Supplementary material to "Select earthquake forecasting models demonstrate consistency with decadal prospective observations in California"

José A. Bayona[1, *], Francesco Serafini[1], Fábio Silva[2], Pablo Iturrieta[3], William H. Savran[4], Marcus Herrmann[5], Warner Marzocchi[5], Philip J. Maechling[2], and Maximilian J. Werner[1]

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

[2]Statewide California Earthquake Center, United States of America.

[3]GFZ German Research Centre for Geosciences, Germany.

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

[5]University of Naples Federico II, Italy.

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

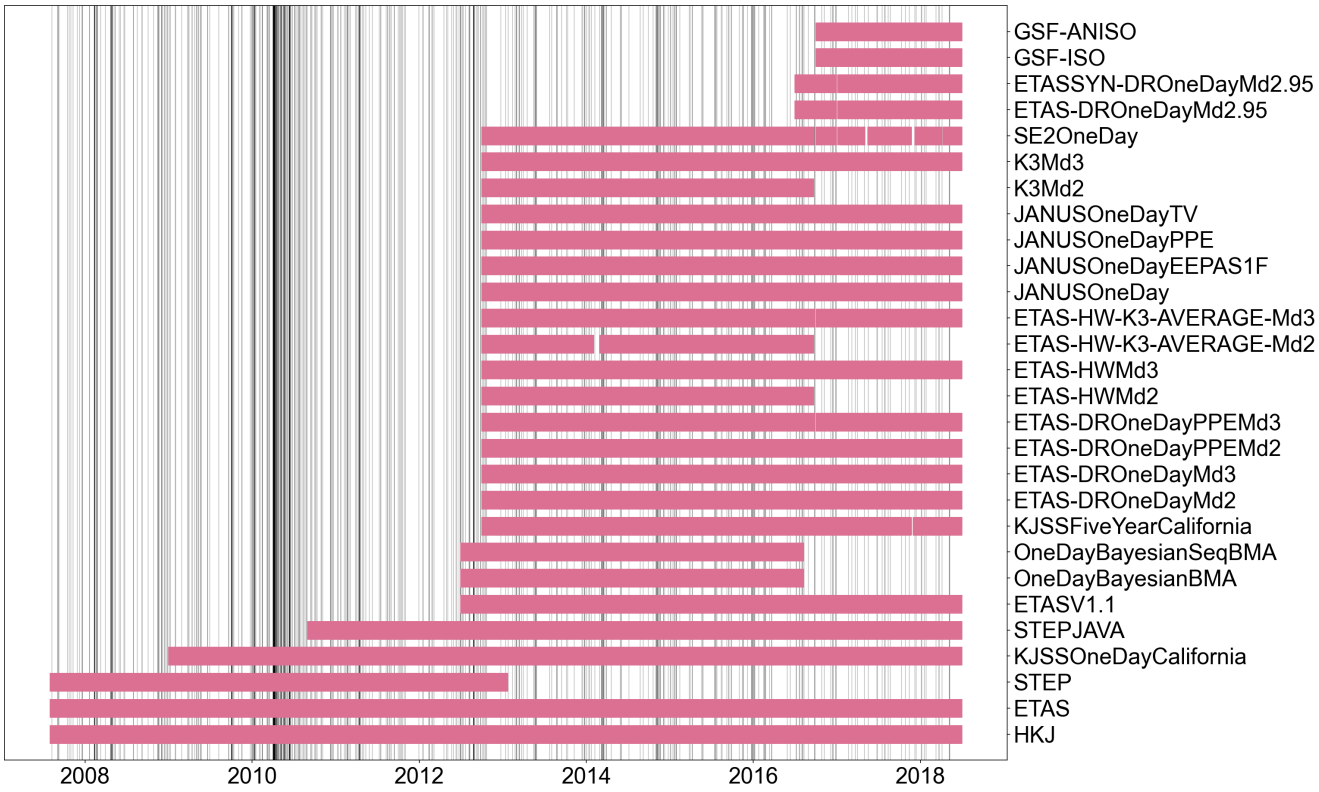

Figure 1. Predictive pool of next-day $M_w \geq 3.95$ earthquake forecasting models for California. The availability of daily forecasts generated by these time-varying models and the time-invariant HKJ benchmark model during the test period (August 1, 2007 to August 30, 2018) is shown in pink. Prospective earthquakes with moment magnitude $M_w \geq 3.95$ and depths ≤ 30 km occurring during this period are shown as vertical gray lines. The number of models increases with time due to new models being developed and submitted to CSEP for prospective testing. Some forecasts have been irretrievably lost due to system processing failures, especially since 2016. E.g., STEP was decommissioned due to software licensing issues. For more information on the exact dates of missing forecasts, see the accompanying data publication[1].

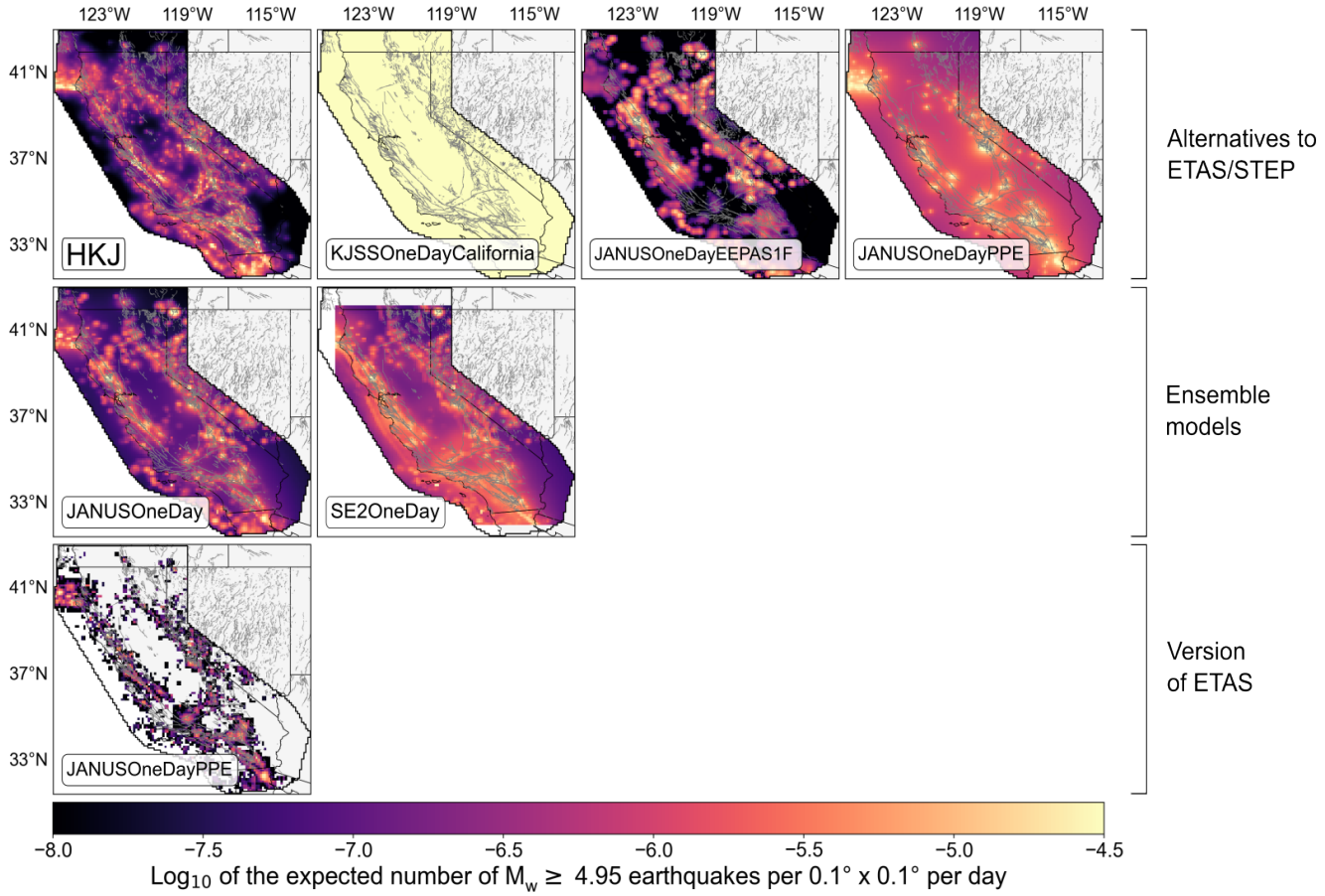

Figure 2. Forecast maps generated by 6 next-day and one time-invariant model (HKJ) $M_w \geq 4.95$ earthquake forecasting models on August 30, 2018, at midnight, for California. This model subset includes a version of ETAS, three alternative models to ETAS and STEP, and two ensemble models. Expected rates of $M_w \geq 4.95$ seismicity are expressed per 0.1° x 0.1° cell per day. Warm colors denote regions where seismic activity is comparatively high, while dark colors indicate comparatively low rates. The HKJ benchmark model is the only model that provides daily earthquake rates that do not change over time. Fault traces (shown in gray) are obtained from the USGS Quaternary Fault and Fold Database (see the Data Availability section).

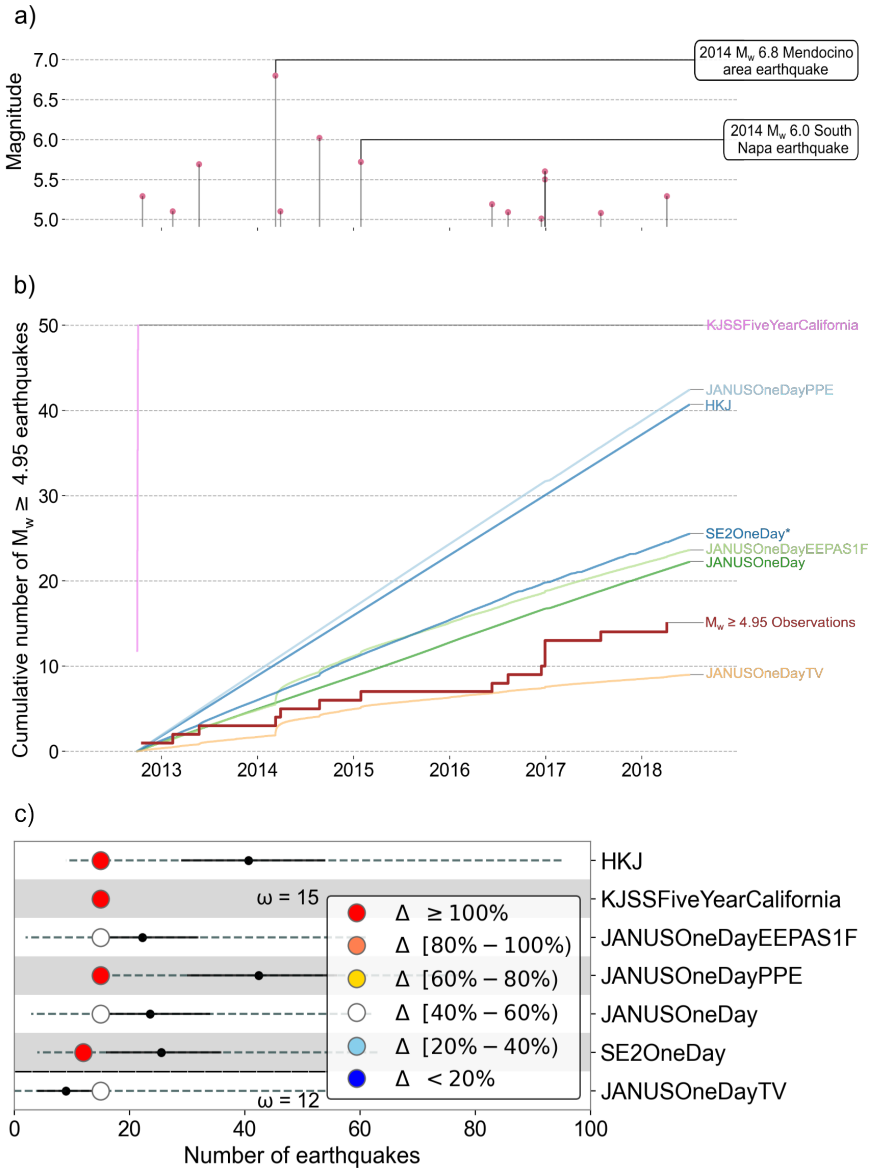

Figure 3. a) Magnitude-time series of observed $M_w \geq 4.95$ earthquakes in California between October 2012 and August 2018. b) Cumulative distributions of observed and forecasted target earthquakes during the prospective evaluation period. The asterisk highlights the SE2OneDay ensemble model, which is defined in a region smaller than the CSEP-California testing region (see Fig. 2) and should therefore not be directly compared to the number of earthquakes observed across the entire CSEP-California test region (brown curve). c) Prospective test results of the cumulative Poisson and Negative Binomial Distribution (NBD) number tests for next-day earthquake forecast models in California. The circles represent the number of observed earthquakes ω, while colors denote their percentage discrepancies (Δ) with the total number of earthquakes predicted by the models. White colors indicate discrepancies between forecasts and observations of less than 60%, and red colors denote discrepancies greater than 100%. Solid black bars and dashed grey bars depict the 95% predictive intervals of the Poisson and Negative Binomial number distributions of the models, respectively, using their predicted number of earthquakes (black dots) as the mean of each distribution. The KJSSFiveYearCalifornia predictive intervals are too large to be shown.

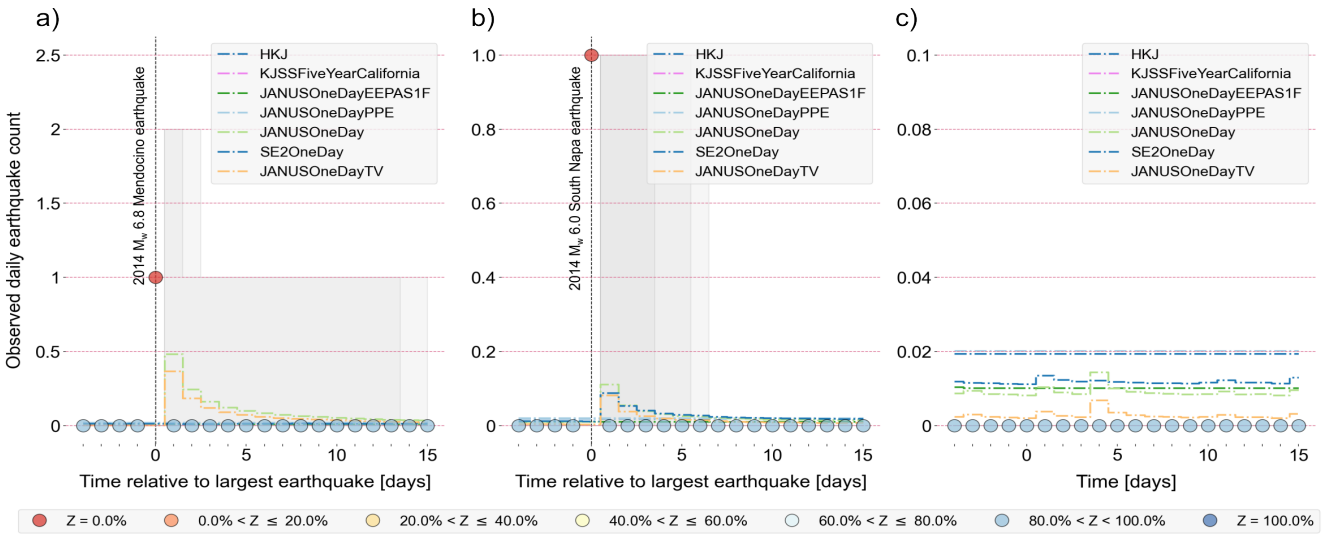

Figure 4. Daily ability of next-day $M_w \geq 4.95$ seismicity models to forecast the observed number of $M_w \geq 4.95$ earthquakes (circles) during the a) 2014 M_w 6.8 Mendocino and b) 2014 M_w 6.0 South Napa earthquake sequences, as well as c) during a randomly selected two-week period in January 2018 of relative seismic calm in California. The color of the circles represents the number N of models (in percentage) that are statistically consistent with daily M_w magnitude 4.95 observed earthquakes. Blue colors indicate that over 60% of the models agree with the observed data. Gray shades denote the 95% predictive intervals of the models. The grayer the shades, the greater the overlap between models' predictive intervals. The KJSSFiveYearCalifornia predictive intervals are too wide to display. Daily expected earthquake rates in c) are so small that the assumed Poisson number distribution are skewed sharply to the right, making the predictive intervals to tend to zero.

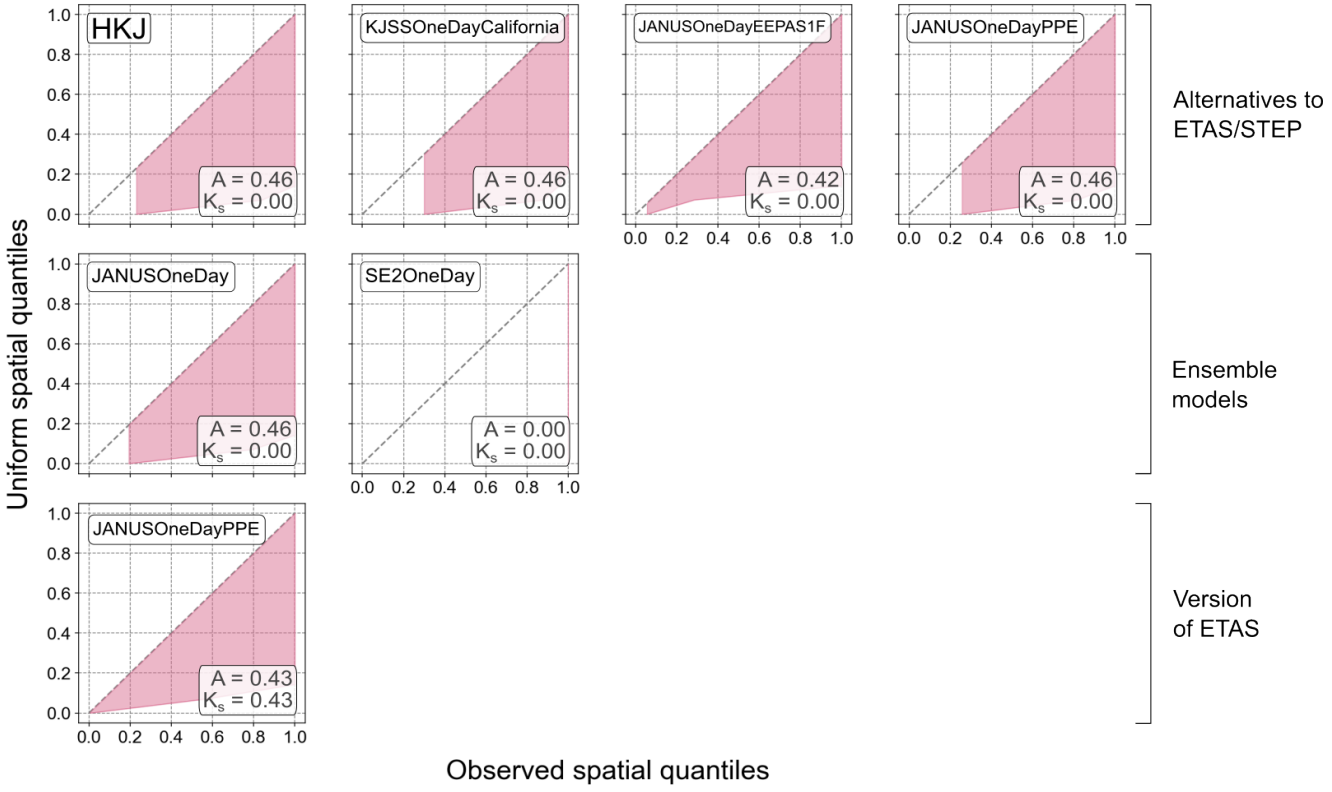

Figure 5. Quantile-vs-quantile (Q-Q) plots comparing the distribution of observed spatial quantiles (in ascending order) with spatial quantiles expected from a uniform distribution. Since we use the $M_w \geq 4.95$ seismicity models as earthquake simulators, the expected and observed spatial quantiles 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. Pink shades denote 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 *p*-values of non-parametric Kolmogorov-Smirnov tests, which assess whether the (continuous) observed quantile distributions are drawn from uniform quantile distributions. $K_s < 0.05$ indicate that spatial forecasts can be considered statistically poorly calibrated.

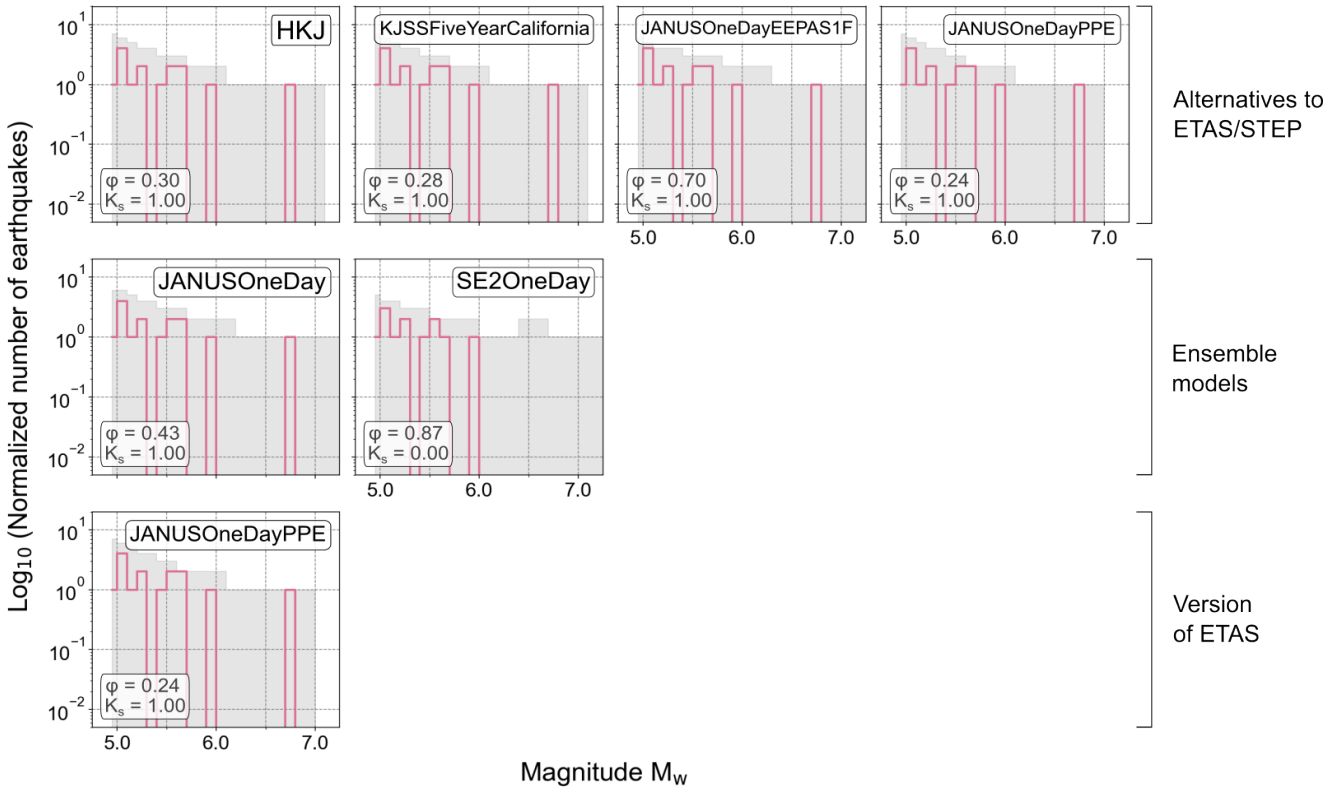

Figure 6. Comparisons between the observed and predicted frequency-magnitude distributions of $M_w \geq 4.95$ earthquakes in California. The *x*-axis shows the earthquake magnitude range divided into 0.1 unit intervals, while the *y*-axis shows the number of observed events within each magnitude bin (pink curves). The 95% model predictive intervals, shown as gray shading, are obtained by normalizing expected earthquake rates to the number of observed earthquakes and assuming a Poisson number distribution. Annotated φ and K_s values are *p*-values of the CSEP magnitude test[2] 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.

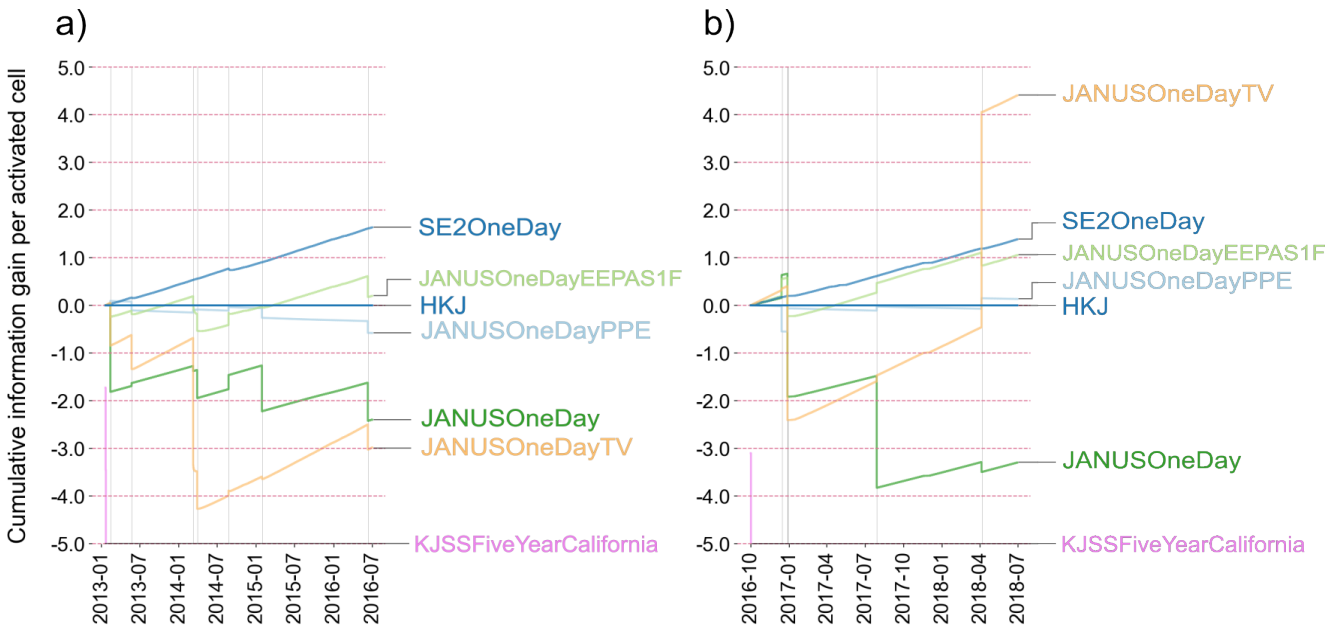

Figure 7. Cumulative binary information gains per activated bin (IGPA) obtained by $M_w \geq 4.95$ time-varying seismicity models over the HKJ time-invariante model between a) January 2013 and July 2016 [7 bins] and b) October 2016 and August 2018 [6 bins]. These gains fluctuate dramatically at the beginning of each evaluation period due to the relatively low number of space-magnitude activated bins. If the cumulative IGPA is positive, the competing model can then be considered more informative than HKJ. Grey vertical lines indicate the occurrence of $M_w \geq 4.95$ earthquakes during each evaluation period.

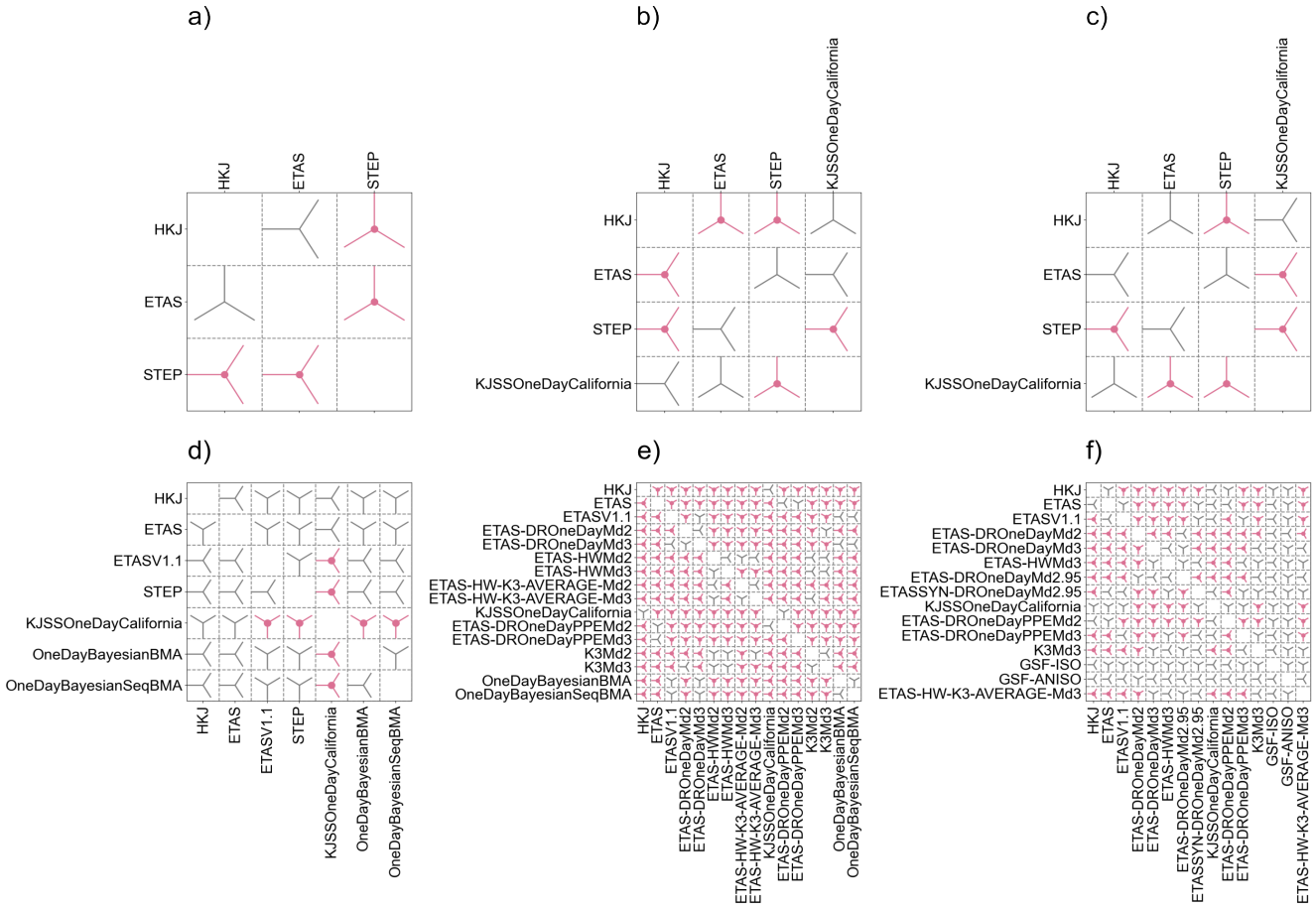

Figure 8. Confusion matrices showing the significance of the cumulative information gain per activated spatial forecast cell obtained by competing $M_w \geq 3.95$ seismicity models (rows) over reference models (columns) between a) August 2007 and January 2009 [69 earthquakes; 47 activated cells], b) January 2009 and September 2010 [214 earthquakes; 89 cells], c) September 2010 and July 2012 [81 earthquakes; 58 cells], d) July 2012 and October 2012 [27 earthquakes; 15 cells], e) January 2013 and July 2016 [131 earthquakes; 81 cells], and f) October 2016 and August 2018 [38 earthquakes; 33 cells]. For each pairwise comparison, the aircraft symbol points to the most informative model. Pink aircraft symbols with midpoints indicate that the IGPA is statistically significant, i.e., the *p*-value of the comparative t-test is smaller than 0.05, while gray aircraft symbols denote the opposite. In the main text, we only discuss IGPAs obtained by next-day models over the time-invariant HKJ model (first column). Note that all matrices are symmetric with respect to the diagonal.

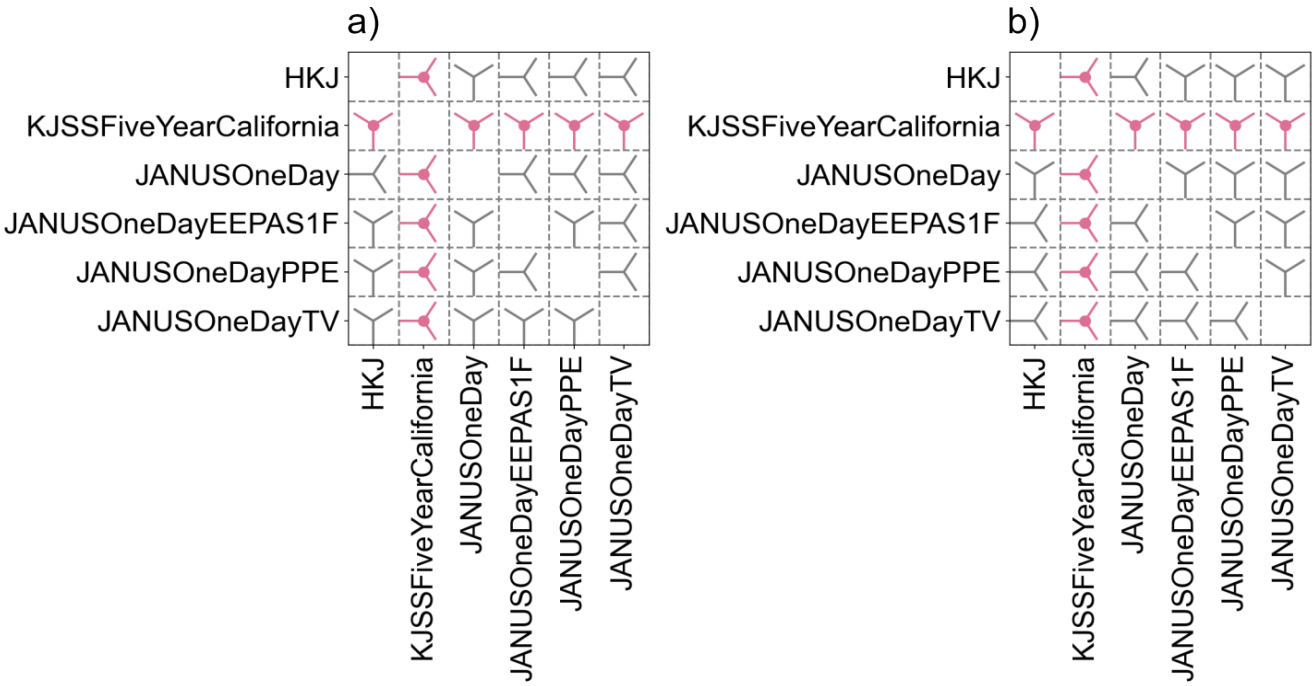

Figure 9. Confusion matrices showing the significance of the cumulative IGPAs scores obtained by competing $M_w \geq 4.95$ seismicity models over reference models between a) January 2013 and July 2016 [7 earthquakes; 7 cells] and b) October 2016 and August 2018 [6 earthquakes; 4 cells]. For each pairwise comparison, the aircraft symbol points to the most informative model. Pink aircraft symbols with midpoints indicate that the IGPA is statistically significant, while gray aircraft symbols indicate the opposite. Note that both matrices are symmetric with respect to the diagonal.

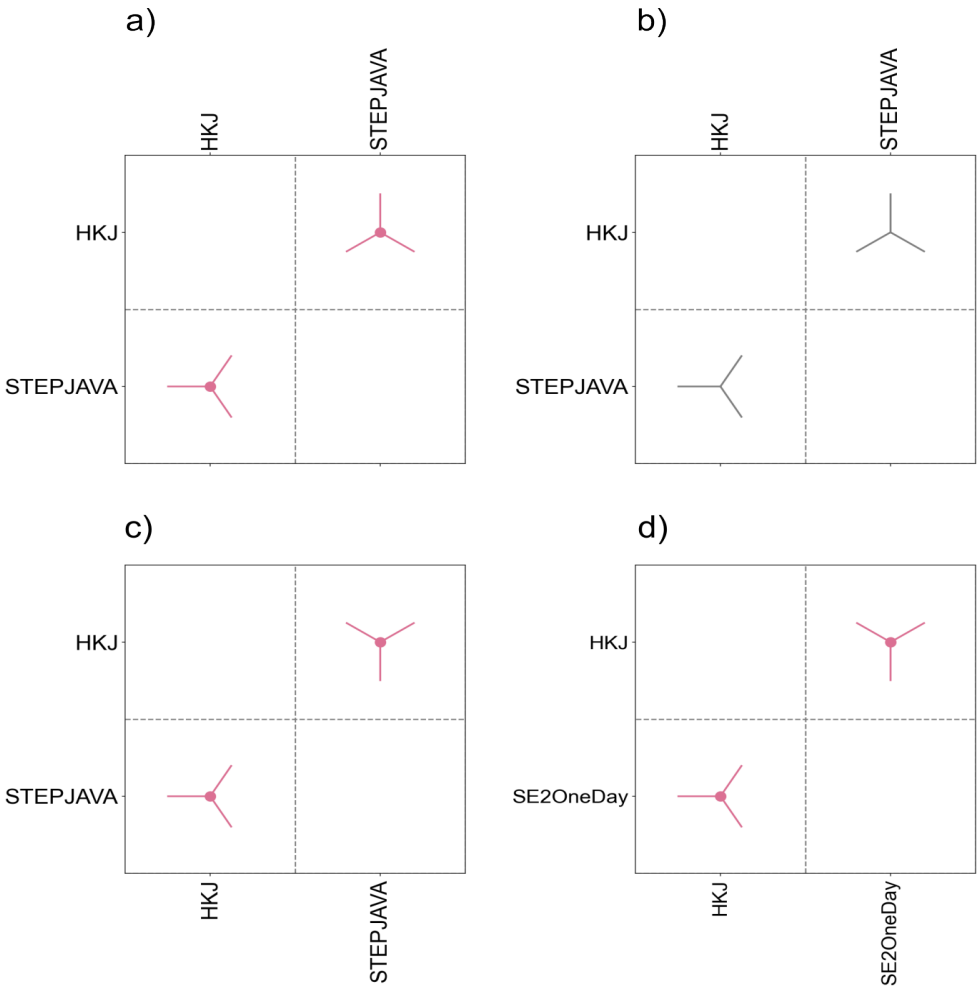

Figure 10. Confusion matrices showing the significance of the IGPA scores obtained by the STEPJAVA and SE2OneDay models over HKJ. We present these results separately, because STEPJAVA and SE2OneDay are defined within a test region smaller than the entire CSEP-California test region. In the case of the $M_w \geq 3.95$ STEPJAVA seismicity model, the significance of its cumulative IGPAs over HKJ are estimated between a) September 2010 and July 2012 [72 earthquakes; 49 cells], b) July 2012 and October 2012 [24 earthquakes; 12 cells], and c) January 2013 and July 2016 [108 earthquakes; 65 cells]. In the case of the $M_w \geq 4.95$ SE2OneDay seismicity model, such a statistical significance is estimated between d) October 2016 and August 2018 [5 earthquakes; 3 cells]. For each pairwise comparison, the aircraft symbol points to the most informative model. Pink aircraft symbols with midpoints indicate that the IGPA is statistically significant, while gray aircraft symbols indicate the opposite. In the main text, we only discuss IGPAs obtained by next-day models over the time-invariant HKJ model (first column). Note that all matrices are symmetric with respect to the diagonal.

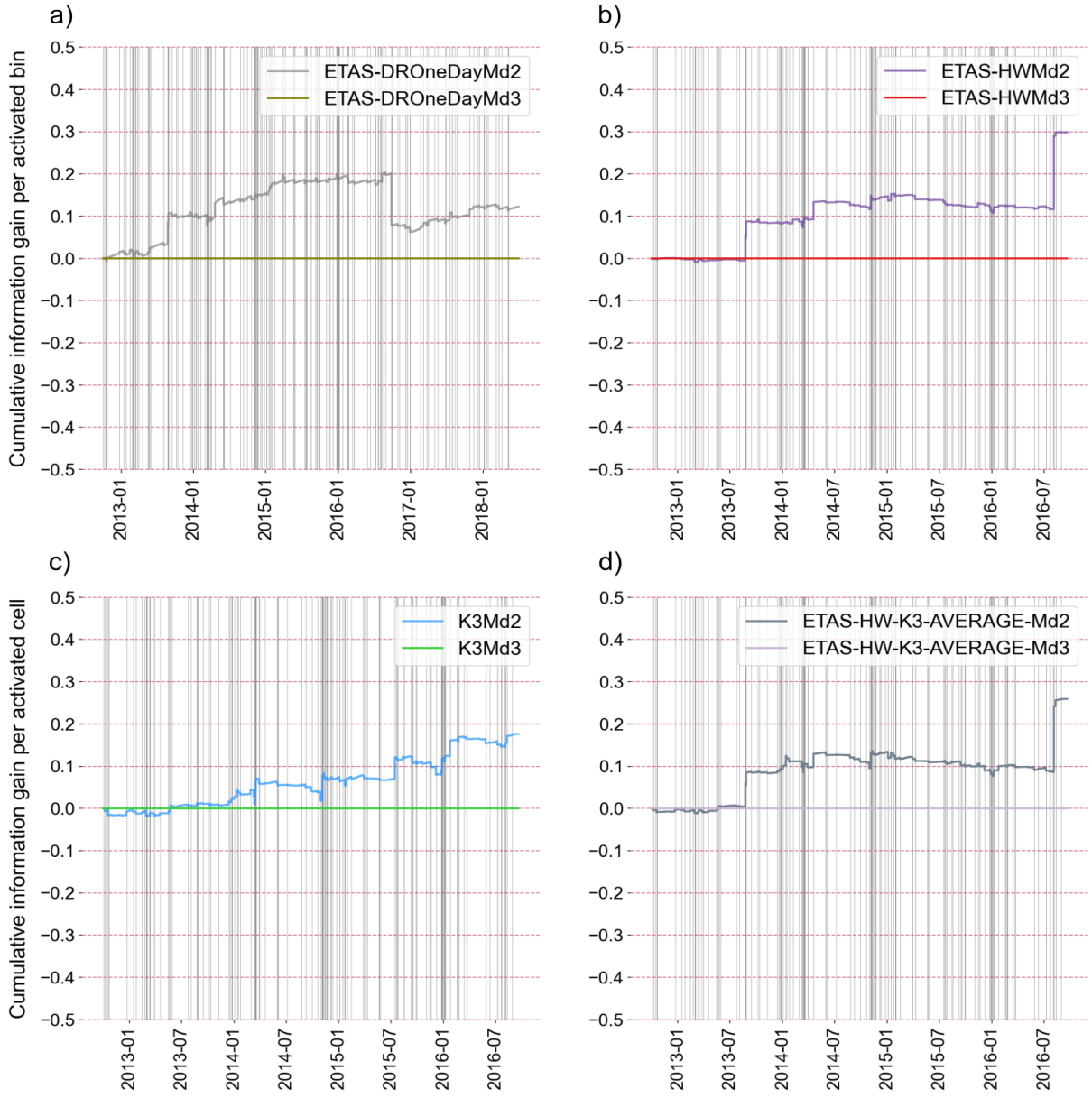

Figure 11. Cumulative binary IGPAs obtained by time-varying $M_w \geq 3.95$ seismicity models calibrated on $M \geq 2.0$ seismicity over their analog models based on larger ($M \geq 3.0$) earthquakes during a) October 2012 and July 2018 [188 earthquakes; 119 cells] and b–d) October 2012 and September 2016 [147 earthquakes; 92 cells]. Positive cumulative IGPAs indicate that the competing models are statistically more informative than the reference models. However, these gains are statistically insignificant as the *p*-values of the underlying t-tests are smaller than 0.05. Gray vertical lines indicate the occurrence of $M_w \geq 4.95$ earthquakes during each evaluation period.

References

1. 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).

2. 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).