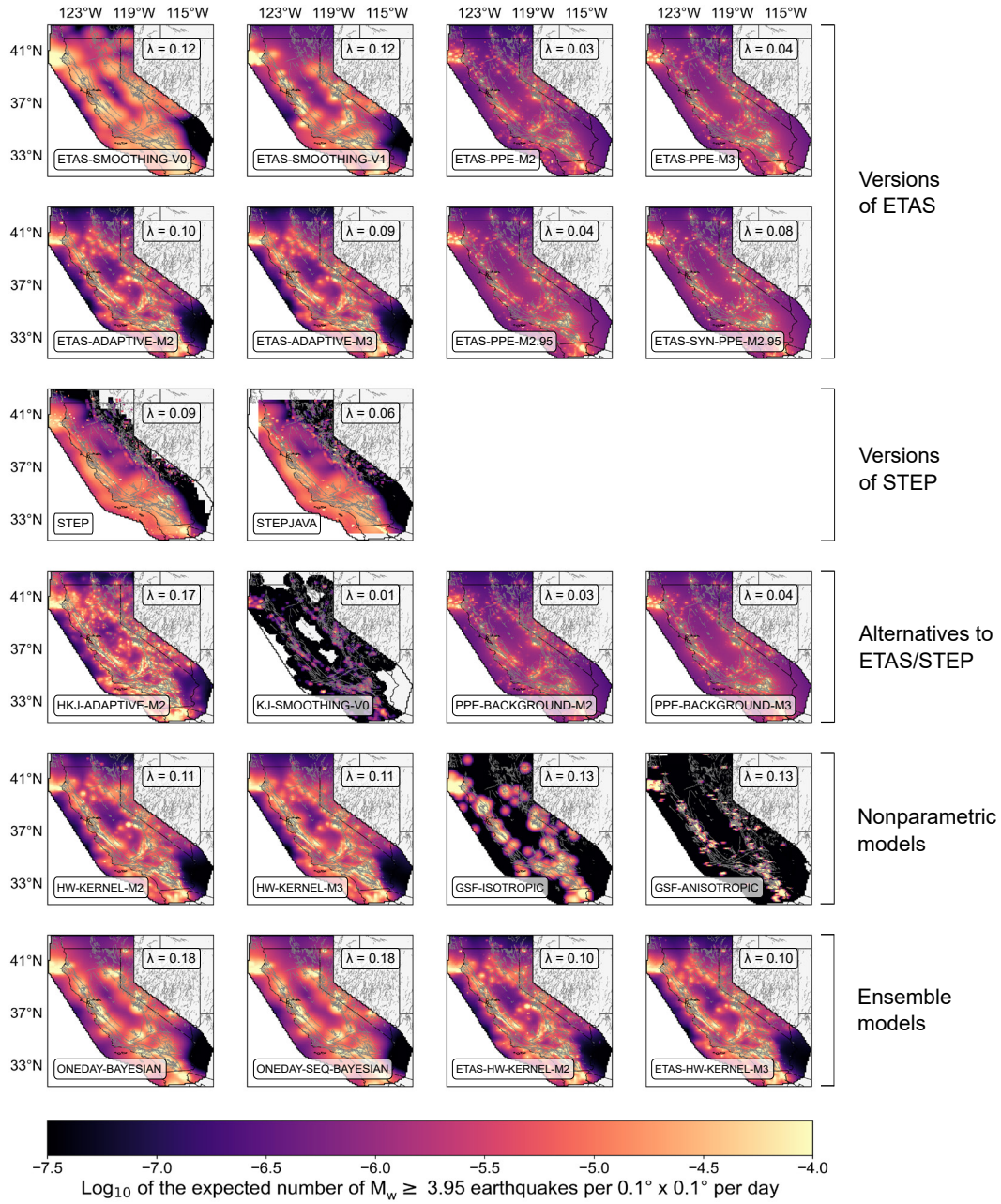


Figure 1. Forecast maps generated by 21 $M_w \geq 3.95$ next-day seismicity models and one time-invariant model (HKJ-ADAPTIVE-M2) for 4 different seismically calm days in California. Earthquake rates are expressed per $0.1^\circ \times 0.1^\circ$ cell per day, with bright colors denoting regions where seismic activity is comparatively high, dark colors indicating comparatively low rates, and white areas denoting regions where the models are not defined. Since the evaluated models were developed and submitted to CSEP at different times⁷, this example shows a STEP forecast issued on January 21, 2013, ONEDAY(-SEQ)-BAYESIAN forecasts issued on August 7, 2016, and ETAS-ADAPTIVE-M2, ETAS-HW-KERNEL-M2, and HW-KERNEL-M3 forecasts created on September 19, 2016. The remaining forecasts were created on August 30, 2018. Annotated λ values indicate the number of $M_w \geq 3.95$ earthquakes expected by the models during those specific days. Fault traces (shown in gray) are obtained from the USGS Quaternary Fault and Fold Database (see the Data Availability section).

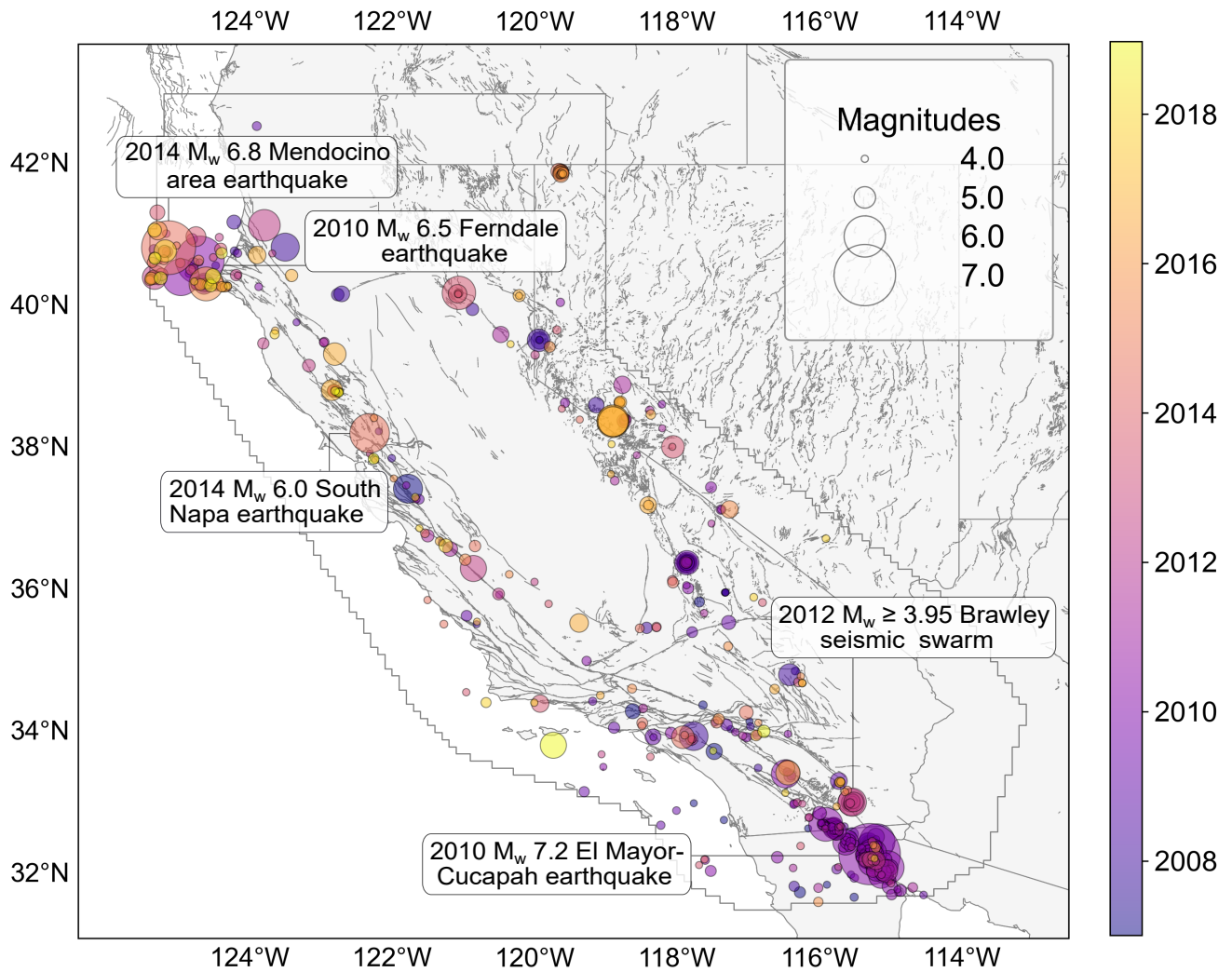


Figure 2. Spatiotemporal distribution of 597 earthquakes with magnitudes $M_w \geq 3.95$ and depths ≤ 30 km observed in the CSEP-California testing region (gray solid polygon) between August 1, 2007, and August 30, 2018. During this period, the evaluated models prospectively generated next-day $M_w \geq 3.95$ seismicity forecasts for California. This evaluation dataset includes the 2010 M_w 6.5 Ferndale, 2010 M_w 7.2 El Mayor–Cucapah, 2014 M_w 6.8 Mendocino, and 2014 M_w 6.0 Napa earthquake sequences, as well as the 2012 $M_w \geq 3.95$ Brawley seismic swarm.

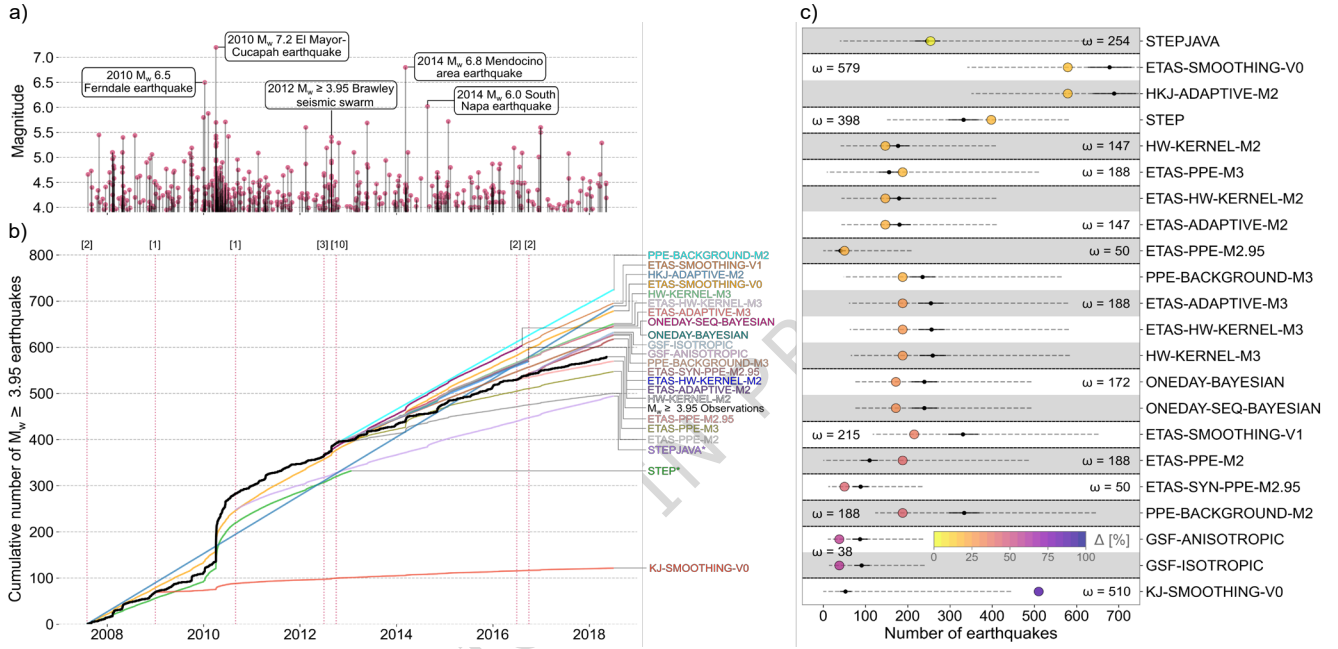


Figure 3. Prospective results of cumulative number tests for next-day $M_w \geq 3.95$ earthquake forecasting models for California. a) Magnitude-time series of observed $M_w \geq 3.95$ seismicity in California between August 2007 and August 2018. b) Cumulative distributions of observed and expected target earthquakes throughout the prospective evaluation period. The vertical dashed lines indicate the dates on which new models became operational (see the specific number above each line). Label asterisks highlight versions of STEP that are defined in regions smaller than the CSEP-California test region (see Fig. 1) and should therefore not be directly compared to the number of earthquakes observed in that region (black curve). c) Prospective results of the cumulative Poisson and Negative Binomial Distribution (NBD) number tests for the next-day seismicity models. The circles represent the number of earthquakes (ω) observed during the different periods in which the models were prospectively generating forecasts. The colors denote (symmetric) percentage discrepancies Δ between forecasts and observations (see the colorbar). The models are ordered from lowest to highest percentage differences. Solid black bars and dashed gray bars depict the 95% confidence intervals of the Poisson and NBD number distributions of the models, respectively, using their total number of expected earthquakes (black points) as the mean of each distribution.

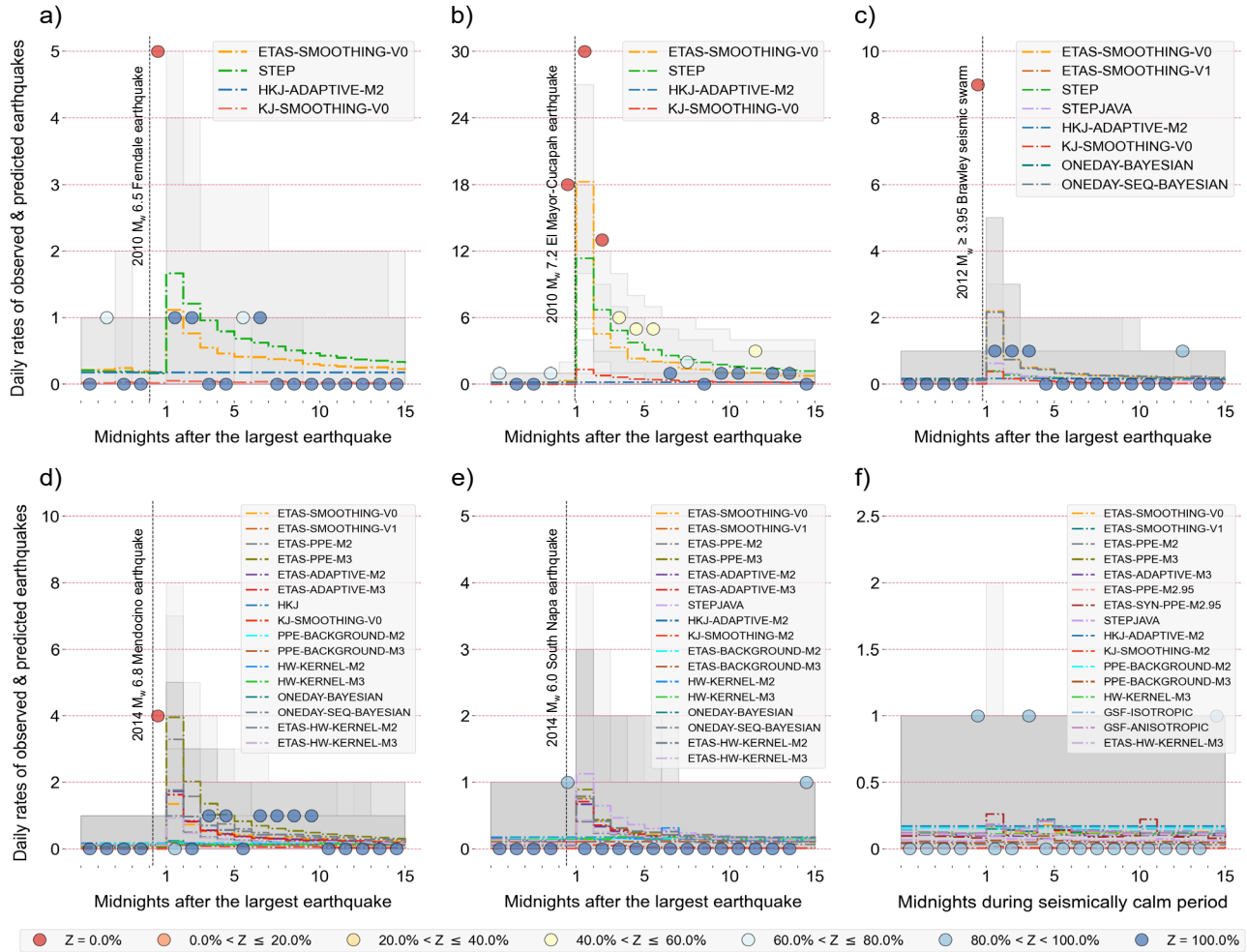


Figure 4. Ability of the next-day models to forecast the daily number of $M_w \geq 3.95$ earthquakes (circles) observed during the first two weeks of specific earthquake sequences/scenarios in California. These include: a) the 2010 M_w 6.5 Ferndale; b) 2010 M_w 7.2 El Mayor–Cucapah; c) 2012 $M_w \geq 3.95$ Brawley; d) 2014 M_w 6.8 Mendocino; and e) 2014 M_w 6.0 South Napa earthquake sequences, as well as f) a randomly selected two-week period in January 2018 of relative seismic calm in California. The vertical dashed lines denote the occurrence of the largest events in relation to specific midnights at which next-day forecasts were issued. If the daily number of observed earthquakes falls within the 95% models' confidence intervals (gray shades), observations and forecasts can be considered statistically consistent. The colors of the circles represent the percentage (Z) of available models that capture the observed daily counts (see the colorbar). The grayer the shades, the greater the overlap between models' confidence intervals.

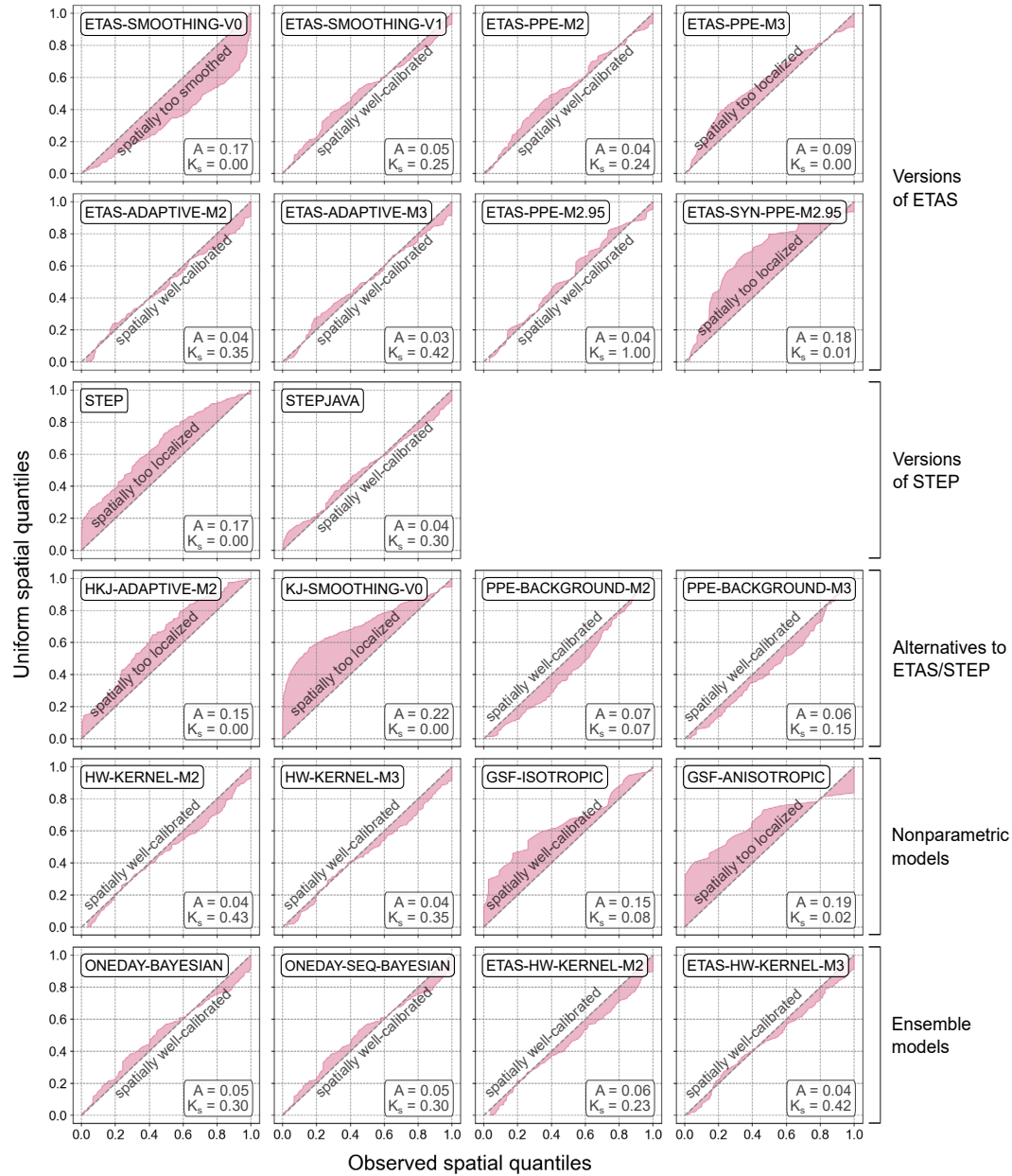


Figure 5. Quantile-vs-quantile plots, or so-called calibration plots, comparing the distribution of observed spatial quantiles (in ascending order) with spatial quantiles drawn from a uniform distribution. If the models were the data generator, the simulated and observed spatial quantiles should maintain a one-to-one relationship, i.e., the solid pink curves (the observed quantile distributions) should lie along the dashed gray diagonals (the uniform quantile distributions). Deviations indicate inconsistencies between forecasts and observations (see the annotated explanations). The pink shades are the annotated areas (A) between the diagonals and the curves. The larger the area, the greater the discrepancies between forecasts and observations. Annotated K_s values are the p -values of nonparametric Kolmogorov–Smirnov tests, which assess whether the (continuous) observed quantile distributions are drawn from uniform quantile distributions. If $K_s > 0.05$, this indicates that models generate spatial forecasts that are statistically consistent with the spatial distribution of observed earthquakes.

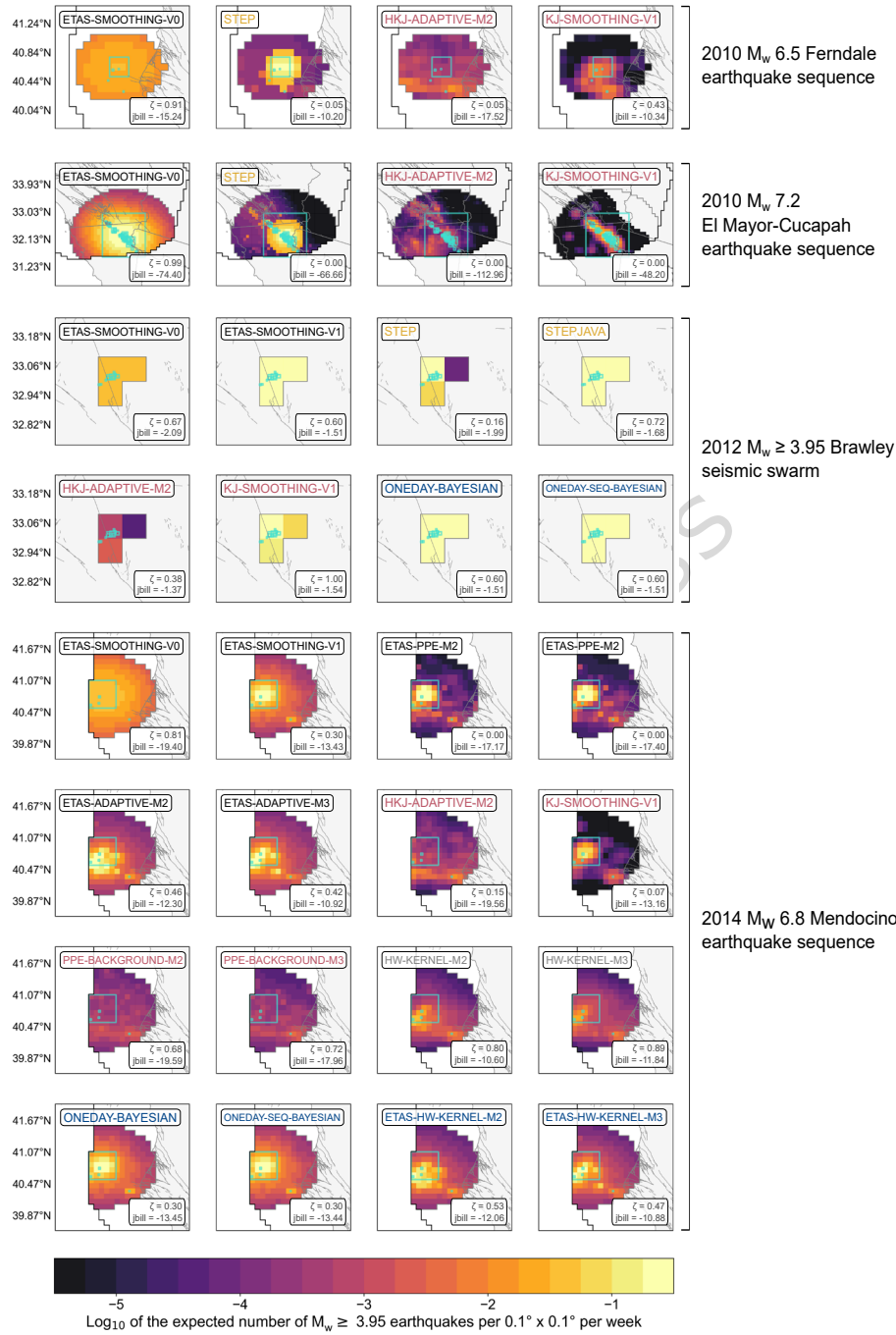


Figure 6. Earthquake and aftershock forecasts generated by the time-invariant HKJ-ADAPTIVE-M2 model and 17 $M_w \geq 3.95$ seismicity models available during the first seven days of 4 major earthquake sequences in California. These sequences are those of the 2010 M_w 6.5 Ferndale, M_w 7.2 El Mayor-Cucapah, and 2014 M_w 6.8 Mendocino earthquakes, and the 2012 $M_w \geq 3.95$ Brawley seismic swarm. The spatial forecasts result from stacking seven consecutive next-day earthquake forecasts, starting on the specific dates of the largest events. Observed earthquakes are shown as turquoise squares and expected seismicity within the aftershock zones of the larger events, defined as circular regions with radii equal to three times their average fault lengths⁷⁰, is color-coded. Annotated 'jbill' values are binary log-likelihoods of observing the data under the models. Greater jbill values indicate greater consistency with observations. If the annotated spatial quantile ζ is greater than 0.05, the spatial forecasts are statistically consistent with the spatial distribution of observed seismicity. The name of each model is colored according to its type: black: ETAS versions; gold: STEP versions; pink: alternative models to ETAS and STEP; blue: ensemble models; and gray: nonparametric models.

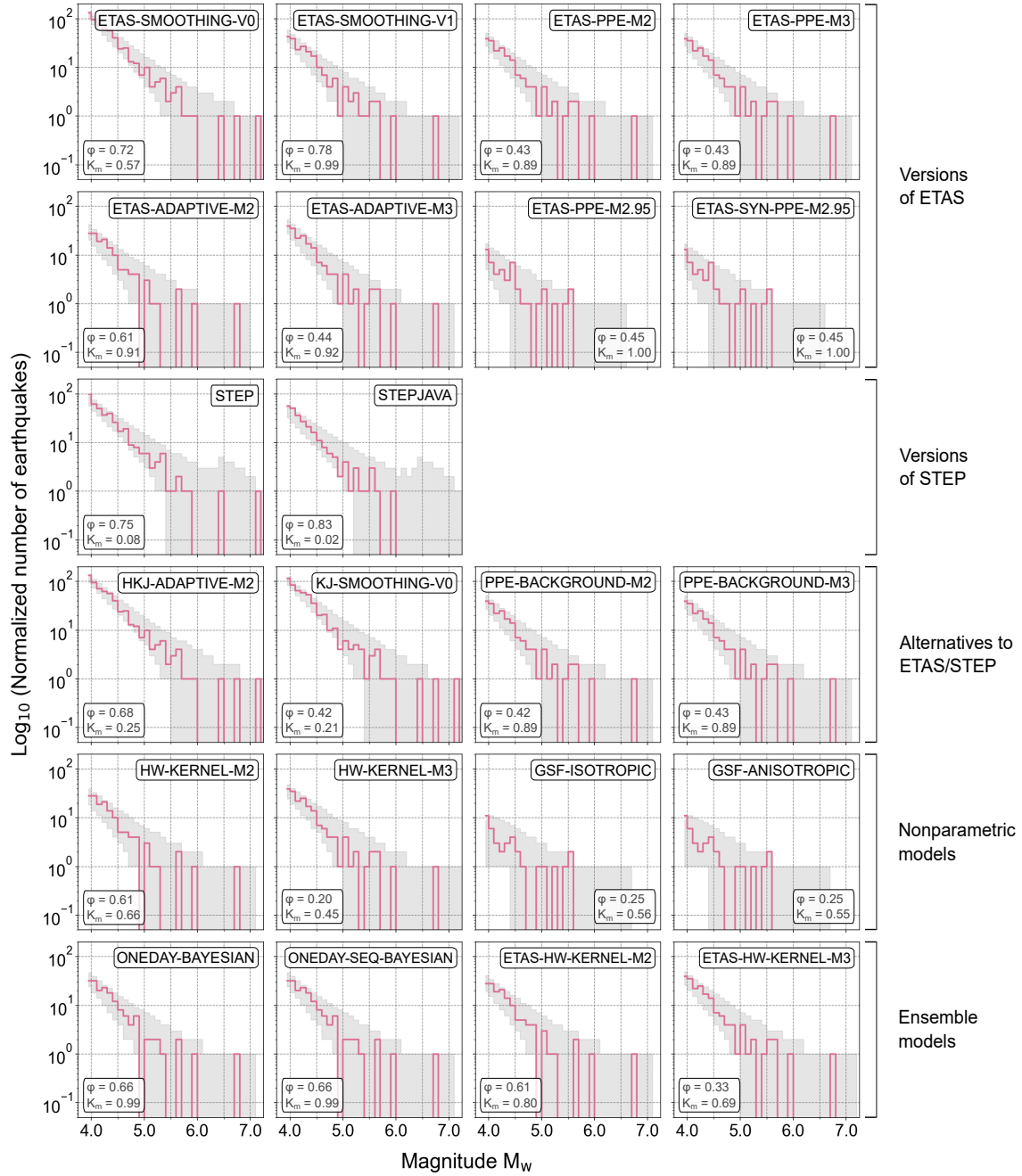


Figure 7. Comparisons between observed and expected frequency-magnitude distributions of $M_w \geq 3.95$ earthquakes in California during the prospective evaluation periods in which the models were operational. For each pink curve, the x -axis shows the earthquake magnitude range divided into 0.1 unit intervals, while the y -axis shows the number of earthquakes observed within each magnitude bin. The models' 95% confidence intervals (gray shading) are obtained by normalizing the expected earthquake rates to the total number of observations and assuming a Poisson number distribution. Annotated ϕ and K_m values are p -values of the CSEP magnitude test²⁶ and two-sample Kolmogorov–Smirnov tests, respectively. Models obtaining p -values larger than 0.05 are considered statistically consistent with the magnitude distribution of observed earthquakes.

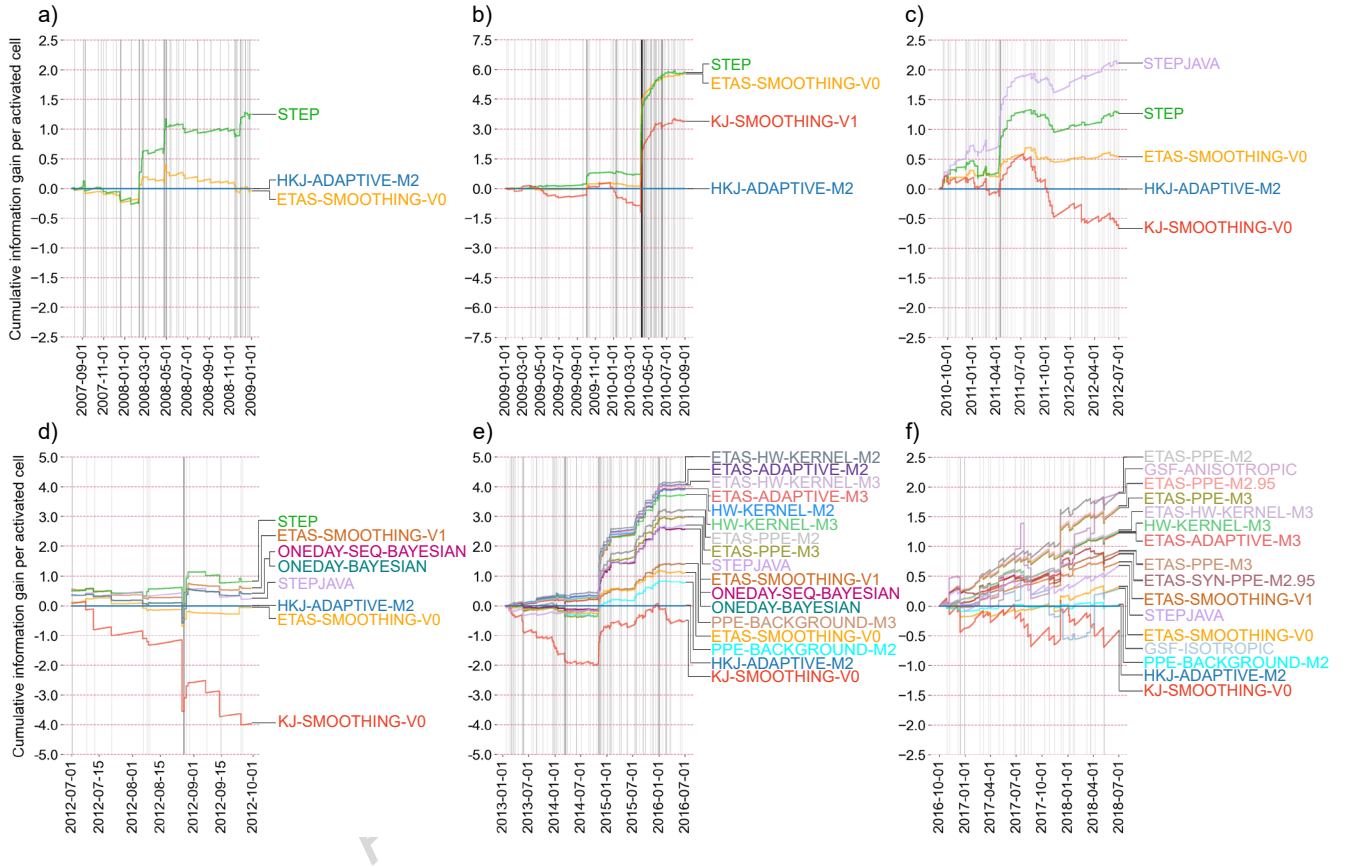


Figure 8. Cumulative binary information gains per activated spatial forecast cell (IGPA) obtained by each $M_w \geq 3.95$ time-varying seismicity model over the time-invariant HKJ-ADAPTIVE-M2 model. Given the staggered availability of models over time, information gains are calculated over across six continuous, non-overlapping time windows spanning from: a) August 2007 to January 2009 [69 earthquakes in 47 cells]; b) January 2009 to September 2010 [214 earthquakes in 89 cells]; c) September 2010 to July 2012 [81 earthquakes in 58 cells]; d) July 2012 to October 2012 [27 earthquakes in 15 cells]; e) January 2013 to July 2016 [131 earthquakes in 81 cells]; and f) October 2016 to August 2018 [38 earthquakes in 33 cells]. Within the comparatively smaller STEPJAVA test region (Fig. 1), the corresponding numbers of observed earthquakes and activated cells during the c–f periods are 72 and 49, 24 and 12, 108 and 65, and 33 and 28, respectively. Larger cumulative IGPA values indicate more informative next-day model relative to HKJ-ADAPTIVE-M2. Gray vertical lines indicate the occurrence of target earthquakes within each evaluation period.