Deliverable

D7.6 Report presenting first results of prospective studies

Deliverable information	
Work package	WP7 Testing
Lead	University of Bristol
Authors	Max Werner, Bristol
Reviewers	Danijel Schorlemmer, GFZ
Approval	[Management Board]
Status	Final
Dissemination level	Public
Delivery deadline	28.02.2023
Submission date	30.05.2023
Intranet path	[DOCUMENTS/DELIVERABLES/File Name]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821115

Table of contents

Summary

This report summarises prospective testing activities of the Collaboratory for the Study of Earthquake Predictability (CSEP)/RISE working group and first results of prospective tests. Prospective evaluations include: (i) a Coulomb stress model and the (USGS) UCERF3-ETAS (Uniform California Earthquake Rupture Forecast version 3, Epidemic-type Aftershock) model during the 2019 Ridgecrest, California, earthquake sequence, (ii) the seismic gap hypothesis in Guerrero, Mexico, (iii) hybrid models versus individual models, (iv) global versus regional models, and (v) 10-year forecasts in Italy. In addition, we summarise the community-based open-source development of CSEP toolkits and applications to modernize CSEP infrastructure towards open and reproducible experiments (i.e., the pyCSEP and floatCSEP packages). Finally, we report on the launch of a new experiment in Italy, which includes an open call for model submission from outside the RISE collaboration.

1. Pseudo-Prospective Evaluation of the Coulomb Stress Model during the 2019 Ridgecrest, California, Earthquake Sequence

Operational earthquake forecasting protocols commonly use statistical models for their recognized ease of implementation and robustness in describing the short-term spatiotemporal patterns of triggered seismicity. However, recent advances on physics-based aftershock forecasting reveal comparable performance to the standard statistical counterparts with significantly improved predictive skills when fault and stress-field heterogeneities are considered. Mancini et al. (2020) performed a pseudo-prospective forecasting experiment during the first month of the 2019 Ridgecrest (California) earthquake sequence. They developed seven Coulomb rate-and-state models that couple static stress-change estimates with continuum mechanics expressed by the rate-and-state friction laws (Figure 1). The model parameterization supports a gradually increasing complexity; they start from a preliminary model implementation with simplified slip distributions and spatially homogeneous receiver faults to reach an enhanced one featuring optimized fault constitutive parameters, finite-fault slip models, secondary triggering effects, and spatially heterogenous planes informed by pre-existing ruptures. The data-rich environment of southern California allows them to test whether incorporating data collected in near-real time during an unfolding earthquake sequence boosts our predictive power. They assess the absolute and relative performance of the forecasts by means of statistical tests used within the CSEP and compare their skills against a standard benchmark epidemic-type aftershock sequence (ETAS) model for the short (24 hr after the two Ridgecrest mainshocks) and intermediate terms (one month). Stress-based forecasts expect heightened rates along the whole near-fault region and increased expected seismicity rates in the central Garlock fault. Their comparative model evaluation not only supports that faulting heterogeneities coupled with secondary triggering effects are the most critical success components behind physics-based forecasts, but also underlines the importance of model updates incorporating near-realtime available aftershock data reaching better performance than standard ETAS. In their publication, Mancini et al. explore the physical basis behind the results by investigating the localized shut down of pre-existing normal faults in the Ridgecrest near-source area.

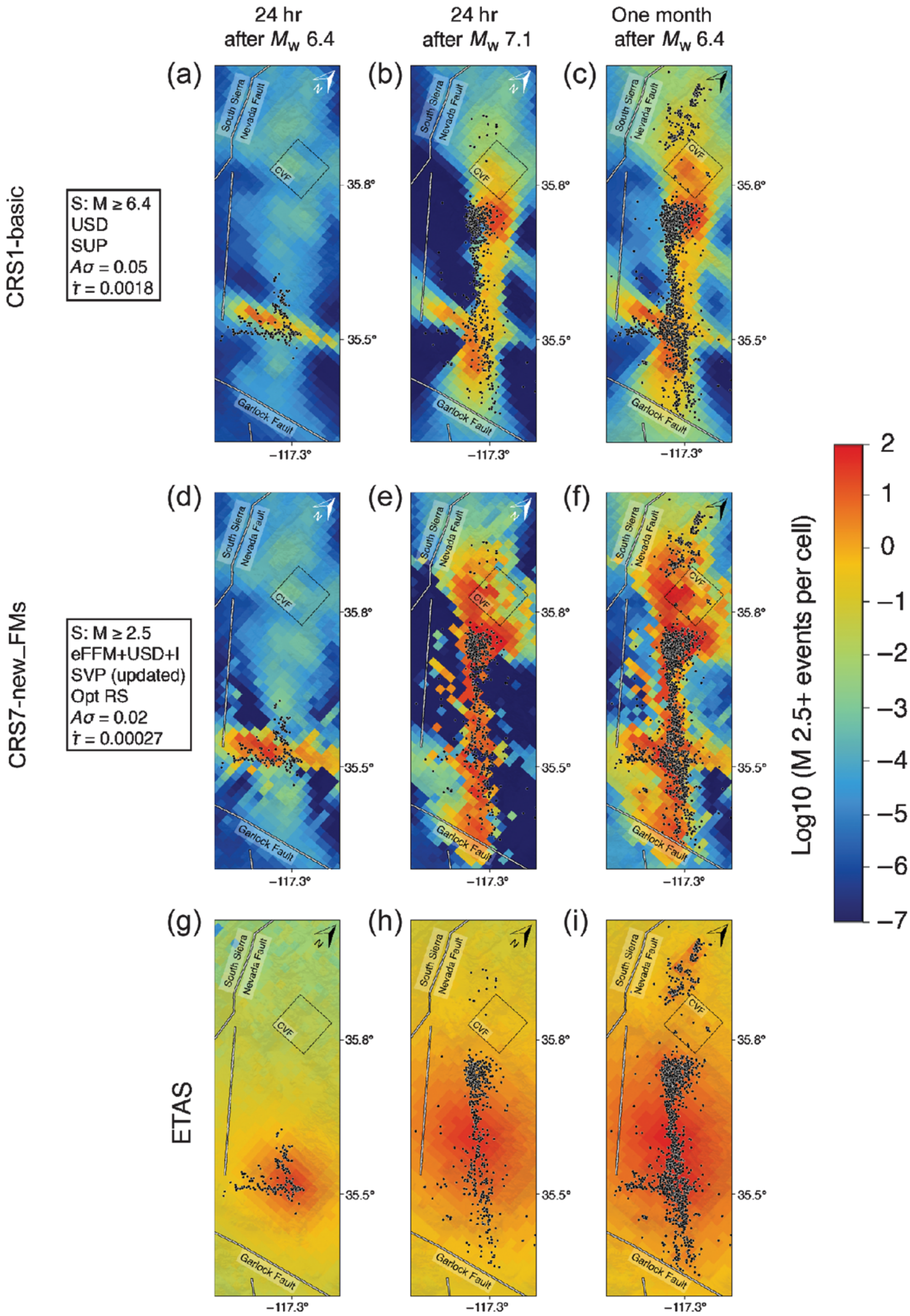

Figure 1: **Maps of expected seismicity rates for (a–c) CRS1, (d–f) CRS7, and (g–i) ETAS in the area of main aftershock productivity for the first 24 hr following the two mainshocks and for the first month of the 2019 Ridgecrest sequence. Observed events in each time window are represented as circles. The dashed-line square indicates the area of the Coso volcanic field (CVF). eFFM, edited finite-fault slip model; I, isotropic stress field; Opt RS, optimized rate-and-state parameters; S, sources (minimum magnitude); SUP, spatially uniform receiver planes; SVP, spatially variable planes; USD, uniform slip distribution. Aσ values are in MPa and $\dot{\tau}$ values are in MPa/yr [from Mancini et al., 2020].**

Publication outputs:

Mancini, Segou, Werner & Parsons (2020): The Predictive Skills of Elastic Coulomb Rate-and-State Aftershock Forecasts during the 2019 Ridgecrest, California, Earthquake Sequence. Bulletin of the Seismological Society of America 110 (4): 1736–1751. https://doi.org/10.1785/0120200028

2. Pseudo-Prospective Evaluation of the UCERF3-ETAS Model During the 2019 Ridgecrest, California, Earthquake Sequence

Savran et al. (2020) developed a tailored pseudo-prospective experiment to test the hypothesis that large supra-seismogenic aftershocks occur on (mapped) faults and control the overall aftershock patterns. The 2019 Ridgecrest sequence provided the first opportunity to evaluate the Uniform California Earthquake Rupture Forecast v.3 with epidemic-type aftershock sequences (UCERF3-ETAS) in a pseudo-prospective sense. For comparison, they include a version of the model without explicit faults more closely mimicking traditional ETAS models (UCERF3-NoFaults). They evaluate the forecasts with new metrics developed within CSEP. The metrics consider synthetic catalogs simulated by the models rather than synoptic probability maps, thereby relaxing the Poisson assumption of previous CSEP tests. Their approach compares statistics from the synthetic catalogs directly against observations, providing a flexible approach that can account for dependencies and uncertainties encoded in the models. RISE developed bespoke software and testing methods that enabled comparing the simulated catalogs with the observed catalog, thereby circumventing the need for approximating likelihood functions. The testing methods were developed as part of the open-source community software toolkit PyCSEP (see deliverable D7.1), which is available from https://github.com/SCECcode/pycsep.

Savran et al. (2020) find that, to the first order, both UCERF3-ETAS and UCERF3-NoFaults approximately capture the spatio-temporal evolution of the Ridgecrest sequence, adding to the growing body of evidence that ETAS models can be informative forecasting tools. However, they also find that both models mildly overpredict the seismicity rate, on average, aggregated over the evaluation period. More severe testing indicates the overpredictions occur too often for observations to be statistically indistinguishable from the model. Magnitude tests indicate that the models do not include enough variability in forecasted magnitude-number distributions to match the data. Spatial tests highlight discrepancies between the forecasts and observations, but the greatest differences between the two models appear when aftershocks occur on modeled UCERF3-ETAS faults. Therefore, any predictability associated with embedding earthquake triggering on the (modeled) fault network may only crystalize during the presumably rare sequences with aftershocks on these faults. Accounting for uncertainty in the model parameters could improve test results during future experiments.

Publication outputs:

Savran, W. H., Werner, M. J., Marzocchi, W., Rhoades, D. A., Jackson, D. D., Milner, K., Field, E. H. & Michael, A. (2020). Pseudoprospective Evaluation of UCERF3-ETAS Forecasts during the 2019 Ridgecrest Sequence. Bulletin of the Seismological Society of America, 110(4), 1799-1817. https://doi.org/10.1785/0120200026

3. A Prospective Test of the Seismic Gap Hypothesis in the Guerrero Gap, Mexico

The seismic gap hypothesis has a long and controversial history, but continues to be popular and is frequently cited in the media. In particular, the seismic gap hypothesis has been widely cited in Mexico to predict the location of future earthquakes and to assess seismic hazard, specifically in the context of the so-called 'Guerrero gap' (Figure 2). However, no analysis of the outcome of any predictions of the hypothesis in Mexico has been done to-date. Husker et al. (2023) analyzed the outcome of the formal seismic gap prediction by Nishenko and Singh (1987). The prediction has well-defined probabilities, areas and timeframes that allow for its evaluation. Those timeframes were 5 years, 10 years and 20 years after 1986. The prediction relies on the precise repeat times of characteristic earthquakes to define segments, but the catalog that the authors use relies on an imprecise definition of characteristic earthquakes. Husker et al. discuss some of their decisions in building their catalog to explain how they analyze the outcome of the prediction. They create

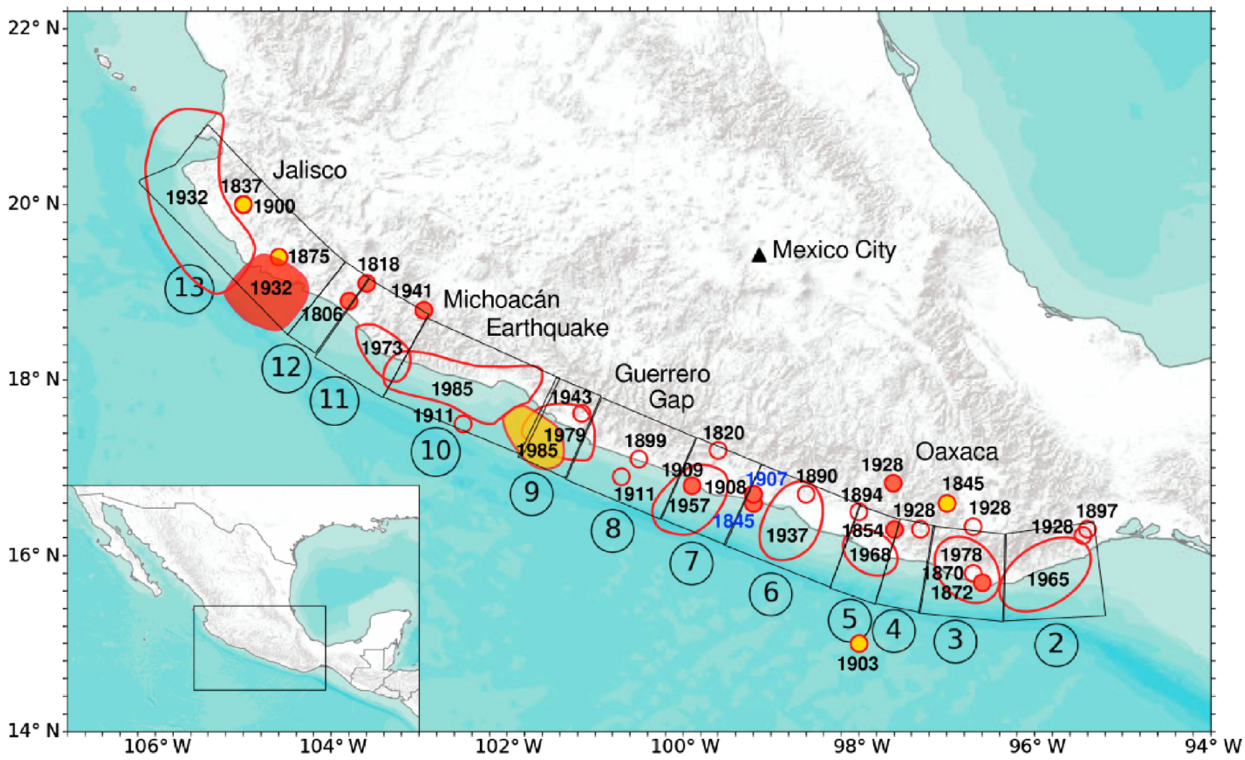

Figure 2: Earthquakes Ms≥7.4 are shown with segments according to Nishenko & Sykes (1987) [NS1987]. Segment 1 extends to the east but has never had an M 7+ earthquake recorded at the plate interface and so was not evaluated by NS1987. We do not evaluate it here. Earthquakes listed in Singh et al. (1981, 1984) that lie in and around the segments and are shallow, but not included in NS1987, are shown in yellow. The red shading indicates that the earthquake's epicenter is outside of the longitude assigned to the segment. The blue lettering indicates two different earthquakes that span more than 1 segment and yet were considered by NS1987 to be characteristic in each segment. All the epicenters come from Singh et al. (1981, 1984). Earthquake rupture outlines are available for most earthquakes that occurred after 1930. The inset shows the location of this area [from Husker et al., 2023].

catalogs of earthquakes based on the probabilities of earthquake occurrence for each segment. They also generate null model earthquake catalogs using the average number of earthquakes that occur in the subduction zone, and randomly distribute these along the distance of the segments. They find that null model performed better than the seismic gap hypothesis prediction. The prediction over the longest time frame of 20 years correctly predicted the outcome in only 48% of the segments compared to 91% coinciding for the null model. The gap hypothesis also greatly over predicted the total number of segments with a characteristic earthquake. Ms ≥ 7.4 earthquakes were predicted to occur in 6 of the 11 segments over the 20-year timeframe, but only 1 actually occurred. That lone earthquake was an Mw 8.0 which occurred in a segment with a 0% chance of an earthquake in one of their models and 16% change in another. Husker et al. conclude that the gap hypothesis did not perform well at predicting earthquakes in Mexico and, in fact, its predictions were worse than predicting earthquakes by chance. There is thus no evidence to suggest earthquakes are overdue in the Guerrero gap, and therefore Husker et al. recommend taking special care in invoking the gap hypothesis to communicate earthquake hazards in Mexico.

Publication Outputs:

Husker, Bayona, Werner, Santoyo & Corona-Fernandez (2023): A Test of the Earthquake Gap Hypothesis in Mexico: the case of the Guerrero Gap. Bulletin of the Seismological Society of America 113 (1), 468-479. https://doi.org/10.1785/0120220094

4. Do Hybrid Models Achieve Greater Predictive Skill than Individual Models?

The Regional Earthquake Likelihood Models (RELM) experiment, conducted within CSEP, showed that the smoothed seismicity (HKJ) model by Helmstetter et al. (2007) was the most informative time-independent earthquake model in California during the 2006–2010 evaluation period. The diversity of competing forecast hypotheses and geophysical datasets used in RELM was suitable for combining multiple models that could provide more informative earthquake forecasts than HKJ. Thus, Rhoades et al. (2014) created multiplicative hybrid models that involve the HKJ model as a baseline and one or more conjugate models. In retrospective evaluations, some hybrid models showed significant information gains over the HKJ forecast. Bayona et al. (2022) prospectively assess the predictive skills of 16 hybrids and 6 original RELM forecasts at a 0.05 significance level, using a suite of traditional and new CSEP tests that rely on a Poisson and a binary likelihood function. In addition, they include consistency test results at a Bonferroni-adjusted significance level of 0.025 to address the problem of multiple tests. Furthermore, they compare the performance of each forecast to that of HKJ (Figure 3). The evaluation dataset contains 40 target events recorded within the CSEP California testing region from 1 January 2011 to 31 December 2020, including the 2016 Hawthorne earthquake swarm in southwestern Nevada and the 2019 Ridgecrest sequence. Consistency test results show that most forecasting models overestimate the number of earthquakes and struggle to explain the spatial distribution of epicenters, especially in the case of seismicity clusters. The binary likelihood function significantly reduces the sensitivity of spatial log-likelihood scores to clustering, however; most models still fail to adequately describe spatial earthquake patterns. Contrary to retrospective analyses, our prospective test results show that none of the models are significantly more informative than the HKJ benchmark forecast, which they interpret to be due to temporal instabilities in the fit that forms hybrids. These results suggest that smoothing high-resolution, small earthquake data remains a robust method for forecasting moderate-to-large earthquakes over a period of five to fifteen years in California.

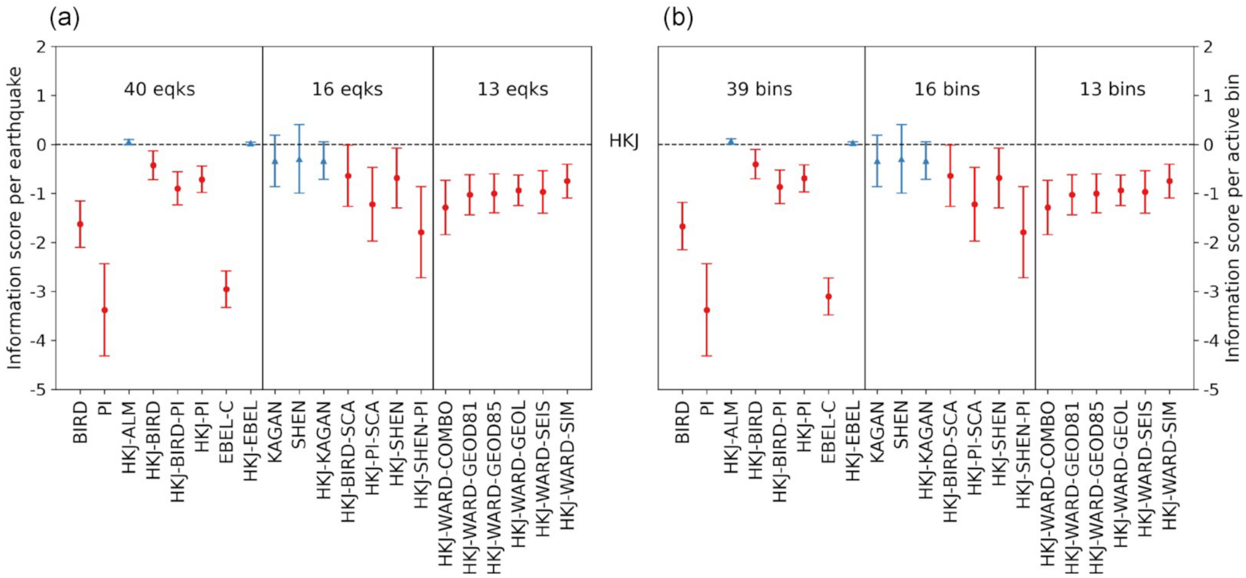

Figure 3: Comparison of information gains T-test results. Information gains per (a) earthquake and (b) active bin are presented as triangles and circles, depending on their relative values to that of the benchmark Helmstter et al. (2007) HKJ model: blue triangles denote that the model is statistically as informative as HKJ and red circles indicate that a model is significantly less informative than HKJ. Hybrid models are generally less informative than the benchmark, despite their promising performance in retrospective studies (that included a penalty term to account for free parameters) [from Bayona et al., 2022].

Publication outputs:

Bayona, Savran, Rhoades & Werner (2022), Prospective evaluation of multiplicative hybrid earthquake forecasting models in California, Geophysical Journal International, ggac018, *https://doi.org/10.1093/gji/ggac018*

5. New Community-Based Prospective CSEP Testing Infrastructure (pyCSEP and floatCSEP)

In collaboration with the Southern California Earthquake Center (SCEC), we have overhauled the CSEP design for prospective testing. The previous testing centers with hardware and monolithic software for conducting blind and automated forecast generation and testing created a substantial maintenance burden in the form of (i) failed testing center workflows that required lengthy reprocessing, (ii) systems administration to maintain three servers and (iii) personnel risk (a single individual had developed the software).

We redesigned CSEP in a modular and open fashion, such that a Python toolkit (pyCSEP) provides the evaluation functionality, and an application (floatCSEP) provides the configuration for running open and reproducible experiments. pyCSEP has been code-reviewed and published, with extensive documentation and online tutorials. Both RISE and SCEC have provided online training of pyCSEP. floatCSEP has completed development and a publication is in development (Savran et al,. 2023, in preparation).

pyCSEP (https://docs.cseptesting.org/) is a Python package to help earthquake forecast developers embed model evaluation into the model development process. The package contains the following modules: (1) earthquake catalog access and processing, (2) data models for earthquake forecasts, (3) statistical tests for evaluating earthquake forecasts, and (4) visualization routines. pyCSEP can evaluate earthquake forecasts expressed as expected rates in space-magnitude bins, and simulation-based forecasts that produce thousands of synthetic seismicity catalogs. Most importantly, pyCSEP contains community-endorsed implementations of statistical tests to evaluate earthquake forecasts, and provides well defined file formats and standards to facilitate model comparisons. The toolkit will facilitate integrating new forecasting models into testing centers, which evaluate forecast models and prediction algorithms in an automated, prospective and independent manner, forming a critical step towards reliable operational earthquake forecasting.

Modernizing CSEP experiments also involved developing an open experiment format that decentralizes the testing process and promotes best-practices in open science software development and data management (using an application called floatCSEP). Savran et al. (2023, in preparation) demonstrate the open experiment format using the Global Earthquake Forecasting Experiment (GEFE). The GEFE addresses the comparability and the stability of test results on quad-tree grids, which provide a significant computational improvement to evaluating global forecasting models.
Additionally, the open experiment format can be the basis for future experiments developed by CSEP and independent researchers.

pyCSEP and floatCSEP are now being used and extended for a new phase of the prospective Italian CSEP experiment (see Section 8).

Publication Outputs:

Savran, Werner, Maechling, Schorlemmer (2022). pyCSEP: A Python Toolkit For Earthquake Forecast Developers. Journal of Open Source Software, 7(69), 3658, https://doi.org/10.21105/joss.03658

Savran, Bayona, Iturrieta, Khawaja, Bao, Bayliss, Herrmann, Schorlemmer, Maechling & Werner (2022), pyCSEP: A Python Toolkit for Earthquake Forecast Developers, Seismological Research Letters, 93 (5), 2858–2870, https://doi.org/10.1785/0220220033

Savran et al. (2023, in preparation), *Modernizing CSEP Earthquake Forecasting Experiments: The Floating Testing Center*.

6. Are regionally calibrated seismicity models more informative than global models?

Earthquake forecasting models express hypotheses about seismogenesis that underpin global and regional probabilistic seismic hazard assessments (PSHAs). An implicit assumption is that the comparatively higher spatiotemporal resolution datasets from which regional models are generated lead to more informative seismicity forecasts than global models, which are however calibrated on greater datasets of large earthquakes. Bayona et al. (2023) prospectively assessed the ability of the Global Earthquake Activity Rate (GEAR1) model and 19 time-independent regional models to forecast *M* 4.95+ seismicity in California, New Zealand, and Italy from 2014 through 2021, using metrics developed by CSEP. Their results (Figure 4) show that regional models that adaptively smooth small earthquake locations perform best in California and Italy during the evaluation period; however, GEAR1, based on global seismicity and geodesy datasets, performs surprisingly well across all testing regions, ranking first in New Zealand, second in California, and third in Italy. Furthermore, the performance of the models is highly sensitive to spatial smoothing, and the optimal smoothing likely depends on the regional tectonic setting. Acknowledging the limited prospective test data, these results provide preliminary support for using GEAR1 as a global reference *M* 4.95+ seismicity model that could inform eight-year regional and global PSHAs.

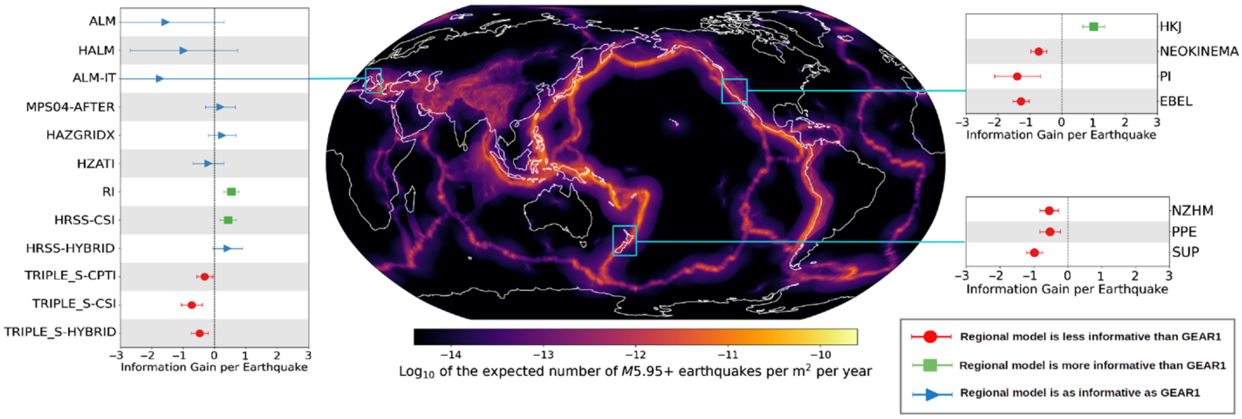

Figure 4: Information gains of regional models over the global GEAR1 model in Italy, California and New Zealand [from Bayona et al., 2023].

Publication Outputs:

Bayona, Savran, Iturrieta, Gerstenberger, Graham, Marzocchi, Schorlemmer & Werner (2023), Are Regionally Calibrated Seismicity Models More Informative than Global Models? Insights from California, New Zealand, and Italy, The Seismic Record, 3 (2): 86–95, https://doi.org/10.1785/0320230006

7. Prospective Evaluation of a decade-long earthquake forecasting experiment in Italy

In 2010, a 10-year CSEP forecasting experiment began in Italy, where modelers collectively agreed on authoritative data sources, testing rules, and formats to independently evaluate a collection of forecasting models submitted by multiple researchers. Iturrieta et al. (2023, in preparation) test the submitted time-independent forecasts with ten years of fully prospective data using a multi-score approach to (i) study the model features that cause a forecast to be consistent or inconsistent with the observations, (ii) evaluate the experiment's results' stability over time, and (iii) quantify the spatial limitations of Poisson models to provide forecasts consistent with the observed seismicity, by using spatial-statistics metrics. Their results (Figure 5) show that the best-performing models use catalogs that span over 100 years in time and incorporate fault information, indicating that these data types should not be overlooked in the future. The experiment's results are stable over time in terms of the models ranking, suggesting a 10-year window is a first order approximation to discriminate between optimal and sub-optimal forecasts. Finally, as expected from Poisson models, no forecast can reproduce earthquakes' spatial clustering. However, the spatial metrics they used, namely Ripley-K and pair correlation functions, can quantify the degree of dissimilarity. The results show that the two best performing models under standard metrics perform spatially better than others, if the correlation between events is analyzed. These spatial metrics could supplement standard tests for model calibration, selection or rejection.

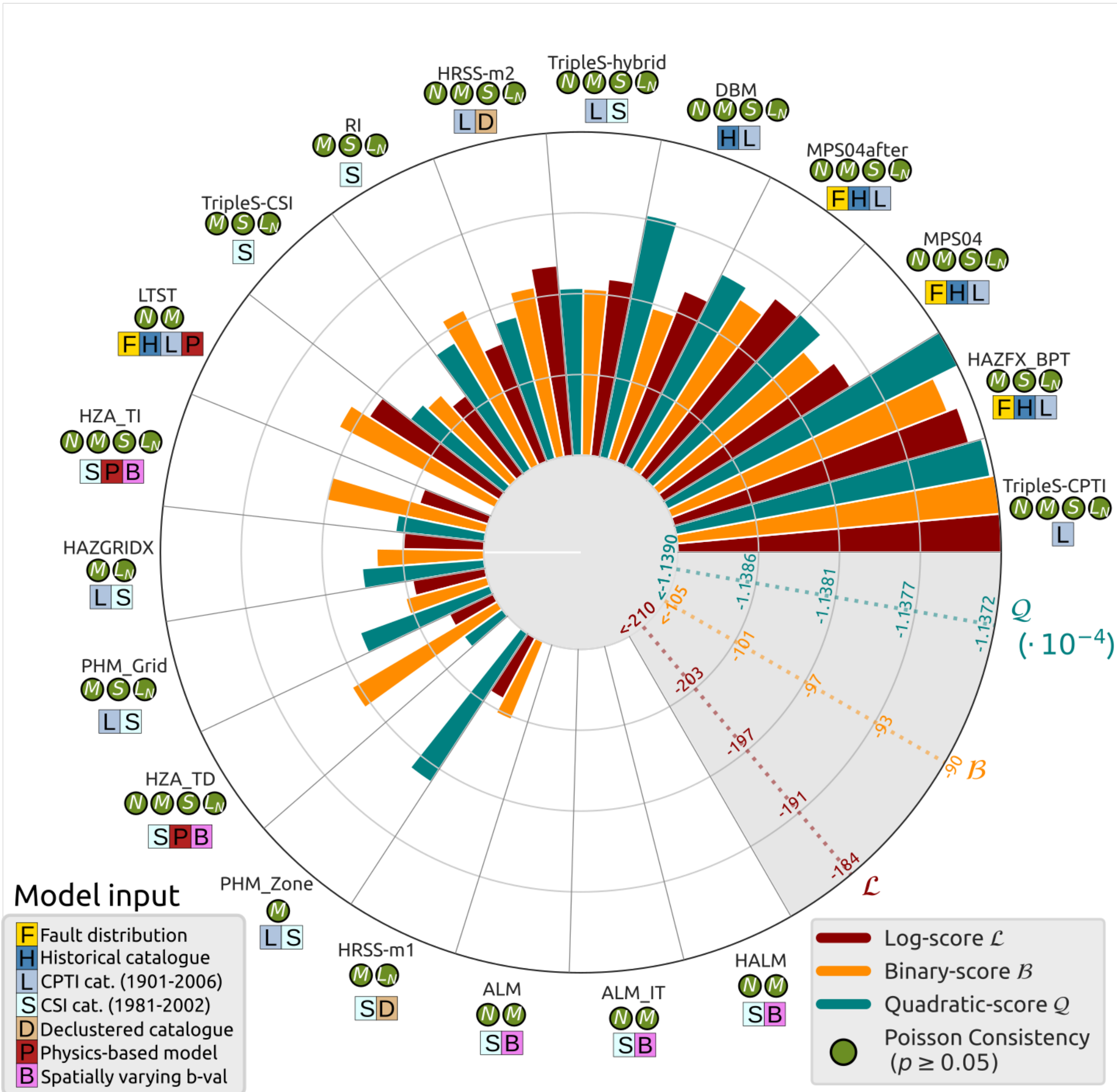

Figure 5: Results from the 10-year prospective Italy experiment. Snail-chart showing the qualitative ranking of the competing models using multiple scores: Log-score, Quadratic(Brier)-Score and the Binary-score. The forecasts are ordered by their ranking of all metrics normalized and averaged. On top of each model name, a green circle is placed if they passed the N-, S-, M- and CL-consistency tests. A qualitative categorization of models according to their components is shown beneath their names [from Iturrieta et al., 2023, in preparation].

Publication Outputs:

Iturrieta et al (2023, in preparation), Evaluating a decade-long earthquake forecasting experiment in Italy: Analysis of model features through a multi-score approach.

8. Launch of a New Phase of the Italian Forecast Experiment

The CSEP group (led by P. Iturrieta) has now finalized all software development, experiment design and modeller instruction materials for a new phase of the Italian forecast experiment (see deliverable D8.7), and the call for model participation has now been made, including to modellers outside of the RISE project. The new phase expands the model class by requiring only sets of simulated

catalogs as forecasts, rather than forecasts specified as Poisson probability maps. The procedures for assessing forecasts specified as simulated catalogs was developed by Savran et al. (2020) [see above], the infrastructure for hosting the experiment in an open manner uses pyCSEP (Savran et al., 2022a,b) and the new application floatCSEP (Savran et al., 2023, in preparation), which was extended to accommodate time-dependent forecast experiments (Iturrieta et al., 2023, in preparation). Modellers can find all relevant information and materials at the following site:
https://git.gfz-potsdam.de/csep-group/rise_italy_experiment/experiment_setup/

Several new models have already been submitted and are undergoing dockerization and backtesting. We anticipate writing an overview paper that describes the new experiment design and the participating model (at the time of writing, as the call for new models will remain open).

Publication Outputs:

Savran et al. (2023, in preparation), Modernizing CSEP Earthquake Forecasting Experiments: The Floating Testing Center.

Iturrieta et al (2023, in preparation), A new CSEP Experiment for Italy.