Khawaja, M. A., Hainzl, S., Schorlemmer, D., Iturrieta, P. C., Bayona, J. A., Savran, W. H., Werner, M., Marzocchi, W. (2023): Statistical power of spatial earthquake forecast tests. - Geophysical Journal International, 233, 3, 2053-2066.

https://doi.org/10.1093/gji/ggad030

Geophysical Journal International

Geophys. J. Int. (2023) **233,** 2053–2066
Advance Access publication 2023 January 24
GJI Seismology

https://doi.org/10.1093/gji/ggad030

Statistical power of spatial earthquake forecast tests

Asim M. Khawaja,[1,2] Sebastian Hainzl,[1,2] Danijel Schorlemmer,[1] Pablo Iturrieta,[1,2] José A. Bayona,[3] William H. Savran,[4,5] Maximilian Werner[3] and Warner Marzocchi[6]

[1] *GFZ German Research Center for Geosciences, Telegrafenberg,* 14473 *Potsdam, Germany. E-mail:* khawaja@gfz-potsdam.de
[2] *Institute of Geosciences, University of Potsdam,* 14476, *Potsdam, Germany*
[3] *School of Earth Sciences, University of Bristol, Queens Road, BS*8 1*RJ, Bristol, UK*
[4] *Southern California Earthquake Center, University of Southern California, Los Angeles, CE* 90089-0742*, USA*
[5] *Nevada Seismological Laboratory, University of Nevada,* 1664 *N Virginia St, Reno, NV* 89557*, United States*
[6] *Department of Earth, Environmental, and Resources Sciences, University of Naples, Federico II* 80126 *Naples, Italy*

Accepted 2023 January 23. Received 2023 January 20; in original form 2022 October 5

SUMMARY

The Collaboratory for the Study of Earthquake Predictability (CSEP) is an international effort to evaluate earthquake forecast models prospectively. In CSEP, one way to express earthquake forecasts is through a grid-based format: the expected number of earthquake occurrences within $0.1^\circ \times 0.1^\circ$ spatial cells. The spatial distribution of seismicity is thereby evaluated using the Spatial test (S-test). The high-resolution grid combined with sparse and inhomogeneous earthquake distributions leads to a huge number of cells causing disparity in the number of cells, and the number of earthquakes to evaluate the forecasts, thereby affecting the statistical power of the S-test. In order to explore this issue, we conducted a global earthquake forecast experiment, in which we computed the power of the S-test to reject a spatially non-informative uniform forecast model. The S-test loses its power to reject the non-informative model when the spatial resolution is so high that every earthquake of the observed catalog tends to get a separate cell. Upon analysing the statistical power of the S-test, we found, as expected, that the statistical power of the S-test depends upon the number of earthquakes available for testing, e.g. with the conventional high-resolution grid for the global region, we would need more than 32 000 earthquakes in the observed catalog for powerful testing, which would require approximately 300 yr to record $M \geq 5.95$. The other factor affecting the power is more interesting and new; it is related to the spatial grid representation of the forecast model. Aggregating forecasts on multi-resolution grids can significantly increase the statistical power of the S-test. Using the recently introduced Quadtree to generate data-based multi-resolution grids, we show that the S-test reaches its maximum power in this case already for as few as eight earthquakes in the test period. Thus, we recommend for future CSEP experiments the use of Quadtree-based multi-resolution grids, where available data determine the resolution.

Key words: Earthquake hazards; earthquake interaction; forecasting and prediction; Statistical seismology; Earthquake forecast testing; Statistical power analysis.

1 INTRODUCTION

Earthquake forecast models are a basic component of probabilistic seismic hazard assessment. They enable us to understand earthquake occurrence better and can lead to building an earthquake-resilient society. Contemporary research about forecasting earthquakes follows numerous approaches, including those based purely on a statistical analysis of the earthquake catalogs and physics-based methods (e.g. Helmstetter *et al.* 2007; Morales-Esteban *et al.* 2010; Martínez-Álvarez *et al.* 2013; Asim *et al.* 2018; Maleki Asayesh *et al.* 2019; Mancini *et al.* 2019; Ahmad *et al.* 2019; Tareen *et al.* 2019; Tariq *et al.* 2019; Asayesh *et al.* 2020; Mignan & Broccardo 2020; Rhoades *et al.* 2020; Sharma *et al.* 2020; Asayesh *et al.* 2022; Ebrahimian *et al.* 2022, etc.).

Earthquake forecast modelling is a complex process, and it is important to assess the skill and performance of forecast models. For that reason, the community-based Collaboratory for the Study of Earthquake Predictability (CSEP) was established (Jordan 2006; Michael & Werner 2018; Schorlemmer *et al.* 2018). The important principle of CSEP is to evaluate earthquake forecasts prospectively against future earthquake observations without any intervention from earthquake forecast modellers, thereby ensuring the reproducibility and transparency of testing experiments and results.

CSEP has conducted experiments in various pre-defined geographical areas referred to as *testing regions* (Schorlemmer *et al.* 2010), e.g. the California testing region (Schorlemmer & Gerstenberger 2007), the Japan testing region (Tsuruoka *et al.* 2012), the Italy testing region (Schorlemmer *et al.* 2010) and the global testing region (Bayona *et al.* 2021). In one type of CSEP forecast experiment design, the testing regions are represented as spatial grids with cells of dimensions of 0.1° × 0.1° in longitude and latitude, where each cell is further subdivided into bins of 0.1 magnitude units. A forecast provides the expected number of earthquakes, assuming a Poisson distribution, for each of the pre-specified space-magnitude bins of the testing region (Schorlemmer & Gerstenberger 2007). The Poisson assumption is debatable, leading to strong biases in evaluating short-term forecasting models (Lombardi *et al.* 2010; Werner *et al.* 2011), and new CSEP efforts are also trying to address this problem (Savran *et al.* 2020; Bayona *et al.* 2022).

CSEP provides a community-endorsed testing suite to evaluate forecast models (Schorlemmer *et al.* 2007; Zechar *et al.* 2010; Werner *et al.* 2011). The suite consists of multiple tests designed to assess the consistency of different aspects of forecasts with observed data assuming a Poisson distribution. These tests evaluate the consistency of the spatial distribution, the magnitude distribution, the combined space-magnitude distribution, and the total number of earthquakes provided by forecast models.

The forecast models mostly use the Gutenberg-Richter frequency-magnitude distribution to provide the expected number of earthquakes per magnitude bin (Bayona *et al.* 2022), leading to similar outcomes of the Magnitude test (M-test) and rendering the evaluation of magnitude consistency less informative than that of the spatial distribution, which is specifically tested with the Spatial test (S-test). Because earthquakes are inhomogeneously distributed around the globe, some regions are seismically quiet and have never witnessed a notable earthquake in recorded history, while others are highly active and frequently experience earthquakes. Therefore, it is important to check the spatial consistency of models with observed earthquake data.

However, the observed data to test earthquake forecast models usually only consist of a small number of earthquakes ranging from a few tens of events to a few hundred earthquakes depending on the testing region and duration of observation. The earthquake forecast evaluation conducted for the California testing region conducted by Bayona *et al.* (2022) used 40 earthquakes covering 10 yr, while Bayliss *et al.* (2022) and Zechar *et al.* (2013) used 32 and 31 $M4.95+$ observed earthquakes covering the duration of 5 yr. The forecast evaluation conducted by Taroni *et al.* (2018) for the Italy testing region is based on 97 $M3.95+$ earthquakes for the duration of 5 yr. Bayona *et al.* (2021) conducted an earthquake forecast evaluation for the global testing region using 651 $M5.95+$ earthquakes for a 6-yr duration. In contrast, the numbers of spatial cells for California, Italy, and the global testing region are 7682, 8993 and 6.48 million, respectively. This means that we have, on average, approximately one earthquake to evaluate the forecast per several hundred spatial cells for the regional testing areas of California and Italy and one earthquake per 10 000 cells on the globe to evaluate the forecast models. The scarcity of observation data can result in tests with low statistical power, leading to a reduced chance of detecting the true performance of models (Bezeau & Graves 2001; Bray & Schoenberg 2013; Button *et al.* 2013). Less powerful forecast evaluations can potentially lead to the use of flawed forecast models for seismic hazard assessment.

The power of the S-test defines the capability of the test to correctly identify whether or not the occurrence of observed seismicity is consistent with the forecast models. Because the true seismicity model is unknown, computing the power of the S-test is not straightforward. The statistical power analysis of the S-test is based on the assumption that the S-test should identify two seismicity models that are different and inconsistent with each other. Therefore, we can determine the power of the S-test to discriminate between two alternative forecast models by using one forecast as data generating model and then using this data to evaluate the other forecast model. The power is computed as the probability that the S-test successfully identifies that the catalogs based on one forecast model are inconsistent with the other forecast model. As a reference for computing the power, we use in this study the Global Earthquake Activity Rate model [GEAR1; (Bird *et al.* 2015)], one of the global forecast models that competed in the aforementioned global experiment (Strader *et al.* 2018). GEAR1 was the most informative forecast model in the global forecast experiment based on 651 $M5.95+$ earthquakes during 2014–2019 (Bayona *et al.* 2021).

In this study, we show two key factors determining the statistical power of the S-test, namely the sample size (number of observed earthquakes) and the grid resolution. While it is not possible to change the amount of observed data for a given duration of time, we can change the test grid, thus the number of cells for which a forecast is made. To explore the S-test's statistical power for different spatial grid resolutions, we use the strategy introduced by Asim *et al.* (2022) to replace the conventional 0.1° × 0.1° grid by the Quadtree-based grids. Quadtree is a hierarchical tiling strategy for recursively dividing the globe into tiles to create easily and elegantly single- or multi-resolution grids. We find that representing earthquake forecast models using fewer cells (low-resolution grid) can improve the power of testing. We observe that for powerful testing for the uniformly gridded global testing region with a reasonable number of earthquakes in the observed catalog (e.g. 1000–2000 earthquakes), we should not have more than 16 000–65 000 cells in the testing region. However, uniformly reducing the resolution results in losing models' spatial information in seismically active regions. Thus, instead of reducing the grid resolution uniformly, we show that using data-driven multi-resolution grids significantly increases the statistical power of the S-test.

The following section explains the CSEP consistency tests in detail, focusing on the S-test. Section 3 introduces the statistical power of a test, and how the S-test's power depends on the number of observed earthquakes and grid resolutions. Section 4 shows the experimental results for the statistical power of the S-test for different grid resolutions, and the global forecasts are re-evaluated in Section 5. In Section 6, we discuss the recommendation for improving the power of the S-test in CSEP experiments and show how this can affect the performance evaluation of forecast models.

2 CSEP FORECAST EXPERIMENT

2.1 CSEP consistency tests

CSEP forecasts are provided as the expected number of earthquakes for a given time horizon in the independent space-magnitude bins. In each bin (indexed by i), the forecasted number of earthquakes

λ_i is then compared with the observed number of earthquakes ω_i in the same cell in various ways depending on the applied test. If earthquakes are assumed to follow a Poisson distribution, the Poisson likelihood of the observation is computed based on the expectation value λ of the model according to

$$Pr(\omega|\lambda) = \frac{\lambda^{\omega}}{\omega!}\exp(-\lambda)\,. \tag{1}$$

It is convenient to work with the logarithm of likelihood values, referred to as log-likelihood, given by

$$L(\omega|\lambda) = \ln(Pr(\omega|\lambda)) = -\lambda + \omega\ln(\lambda) - \ln(\omega!)\,. \tag{2}$$

The joint log-likelihood value, L, for a complete space-magnitude forecast Λ, and observation Ω, is the sum of all bin-wise log-likelihood values:

$$L(\Omega|\Lambda) = \sum_{i=1}^{N_{\text{bin}}} \ln(Pr(\omega_i|\lambda_i) = \sum_{i=1}^{N_{\text{bin}}}(-\lambda_i + \omega_i\ln(\lambda_i) - \ln(\omega_i!)), \tag{3}$$

where N_{bin} refers to the total number of space-magnitude bins. The joint log-likelihood values given by eq. (3) are negative, with higher values (closer to zero) indicating better agreement between forecast and observation. Earthquake catalogs consistent with the forecast model are simulated ($\hat{\Omega}$) to understand the uncertainty of the joint log-likelihood. Currently, in CSEP, the procedure employed to generate ($\hat{\Omega}$) using the earthquake forecast is provided by Schorlemmer & Gerstenberger (2007) and Zechar *et al.* (2010). Numerous simulated catalogs are generated, and log-likelihood values for the forecast given the simulated catalogs are computed, referred to as simulated log-likelihoods, $\hat{L}$, thereby generating a distribution of $\hat{L}$ values:

$$\hat{L} = L(\hat{\Omega}|\Lambda). \tag{4}$$

The acceptance or rejection of the model is decided by comparing the log-likelihood value with the simulated log-likelihood values. The model can be considered inconsistent if the log-likelihood value of the observation falls in the lower tail of the simulated log-likelihood values. Otherwise, the forecast model is considered consistent with the observation.

All consistency tests in the CSEP testing suite follow the procedure above, where forecast rates λ are related to different aspects. The S-test evaluates the consistency of earthquake forecast rate distribution across spatial cells compared to the observed catalog. The magnitude aspect of the forecast and observation is excluded first by summing up the magnitude bins corresponding to every spatial cell, followed by normalizing the forecast rates so that their sum matches the total number of earthquakes in the observed catalog (N_{obs}). The total number of earthquakes in simulated catalogs ($\hat{N}$) is fixed to the N_{obs} for computing $\hat{L}$ values.

Recently, Bayona *et al.* (2022) proposed a new version of the S-test based on the Binary likelihood function instead of the Poisson likelihood. This new test is proposed to capture the non-Poisson distribution of data better. The Binary likelihood function treats observations in terms of active or non-active cells, unlike considering the number of earthquakes for each cell in the grid, and calculates the Poisson probabilities of observing $\omega_i = 0$ and $\omega_i > 0$ earthquakes in a spatial cell given the forecast λ_i. The whole procedure of the Binary S-test remains the same except for $\hat{N}$ being fixed to the total number of active cells in the observed catalog during the generation of simulated catalogs and the forecast being normalized to the number of active cells instead of N_{obs}. The Binary S-test is meant to reduce the sensitivity of the S-test to the presence of seismicity clusters (or the presence of multiple earthquakes in cells) in the observed catalog.

Given that the underlying procedure of all consistency tests available in the CSEP testing suite remains similar, weaknesses or low power identified in the S-test might hint at similar issues in other CSEP consistency tests. Therefore, it is important to understand the conditions in which the outcome of the S-test is statistically powerful, thus, more reliable.

2.2 Global forecast experiment

In CSEP, numerous regional and global testing experiments have been conducted for various tectonic settings. Compared to the regionally calibrated forecast models, global forecast models offer greater testability due to the availability of a higher number of earthquakes in the observed catalog despite the huge disparity in the number of spatial cells and events.

A prospective global earthquake forecast experiment to evaluate forecast models with 2 yr of observations from 2015 October to 2017 September has been conducted by Strader *et al.* (2018). The competing forecast models in this experiment were the global hybrid GEAR1 (Bird *et al.* 2015), the tectonic SHIFT_GSRM2f (Bird & Kreemer 2015) and the seismicity KJSS (Kagan & Jackson 2010, 2011) models. Later on, Bayona *et al.* (2021) constructed two more global earthquake forecast models, named TEAM (tectonic) and WHEEL (Hybrid), and included them in the global forecast experiment. In this experiment, the forecast testing has been carried out for 6 yr, from 2014 to 2019, using 651 earthquakes of $M \geq 5.95$ (Bayona *et al.* 2021) reported by the Global Centroid Moment Tensor (CMT) earthquake catalog (Ekström *et al.* 2012). The GEAR1, WHEEL and KJSS models passed the S-test and were thus found to be spatially consistent with the observed catalog. In contrast, the geodetic-based TEAM and SHIFT_GSRM2f models were found to be inconsistent with the observations. These experiments concluded that the hybrid GEAR1 model was the most informative earthquake forecast model during 2014–2019 evaluation period. We conduct the S-test for these five global forecasts using pyCSEP. Previously, the CSEP testing suite was available as a monolithic and tightly coupled code base. Recently, it has been redesigned into an object-oriented and open-source toolkit in Python, known as pyCSEP (Savran *et al.* 2022b, a). This toolkit provides an independent module containing all the community-endorsed statistical tests in the CSEP testing suite. We obtain the same performance results as reported by Bayona *et al.* (2021), thereby demonstrating the reproducibility of the test results and ensuring the credibility of the testing infrastructure used.

In addition to testing the previously competing global forecast models, we also generate and test a uniform global forecast model. The uniform forecast model is created by assigning the same forecast density to every cell in the grid, thus yielding forecasts proportional to the area of each cell. It is a simple forecast model and does not involve any information about the actual seismicity with its variability; therefore, it is a non-informative forecast model. The uniform forecast model expects independent, unclustered and evenly distributed seismicity across the whole region in contrast to the heterogeneous distribution of the actual seismicity (Kagan 2007). Consequently, the desired outcome of the S-test is to reject the uniform model based on the observed catalog. Fig. S1 shows the uniform forecast model along with the spatial distribution of observed seismicity, while Fig. 1 shows the outcome of the S-test

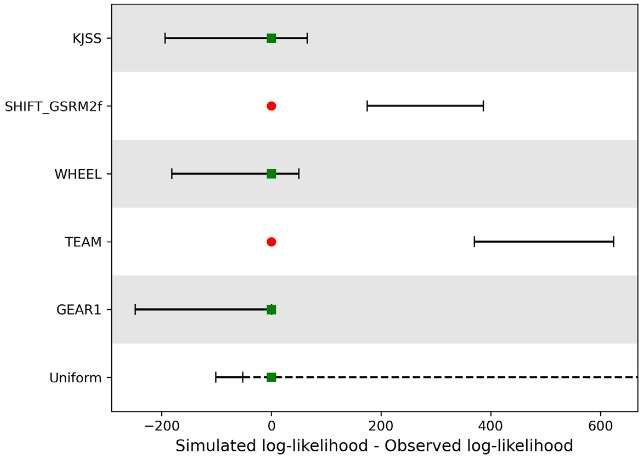

Figure 1. Performance of the S-test for models participating in the global forecast experiment, along with the performance of the uniform forecast model, which is passing the S-test. The error bars show the log-likelihood confidence interval relative to the observed log-likelihood score. The red dots indicate that the observed log-likelihood falls in the lower tail of the simulated log-likelihood and fail the S-test, while the green squares indicate that the S-test is passed.

for all the competing forecast models. The uniform global forecast model passes the S-test contrary to the expectations, showing that the non-informative uniform forecast is consistent with the observation, which also raises concerns regarding the S-test's capability to evaluate forecast models objectively.

The S-test tends to favour more uniform forecast models than other models that try to forecast the precise locations of earthquakes (i.e. provide higher forecast rates in certain cells). In the case of the uniform forecast model, the forecasts in the cells are similar everywhere; thus, it does not matter where the earthquakes are occurring, and the log-likelihood value is not controlled by the spatial distribution of the observed earthquakes but rather by the number of earthquakes occurred in the cells. For a grid with high resolution as the 0.1° × 0.1° grid for the global testing experiment, the size of spatial cells is so small that each cell usually observes only one earthquake. This also happens to be the case for the simulated catalogs. Thus, the likelihood for observed data will fall in the range of the simulated likelihoods, and the S-test accepts the forecast as consistent with the observation. In other words, the global uniform forecast states, obviously incorrect, that earthquakes are everywhere similar likely on the globe, and the S-test is unable to reveal this inconsistency with the observed seismicity. Thus, the uniform forecast expressed on such a high-resolution grid can only be rejected by the S-test if the observed catalog contains clusters of seismicity, which means that some spatial cells in the grid receive multiple earthquakes.

To be useful, the S-test should be powerful enough to reject a non-informative forecast model irrespective of the occurrence of clustered seismicity. Fig. 1 shows that the S-test rejects the SHIFT_GSRM2f and TEAM forecast models, while other models are found consistent with the observed seismicity. However, the acceptance of the S-test can be meaningless if the power of the test is low. To address this problem, we conduct a detailed analysis of the statistical power of the S-test to find out how (i) the sample size (earthquakes in the test catalog) and (ii) the definition of the spatial test grid contribute to the forecast evaluation.

Table 1. Statistical power of S-test for different number of earthquakes in the test catalogs generated using Λ_1 = GEAR1 and 100 simulations.

Λ_2	64	128	256	512	1024	2048
KJSS	0.18	0.28	0.56	0.86	0.99	1
SHIFT_GSRM2F	0.75	0.97	1	1	1	1
TEAM	0.98	1	1	1	1	1
WHEEL	0.2	0.32	0.58	0.83	0.99	1
Uniform	0	0	0	0	0	0

3 STATISTICAL POWER ANALYSIS

The power of a test is defined as the probability of rejecting a null hypothesis (H_0) when it is false (Lehmann *et al.* 2005) or, in other words, the probability of making a correct decision when rejecting a hypothesis (Lehmann *et al.* 2005),

$$\text{Power} = \Pr(\text{Correctly rejecting } \mathrm{H}_0). \tag{5}$$

The value of statistical power ranges from 0 to 1, and as the power of a test increases, the probability of wrongly failing to reject the null hypothesis decreases.

In CSEP forecasting experiments, there is no accepted true model for seismicity, and the likelihood tests evaluate equipollent hypotheses (Schorlemmer *et al.* 2007). Therefore, we use an indirect way to calculate the power of the test based on simulations (Zechar *et al.* 2010). We assume one earthquake forecast model as the true model of seismicity (Λ_1) and use its simulations as observed catalogs. This observation is then used to evaluate a different forecast model (Λ_2) using the S-test. The process of generating observations based on Λ_1 to evaluate forecast models Λ_2 is repeated multiple times, and the power is estimated as the fraction of instances in which the simulated Λ_1 catalogs (assumed observation) are inconsistent with the Λ_2 forecast, i.e.,

$$\text{Power of S-test} = \frac{\text{Number of S-test failures}}{\text{Total number of simulations}}. \tag{6}$$

The flowchart for computing the statistical power of the S-test is shown in Fig. S2.

The GEAR1 forecast model was found to be the most informative forecast model in the CSEP global forecast experiment based on the pair-wise comparison with other models in terms of information gain per earthquake (Bayona *et al.* 2021). Thus, in statistical power analysis of the S-test, we consider GEAR1 as the seismicity generator (Λ_1) and generate earthquake catalogs with a different number of earthquakes [N_{obs} =(64, 128, 256, 512, 1024, 2048)] in the 0.1° × 0.1° global test grid. Then, we use these observed catalogs to evaluate all the competing earthquake forecast models, including the uniform forecast. This experiment is repeated 100 times, and the statistical power of the S-test is calculated and provided in Table 1 for each forecast. The power of the S-test is found to be directly correlated to the number of earthquakes in the observed catalog. This finding agrees with the general understanding that the statistical power of tests used to evaluate forecast models is directly related to the quantity of data available for testing (Bezeau & Graves 2001).

Table 1 demonstrates that the S-test has a different power to reject different forecast models for synthetic GEAR1 catalogs, which is due to different levels of similarity between the GEAR1 and the other forecast models. The TEAM and SHIFT_GSRM2f are rejected for small test samples, while KJSS and WHEEL require larger data sets. The latter can be explained by the fact that WHEEL and GEAR1 are both using KJSS as one model component, and thus the models are similar. However, the power of the S-test to correctly reject the models increases with the number of earthquakes. In contrast, the S-test remains powerless against the uniform forecast model with zero power, even with more than 2000 earthquakes in the observed catalog. In this study, we use the uniform forecast model as Λ_2 for the statistical power analysis because it is a simple model that is neither based on any meaningful dataset (i.e. earthquake catalogs, etc.) nor physics and statistics. Thus, it can be assumed that the uniform forecast model is different from the GEAR1 and should be rejected.

The power of a test is not a fixed value but depends on the hypotheses (i. e. forecast models) being considered for the power analysis. The number of earthquakes in the observed catalog is an important factor to consider when conducting a forecast evaluation experiment, but it is not the only factor contributing to the power of the test. Thus, we further need to explore the statistical power of the S-test for different representations of forecast in terms of the spatial grid, where we can reduce the number of cells.

3.1 Showcasing power of S-test

Real forecast models may involve multiple complexities, such as regional (tectonic) variations, data dependence, etc. Therefore, before exploring the effect of different grid resolutions on the statistical power of the S-test results for real forecast models, we perform a simple synthetic experiment to understand the behaviour of the S-test in a fully controlled environment. We create a hypothetical one-dimensional scenario where we test the uniform forecast model against earthquakes inhomogeneously distributed in space. In particular, we used a Gaussian distribution with zero mean and a standard deviation of 1.0 normalized in the spatial range between −3 and 3 to select the events randomly. The forecasts and test data are binned in one-dimensional grid cells equivalently to the two-dimensional grid cells in real CSEP forecast experiments. In the experiment, we create two different types of discretizations in space, using either (i) uniform bins or (ii) density-based bins. The uniform bins are equally spaced, while density-based bins are created in a way that all bins have, on average, the same event rate. The locations of the test samples drawn from the Gaussian distribution are associated with their corresponding bin. Thus, each bin gets either zero, one, two or more *observed* events per bin. Using the forecast and the test sample, we have all the ingredients of an earthquake forecast experiment to run the S-test. We use the S-test to determine whether or not the uniform distribution is consistent with the data generated by the Gaussian distribution. The experiment is repeated multiple times, and the statistical power of the S-test is calculated for different combinations of uniform bins, density-based bins, the total number of bins and the sample size.

Fig. 2 summarizes the whole synthetic experiment and showcases the statistical power of the S-test. As an example, panel (a) shows 20 uniform bins with 10 events (blue points) randomly selected from the Gaussian distribution (blue line), while panel (b) shows the same case for density-based bins, having small-sized bins in the centre and bigger bins towards the edges. As the density-based bins are designed to contain, on average, an equal rate (or probability) for each bin, thus the true rate becomes a horizontal line instead of the bell-shaped curve, while the uniform forecast (black line) becomes an inverted bell-shaped as more data are expected in the larger cells towards the edges. By repeating the S-test for 1000 random simulations, we first explore the dependence of the test power on the number of spatial grid cells. In particular, we set the sample size to $N_{\mathrm{eq}} = 10$ and change the number of grid cells, N_{cell}, from 1 to 100. As shown in panel (c), we find that the power reaches for uniform binning (black curve) its maximum (∼0.3) at $N_{\mathrm{cell}} = 4$ and then monotonously decreases with increasing N_{cell}. In contrast, the power converges for the density-based bins (blue curve) to a value of about 0.85 for increasing N_{cell}. In a second step, we explore the dependence of the power on the sample size given $N_{\mathrm{cell}} = 20$ (panel d). We find a systematic increase of the power with N_{eq} for both binning approaches. However, the increase in power is significantly faster for density-based bins than the uniformly spaced binning. A power of 0.9 is already obtained for 15 events in the former case but for 38 events in the latter case. It should be noted that the example shown in Fig. 2 refers to a rather smooth seismicity distribution. Repeating the same experiment for a Gaussian distribution with a standard deviation of 0.5 instead of 1.0 shows that a power of 0.9 is already achieved for $N_{\mathrm{cell}} = 6$ [with $N_{\mathrm{eq}} = 10$, panel (c)] and for five events [with $N_{\mathrm{cell}} = 20$, panel (d)] for the same density-based grids.

This simple experiment, which emulates the earthquake forecast experiment with hypothetical spatial grids and earthquakes, provides meaningful insights into the statistical power of the S-test. It highlights different potential factors affecting the power of the S-test, such as the number of earthquakes in the test catalog and the definition of the spatial grid. The density-based binning (hereafter referred to as multi-resolution grid) shows the capability to increase the statistical power of the test compared to uniformly spaced binning (hereafter referred to as single-resolution grid). This simplified test setup points us to how the choice of the spatial grid can improve the test performance.

4 EFFECT OF SPATIAL GRID ON STATISTICAL POWER OF S-TEST

In CSEP experiments for evaluating forecast models, the choice of 0.1° × 0.1° spatial grid for representation of earthquake forecasts has been the most convenient choice of a grid because it is easy to handle in computer codes, intuitive and simple to understand. However, the outcomes of the experiment conducted in Section 3.1 suggest exploring other resolutions for powerful testing of forecast models. Recently, Asim *et al.* (2022) proposed Quadtree to acquire spatial grids for CSEP forecast experiments as a replacement for 0.1° × 0.1° grid. Quadtrees are not only easy to implement and intuitive but also provide the flexibility to generate grids with desired resolutions. Furthermore, the Quadtree approach is already integrated into pyCSEP, making the grid generation even more straightforward. Thus, we use the recently proposed Quadtree grids to explore the statistical power of the S-test for different resolutions of spatial grids.

4.1 Quadtree grids for CSEP experiments

The Quadtree is a tree-based hierarchical data structure in which each node is allowed to have either zero or four child nodes. The Quadtree, in combination with the Mercator projection of the earth, divides the global map into four quadrants, also called *tiles*, where the prime meridian and the equator define the dividing lines. In this case, we refer to the globe as the root node, representing the globe up to 85.05° latitude north and south. In the first step, the root tile is divided into four square subtiles, the NE, NW, SW and SE regions. These tiles are uniquely identified using *Quadkey*, which are numbers of the base-four system 0, 1, 2 and 3, respectively. Each of these four tiles can be further divided into four square subtiles. The Quadkeys of these subtiles are generated from the Quadkey of

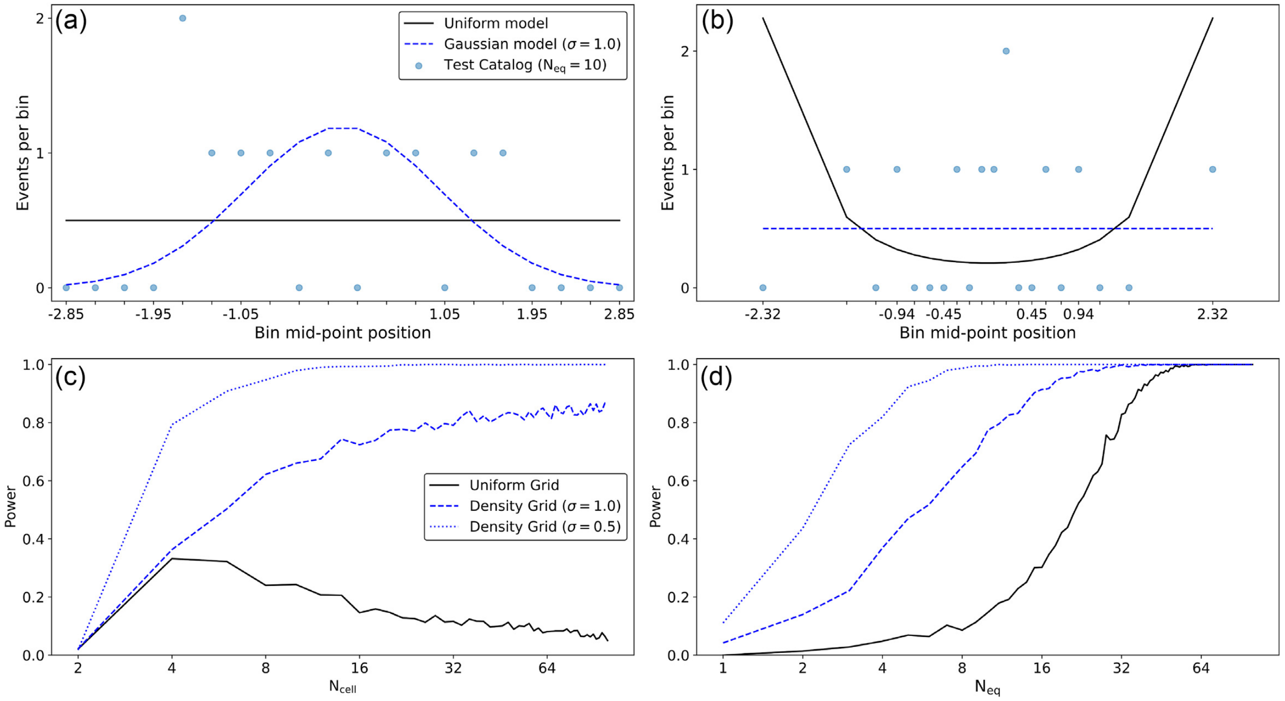

Figure 2. A showcase of the potential factors affecting the statistical power of the S-test using a synthetic experiment. In this synthetic experiment, a uniform forecast model is tested for events randomly sampled from a one-dimensional Gaussian distribution (mimicking observed earthquakes) using the S-test. We analyse the S-test's statistical power as a function of the number of events (N_{eq}) and the number of grid cells (N_{cell}) for uniform and density-based binning (mimicking spatial grids). (a) Illustration of 20 uniformly spaced bins, along with its counts for 10 events randomly sampled from Gaussian distribution with $\sigma = 1$. (b) Same as (a) but for 20 density-based bins. (c) Dependence of the S-test's power on N_{cell}, given $N_{\mathrm{eq}} = 10$. (d) The power of the S-test as a function of N_{eq} in the case of 20 grid cells for uniform and data-based binning.

the parent tile by appending the relative Quadkey of the subtile (0, 1, 2 or 3), e.g. the subtiles of tile 3 are 30, 31, 32 and 33. The number of times a tile is divided is called the zoom level (L). This way, the entire globe can recursively be divided into as many tiles as desired, with a unique Quadkey for every subtile. Once the tiling process reaches the required decomposition, we refer to it as a Quadtree grid, and each tile is referred to as a spatial grid cell.

We can create a data-driven multi-resolution grid by associating the recursive division of tiles subject to the data availability. We can also use other data sets that could potentially be involved in creating the forecast models to define Quadtree spatial grids, such as seismicity, information about global strain rate, Coulomb stress changes, etc. Here, for simplicity, we only use the observed seismicity to generate the Quadtree spatial grids. The idea is to increase the grid's resolution in regions with high earthquake density while keeping cells large for seismically less active regions. To generate such a data-based multi-resolution grid, first, we define a threshold for the maximum number of data points (earthquakes in our case) allowed per cell, N_{max}. We can also introduce an additional criterion, such as a minimum cell area or maximum zoom level (L_{max}), allowed for a cell to ensure that cell size does not get too small for seismically dense regions.

4.2 Statistical power analysis: single-resolution grids

We use the Quadtree for the global testing region to analyse the statistical power of the S-test against the uniform forecast model in the case of different single-resolution grids. We generate spatial grids at different zoom levels [$L = (5, 6, 7, 8, 9, 10 \text{ and } 11)$], which lead to the spatial grids with a different number of cells [(1024, 4096, 16384, 65536, 262144, 1048576 and 4194304)]. We name these single-resolution Quadtree grids based on their zoom level as *L5*, *L6*, *L7*, *L8*, *L9*, *L10* and *L11*. For every grid, a uniform forecast model is generated. The observed catalogs of different sizes [$N_{obs} = 2^i$, where $i = (6, 7, 8, \ldots, 15)$] are simulated using the seismicity model Λ_1 (i. e. GEAR1) in the same way as explained in Section 3. Now the uniform forecast for each spatial grid is evaluated using the simulated observed catalogs, and the statistical power is computed for all the combinations of spatial grids and N_{obs}, as shown in Fig. 3. The figure reveals that the S-test only achieves maximum power, i. e. the uniform forecast is rejected in all 100 simulated catalogs, if the number of earthquakes in the observed catalogs exceeds 32 000 in the case of a single-resolution grid with approximately 4.2 million cells, which is the nearest to the conventional 0.1° × 0.1° grid in terms of the number of cells. The trends of statistical power observed in this figure are consistent with Table 1 and Fig. 2, highlighting that the statistical power of the S-test increases with more earthquakes in the observed catalog. The analysis also explains our result that the S-test cannot reject the uniform model for the observed earthquakes in the 0.1° × 0.1° gridded global test region (see Section 2.2). Since we do not have any true seismicity model for the earth, we rely on simulations from GEAR1 as a proxy. Therefore, the statistical power analysis of the S-test is an approximation for the real scenarios, which should help design the forecast experiment for powerful testing of forecast models.

In the aforementioned global forecast experiment, only 651 earthquakes are available to test the forecast in 6.48 million spatial cells. The decrease in the resolution of the grid can lead to a powerful S-test with fewer earthquakes. Using Fig. 3 as a look-up table, 651

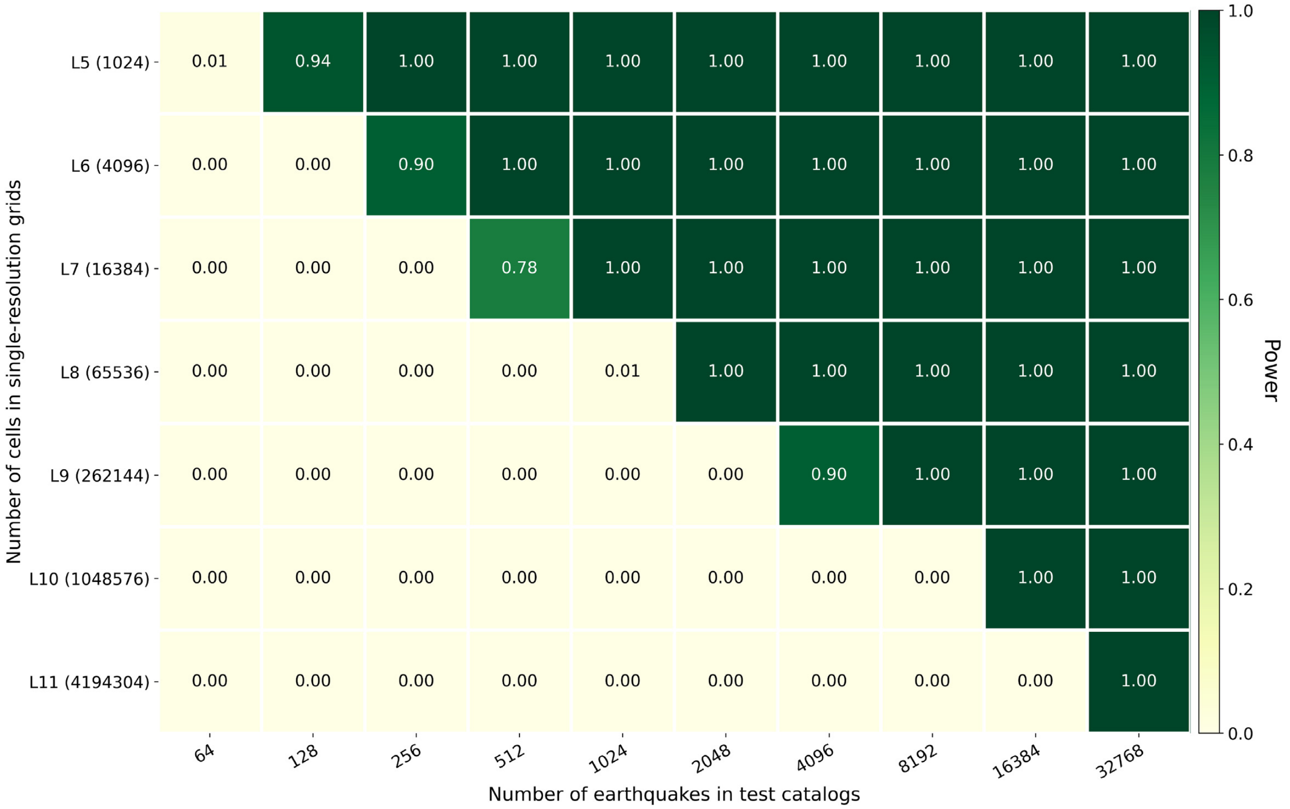

Figure 3. The statistical power of the S-test with GEAR1 as the data-generating model to reject the uniform model forecasts for single-resolution grids indicates that the power of the S-test increases for the grids with lower resolution and a larger test sample size.

observed earthquakes of the actual forecast experiment suggest that we should have at most 16 000 spatial cells in the global test grid to conduct a statistically powerful forecast evaluation. However, simply decreasing the resolution of earthquake forecast models uniformly from a grid of 6.48 million cells to 16 000 cells leads to bigger spatial cells (e.g. more than 300 × 300 km cell dimensions around the equator) and loss of spatial information provided by a model, particularly for regions with dense seismicity. Therefore, we need to explore the density-based grids for testing earthquake forecast models, as also suggested by the outcome of the synthetic experiment in Section 3.1.

4.3 Statistical power analysis: multi-resolution grids

We generate different data-driven multi-resolution grids based on the global CMT catalog (1976–2013) and repeat the rest of the steps to determine the statistical power of the S-test. For this analysis, we use the global CMT catalog without declustering up to the year 2013 to generate data-based Quadtree grids. For demonstrating the use of multi-resolution grids for the S-test analysis, we used the catalog consisting of 28 465 earthquakes covering 37 yr from 1976 to 2013 with $M \geq 5.15$. The corresponding frequency-magnitude distribution of earthquakes is shown in Fig. S3. For the grid generation, we use the criteria that the number of earthquakes in a cell is not allowed to exceed $N_{\rm max}$ as long as the grid resolution is smaller than $L_{\rm max}$. Here, we set $L_{\rm max} = 11$.

We generate grids for different thresholds on the maximum number of earthquakes allowed per cell as $N_{max} = (100, 50, 25, 10, 5$ and $1)$, which results in multi-resolution grids with the number of cells (922, 1780, 3502, 8089, 14 782 and 39 811), respectively. We name these grids as *N100L11*, *N50L11*, *N25L11*, *N10L11*, *N5L11* and *N1L11*. The data-based grids can reduce the number of spatial cells without losing the resolution of the spatial seismicity distribution. The uniform forecast is aggregated to every multi-resolution grid, and the statistical power is computed for each grid using synthetic catalogs based on Λ_1 (GEAR1). The trend of statistical power of the S-test for data-based multi-resolution grids as a function of $N_{\rm obs}$ is recorded in Fig. 4. The figure indicates that the S-test achieves high power in evaluating earthquake forecast models when tested on multi-resolution grids. The S-test rejects the uniform model for all 100 simulations (estimated power of 1.0) for every multi-resolution grid with just four or eight earthquakes in the test catalog. Thus, if we aim to increase the statistical power of the S-test, then we must evaluate the forecast models on the data-driven multi-resolution grids without waiting for more data.

5 RE-EVALUATION OF GLOBAL FORECASTS

Our synthetic study in the previous section shows that the choice of the test grid strongly impacts the S-test's statistical power. The disparity in the number of spatial test cells and the number of observed earthquakes leads to the motivation to analyse the statistical power of the S-test. Furthermore, the S-test cannot even find the non-informative uniform forecast model as inconsistent with the observed seismicity. It raises questions about the significance of the S-test result in previous CSEP forecast experiments. The result of any test with low power can be considered less informative. Thus, we re-evaluate the global models on different single and multi-resolution Quadtree grids to identify powerful testing.

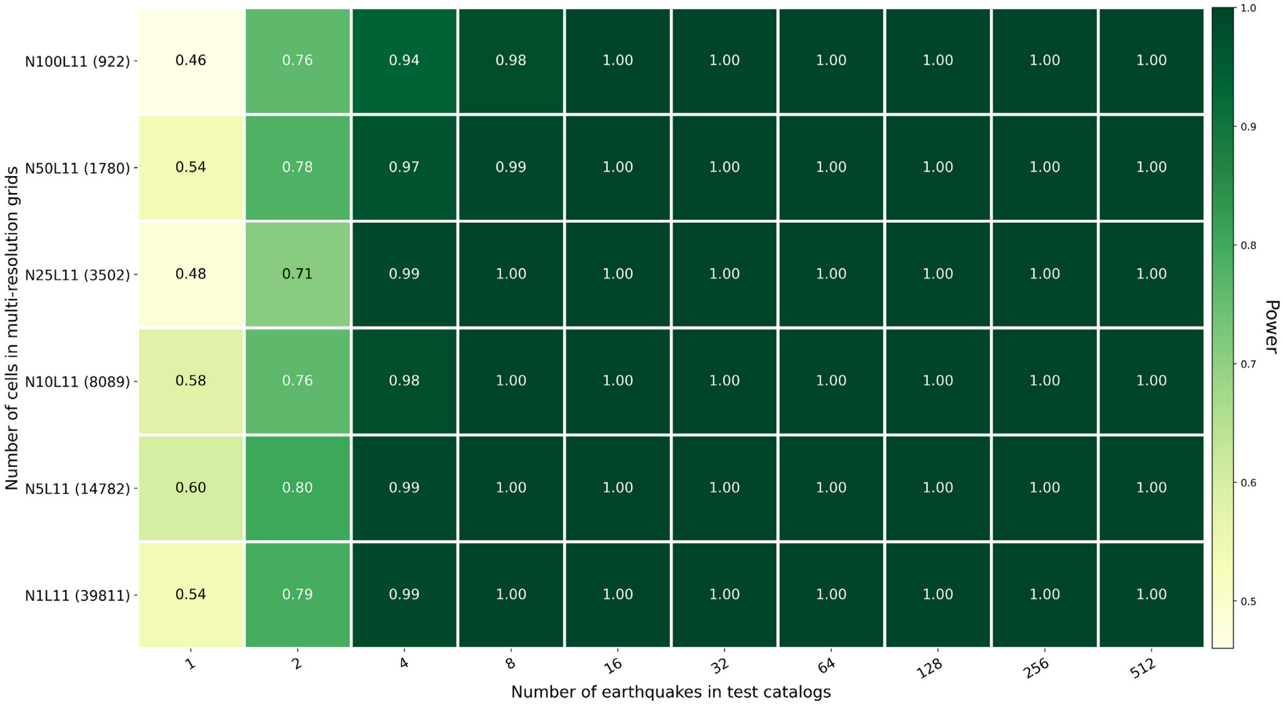

Figure 4. The statistical power of the S-test with GEAR1 as the data-generating model to reject the uniform forecast model for multi-resolution grids indicates that with multi-resolution grids, statistically powerful testing can be performed with as few as four or eight earthquakes in the observed catalog.

The forecast models competing in the global forecast experiment shown in Fig. 1 are available for the 0.1° × 0.1° grid. We need to explore how the performance of these forecast models changes for different grids. Because we do not have the codes to recreate those models for different grids, we create multiple versions of the forecast models by mapping the forecasts from the 0.1° × 0.1° grid onto the multiple Quadtree grids. Mapping forecasts from 0.1° × 0.1° grid to other grids involve aggregation and de-aggregation Asim *et al.* (2022). Aggregation of forecasts of smaller cells to a larger cell is done by summing the smaller cells' rates. The forecast mapping from a bigger cell to smaller cells is referred to as forecast de-aggregation, which is carried out by uniformly distributing the rate of the bigger cell into smaller cells. For forecasts available at a Quadtree grid of any resolution, the forecast (de-)aggregation is fast and computationally inexpensive because all the cells are exactly comparable, and not a single cell from the model grid is shared between two or more cells of the testing grid. However, for aggregating forecasts from 0.1° × 0.1° to Quadtree grids, we also come across cells that are shared between multiple adjacent cells. For such shared 0.1° × 0.1° cells, we also assume a uniform seismicity rate within the cells and distribute the forecast rates to the intersecting test cells according to the overlap area with these cells.

The forecast aggregation from a 0.1° × 0.1° model grid to a Quadtree grid raises the question of whether the change in the forecast grid affects the consistency of the model or not. In theory, the sum of Poisson models is also a Poisson model with the rate $\Lambda = \sum \lambda_i$. Thus, a model passing the consistency test on higher resolution should also pass on the lower resolution grids after aggregation. In order to demonstrate this, we use the GEAR1 forecast model available for 0.1° × 0.1° as a seismicity generator and simulate catalogs with $N_{\mathrm{obs}} = 2^{[6,\,7,\,8,\,\ldots,\,15]}$ events. For each N_{obs}, we generate 100 random catalogs. Then, we aggregate GEAR1 on different multi-resolution Quadtree grids and evaluate them using the S-test against the generated GEAR1 catalogs. We determine the performance of the S-test as the fraction of times the S-test is passing for each N_{obs}. The results are presented in Fig. 5, showing that the catalogs generated by GEAR1 of 0.1° × 0.1° grid and the forecasts aggregated on different grids are consistent with each other. The S-test traditionally uses a 95 per cent confidence interval, with a rejection when the actual value is outside the lower bound. Thus, a true model should theoretically pass the test in 97.5 per cent of the cases. Our results show that the values scatter between 95.6 per cent and 100 per cent. These values are consistent with the theoretical value considering that we use 1000 random catalogs. The results indicate that (de-)aggregation of a true model to the Quadtree grids does not affect the consistency of the forecast models. A forecast model representing the true spatial earthquake distribution will pass the S-test if (de-)aggregated to Quadtree grids. However, as shown by our previous tests, a false model, which passes the test on the 0.1° × 0.1° grid, might be rejected on the new test grids because the statistical power is increased.

We aggregate the forecast models on different Quadtree grids discussed in Section 4 and evaluate them using the S-test based on the same test data used for the recent global forecast experiment, i.e. the 651 earthquakes from the Global CMT catalog observed in 2014–2019 with $M_w \geq 5.95$. As an example, we show GEAR1 aggregated on grid L6 (single-resolution) and grid N50L11 (multi-resolution), along with the observed earthquakes in Figs S4 and S5, respectively. The Poisson S-test results are shown for all the grids and the different forecast models in Fig. 6. The log-likelihood confidence interval is shown relative to the observed log-likelihood score. The outcome of the S-test in terms of non-normalized log-likelihood values is shown in Fig. S6. The forecasts have different test results when evaluated at different grid resolutions. Our previous analysis using GEAR1 as a data-generating model suggests that the observed catalog with 651 earthquakes can probably lead to statistically powerful S-tests for either single-resolution grids with

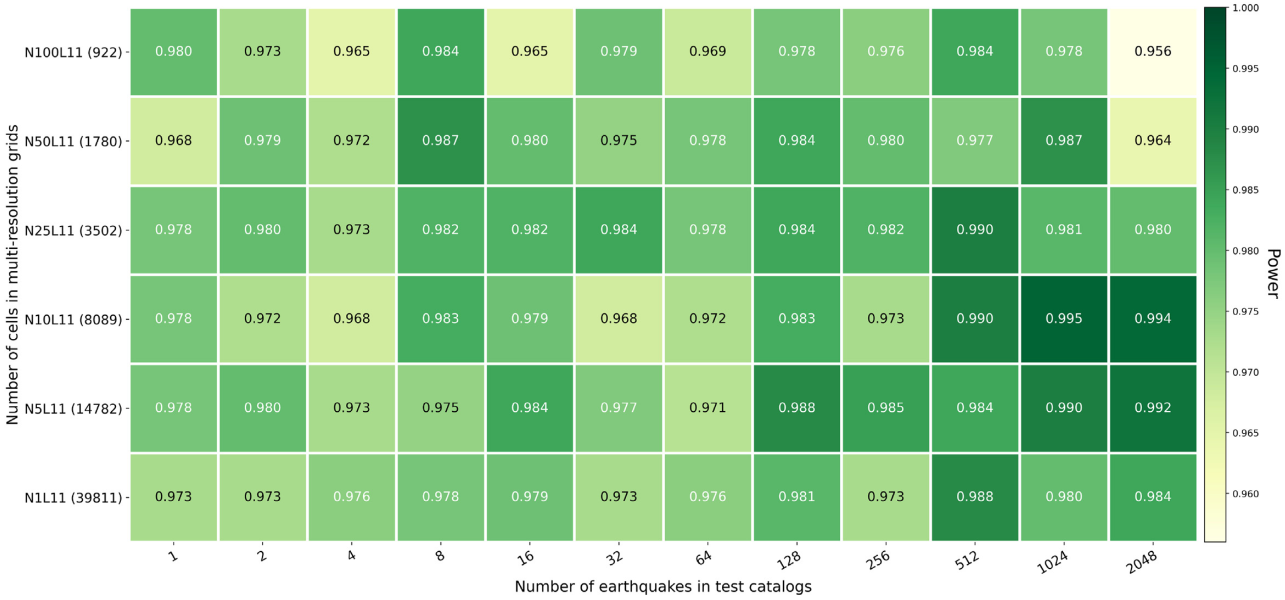

Figure 5. Observed catalogs generated using GEAR1 at 0.1° × 0.1° grid to evaluate GEAR1 aggregated forecasts on different multi-resolution grids, showing that the aggregated forecasts are spatially consistent with the actual forecast provided at 0.1° × 0.1° grid.

less than 16 000 cells or multi-resolution grids. The forecasts of SHIFT_GSRM2f and TEAM were already found inconsistent by the S-test at the 0.1° × 0.1° grid. We find that these models are indeed limited in their ability to forecast the location of observed earthquakes on all other grids as well. In contrast, the forecasts of GEAR1, KJSS and WHEEL, as well as the uniform model, all pass the S-test for the 0.1° × 0.1° grid as well as high-resolution single-resolution Quadtree grids. The uniform models starts to fail the S-test at L8 ($N_{\text{cell}} = 65536$), WHEEL at L7 ($N_{\text{cell}} = 16384$), GEAR1 and KJSS at L5 ($N_{\text{cell}} = 1024$). All the models fail the S-test for all multi-resolution grids as well. Thus, none of the models is spatially consistent with the observations.

To further analyse the S-test's behaviour, we keep on increasing the resolution of spatial grids beyond grid L5, i.e. L4, L3, L2 and L1 with spatial cells of 256, 64, 16 and 4, respectively. For a grid with just one spatial cell, the S-test equals the Number-test, and the test will always be passed due to the S-test condition that the number of earthquakes in the forecast models equals the number of observed earthquakes. The S-test results for the forecasts at the other low-resolution grids are also provided in Fig. S6, along with non-normalized S-test results for the higher-resolution grids. With decreasing the number of cells in single-resolution grids, the confidence interval of simulated distribution is shrinking and drifting towards 0 (the maximum possible log-likelihood value). The observed log-likelihood falls in the lower tail of the confidence interval for all the forecast models, while the uniform model fails the S-test with a relatively big margin compared to other forecast models.

We also apply the newly proposed Binary S-test to all the aggregations of forecast models to analyse whether the failure of the S-test is related to short-time clustering (Bayona *et al.* 2022). The Binary S-test is known to be less sensitive to clustering in contrast to the Poisson S-test. The outcome of the Binary S-test is shown in Fig. 7 in terms of the log-likelihood confidence interval relative to the observed log-likelihood score. The same result is also provided in terms of non-normalized log-likelihood values in Fig. S7. The results show that more forecast aggregations are consistent with the observation based on the Binary S-test than the Poisson S-test, as expected. The forecasts models, GEAR1 and KJSS, pass the Binary S-test for two more grid aggregations, including one multi-resolution grid i.e. *L5* ($N_{\text{cell}} = 1024$) and *N100L11* ($N_{\text{cell}} = 922$) as compared to the Poisson S-test. Similarly, WHEEL passes the Binary S-test for all the single-resolution grids and one multi-resolution grid of *N100L11*. Thus, the Binary S-test makes the S-test less sensitive to the presence of temporal seismicity clusters in the test catalog. However, all forecasts fail the Binary S-test for the rest of the multi-resolution grids, indicating that short-time earthquake clustering might not be the only reason why the models fail the S-test.

We have also explored the performance of the Binary S-test by further reducing the resolution of spatial grids up to L4, L3, L2 and L1 with 256, 64, 16 and 4 spatial cells, respectively. The results are shown in Fig. S7, along with non-normalized Binary S-test results. The results show that all forecasts pass the Binary S-test for these low-resolution grids.

6 DISCUSSION AND RECOMMENDATION

We substitute the lack of availability of a true seismicity model with a forecast model showing the highest information gain (i.e. GEAR1) (Bayona *et al.* 2021) and use it as an earthquake generator to explore the conditions for statistically powerful testing against the uniform forecast model. The resolution of the grid or the number of earthquakes required to conduct the statistically powerful S-test may change slightly with different choices of the earthquake-generation model used in the analysis. However, the essence of the analysis shall remain valid for other reasonable earthquake generation models, with consistent outcomes to those of Sections 3.1 and 4. As an example, we replace the GEAR1 with TEAM for generating seismicity and compute the statistical power of the S-test. The corresponding results are provided in Figs S8 and S9 for single-resolution and multi-resolution grids, showing similar S-test results. This study essentially provides a way forward to conduct

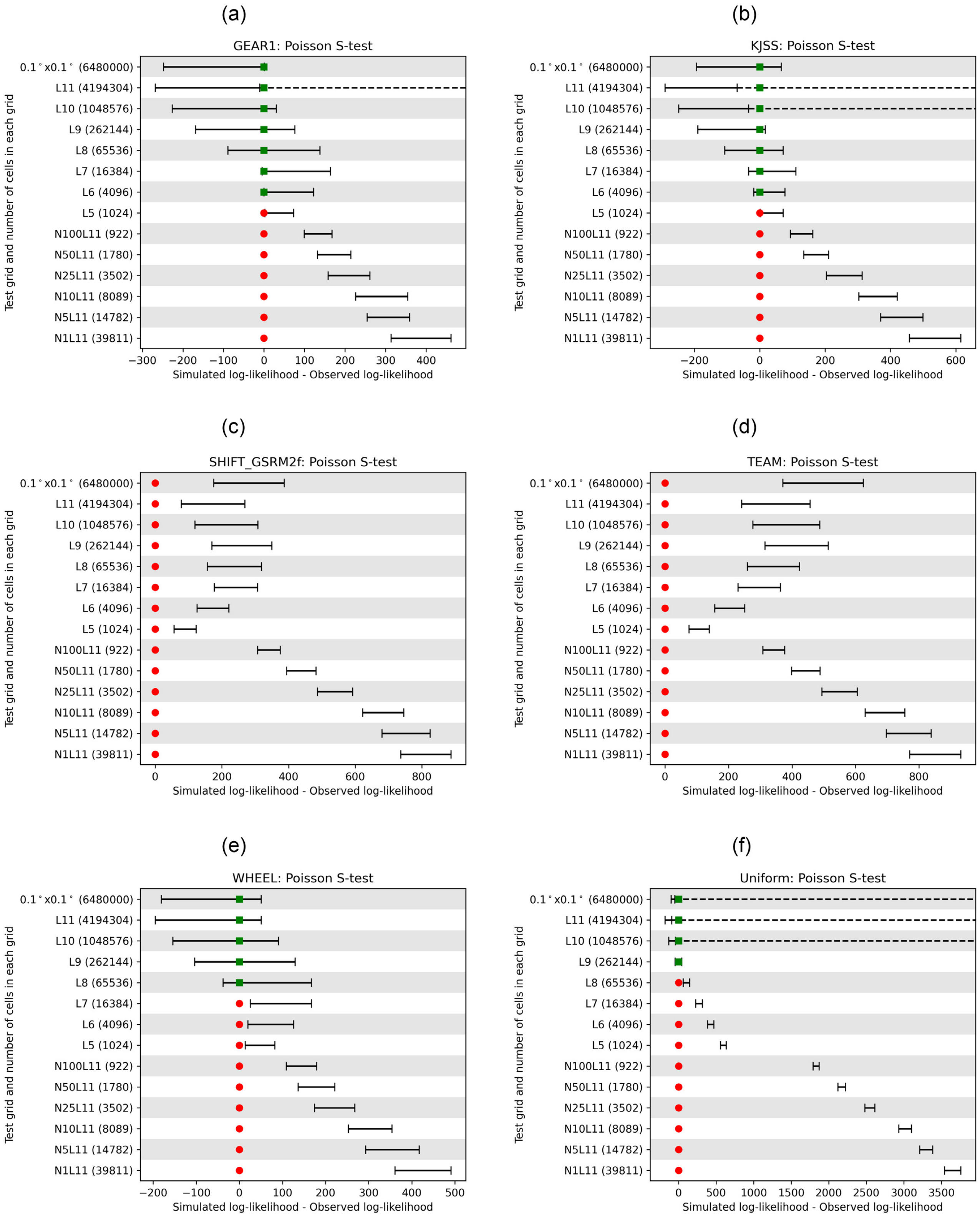

Figure 6. Results of Poisson S-test for earthquake forecast models evaluated at different grids.

powerful and meaningful earthquake forecast experiments. Our investigations show that the earthquake sample size for evaluating forecast models is an important factor contributing to the statistical power of the S-test. However, it is not feasible to sufficiently extend the testing period to achieve a powerful test in the case of a forecast for millions of cells. This study suggests that we would need, in this case, more than 32 000 earthquakes in the observed catalog, which would require approximately 300 yr to record $M \geq 5.95$, 80 yr to record $M \geq 5.5$ and 40 yr to record $M \geq 5.15$ earthquakes on the globe. The other factor that can make the

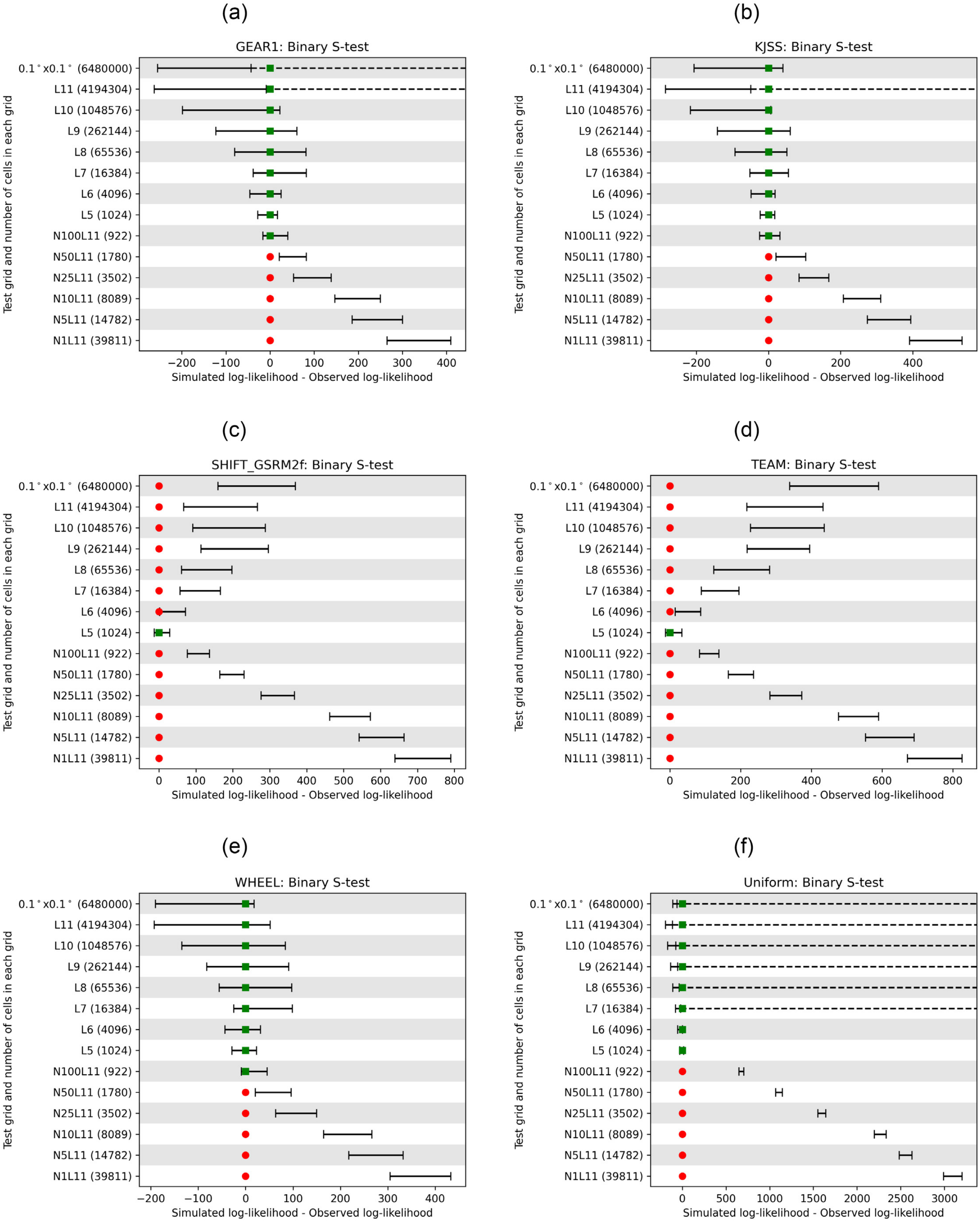

Figure 7. Results of binary S-test for earthquake forecast models evaluated at different grids.

S-test powerful is the representation of earthquake forecast models using different grids. We found that using multi-resolution grids can enhance the statistical power in such a way that instead of 32 000, only eight earthquakes can be sufficient for powerful testing. The technical capability to represent forecast models on different Quadtree grids instead of 0.1° × 0.1° grid is already integrated into the pyCSEP. Additionally, the Quadtree also offers computational advantages and facilitates handling multi-resolution grids (Asim *et al.* 2022).

The trends in statistical power observed in our synthetic experiments are also reflected for all the competing forecast models at different grids in the case of the re-evaluation of the real data. All the forecast models passing the test at the conventional grid fail to pass the same test for some of the other grids. So the question arises what resolution of the grid should be the most suitable resolution for testing the forecast models? Evaluating the forecast models at any single-resolution grid (like a conventional grid) with reduced resolution can be a potential choice for testing earthquake forecast models. However, due to the sparse nature of seismicity, simply reducing the resolution of the grid uniformly everywhere reduces the capability of the grid to capture the spatial variation offered by the forecast models and seismicity. Even after reducing the resolution by a factor of 25 from 6.48 million cells ($0.1^\circ \times 0.1^\circ$) to 0.26 million cells (*L9*), we still need approximately more than 4000 earthquakes in the test catalog for a statistically significant test. So how much should the grid resolution be reduced to maintain the balance between capturing the spatial variation of seismicity, the number of earthquakes required in the test catalog and the statistical power of the S-test? If we further reduce the resolution to grid *L8* with 65536 cells, we still need more than 2000 earthquakes to achieve high statistical power, which may need approximately 20yr of test period to record $M \geq 5.95$ earthquakes. For reliable and statistically powerful testing with approximately 1000 earthquakes, we need a single-resolution grid with less than 16 000 cells (grid *L7*).

Conducting testing experiments using single-resolution grids involves a trade-off. Either we lose statistical power or the capability to capture spatial variations by increasing the size of cells everywhere uniformly, or we have to wait too long for sufficient testing data. Thus, it is desirable to exploit multi-resolution grids to achieve greater power within shorter times.

Asim *et al.* (2022) showed that the multi-resolution data-driven grid is superior for generating an earthquake forecast model. Using the highest available single-resolution grid might not be the best choice for creating forecasts. The available earthquake data can be split into training and test datasets to search for the optimal choice of the grid's resolution to constrain the model's forecast. Thus, we can expect the forecast models to be generated using data-based multi-resolution grids. The present study shows that data-based multi-resolution grids can provide statistically powerful testing with as few as eight earthquakes, which endorses the use of multi-resolution grids also for forecast evaluation. Therefore, we propose using multi-resolution grids to evaluate earthquake forecast models in future CSEP experiments.

We see from Figs 6 and 7 that the S-test is systematically rejecting the forecasts with a higher margin for multi-resolution grids as we locally increase the resolution in the regions of higher seismicity. For example, the S-test rejects forecasts at grid *N100L11* with 922 cells by a smaller margin than for other multi-resolution grids. This margin is increasing as we enhance the ability of the grid to capture the spatial variation of data (earthquakes in this case) by reducing the threshold on the maximum number of earthquakes ($N_{\max}$) allowed per cell during the grid generation process. The largest margins, i.e. the strongest forecast rejections, are obtained for $N_{\max} = 1$, indicating that zooming into the highly active regions offers the most powerful testing.

With the Quadtree approach already integrated into pyCSEP, the forecast modellers can generate multi-resolution grids based on the availability of data at their disposal. A modeller can locally increase the resolution of the grid if higher-resolution data are available for the model's training. Similarly, the resolution of the grid can be reduced for the regions where fewer data are available. Thus, the modellers should make sure that the choice of the grid for every forecast model should reflect the resolution of the data used to create the forecast. Consequently, the provided forecast on the multi-resolution should pass the S-test on the same grid and any aggregations of this grid.

In contrast, Asim *et al.* (2022) suggested that pair-wise comparative testing of earthquake forecast models based on information gain (T-test) should be conducted using the highest available spatial resolution. Thus, to avoid any definition of a particular grid for comparative testing and to save computational resources associated with forecast de-aggregation, the point-process log-likelihood, which is equivalent to the joint likelihood (eq. 3) in the limit of vanishing cell sizes, should be used. In particular, Asim *et al.* (2022) demonstrated that the pair-wise comparison of forecasts with different grids, when tested on a common high-resolution grid, leads to the same results as using the grid-independent point-process log-likelihood. Therefore, the pair-wise testing of competing forecast models on different spatial grids is not an issue.

7 CONCLUSION

Earthquake forecast experiments are currently performed in CSEP for forecasts represented using $0.1^\circ \times 0.1^\circ$ grid. The availability of limited data to evaluate the forecast models causes a huge disparity in the number of earthquakes and the number of cells, leading to statistically powerless testing. It is necessary to have statistically powerful tests for informed decision-making. Thus, we perform a systematic power analysis that can guide the future earthquake forecast experiment. Our analysis reveals that the statistical power of the S-test depends on the available number of earthquakes for testing and the resolution of the spatial grid. To compute the statistical power of the S-test, we substitute the true seismicity model of the earth (as none exists so far) with the GEAR1 forecast model and simulate earthquakes to evaluate the uniform forecast model. Our analysis shows that we need approximately more than 32 000 earthquakes in the global testing region to have a statistically powerful test in the case of the standard $0.1^\circ \times 0.1^\circ$ grid. While we wait so long to achieve a powerful S-test on this grid, we can alter the grid representation of the forecast by changing the resolution, which also affects the statistical power of the S-test. With data-based multi-resolution grids, we can achieve the maximum statistical power of the S-test with as low as eight earthquakes. Therefore, we propose to use multi-resolution grids for future earthquake forecast experiments for evaluating earthquake forecast models.

ACKNOWLEDGMENTS

The authors thank Antonio Morales, Chris Rollins and one anonymous reviewer for reviewing and providing constructive feedback to improve the manuscript. The authors are grateful to friends and colleagues for their valuable suggestions. This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement Number 821115, Real-Time Earthquake Risk Reduction for a Resilient Europe (RISE). This research was also supported by the Southern California Earthquake Center (Contribution No. 12691). SCEC is funded by NSF Cooperative Agreement EAR-1600087 & USGS Cooperative Agreement G17AC00047. In particular, we want to thank the open-source community for the Linux operating system and the many programs used in this study, e.g. Python (https://www.python.org/), Spyder (Raybaut 2009) and PyCharm (https://www.jetbrains.com/pycharm), OpenStreetMap (https://www.openstreetmap.org) and QGIS (https://qgis.org/).

DATA AVAILABILITY

We acquired the global catalog from the CMT webpage (globalCMT; https://www.globalcmt.org/, last accessed December 2021). The Quadtree approach has been integrated as a part of an extensive software package developed for CSEP tests known as pyCSEP. The codes, including documentation and examples, for generating Quadtree grids are available here: https://github.com/SCECcode/pycsep. Finally, the codes to reproduce the results and figures of this manuscript are available here: https://git.gfz-potsdam.de/csep-group/stest-power-analysis/.

AUTHOR CONTRIBUTIONS

This work is a part of the PhD thesis of Asim M. Khawaja. Investigation and research were performed by all authors. Specifically, the ideas and research goals of this manuscript were formulated by Asim M. Khawaja under the leadership of Sebastian Hainzl and Danijel Schorlemmer. Asim M. Khawaja wrote the codes, conducted the statistical analysis and obtained the results. He wrote the original draft under the supervision of all co-authors. All co-authors contributed significantly in many ways, e.g. providing the critical review of the manuscript.

REFERENCES

Ahmad, N., Gurmani, S.F., Qureshi, R.M. & Iqbal, T., 2019. Preliminary results of fair-weather atmospheric electric field in the proximity of Main Boundary Thrust, Northern Pakistan, *Adv. Space Res.,* **63**(2), 927–936.

Asayesh, B.M., Zafarani, H. & Tatar, M., 2020. Coulomb stress changes and secondary stress triggering during the 2003 (mw 6.6) bam (iran) earthquake, *Tectonophysics,* **775,** 228304. https://doi.org/10.1016/j.tecto.2019.228304

Asayesh, B.M., Zafarani, H., Hainzl, S. & Sharma, S., 2022. Effects of large aftershocks on spatial aftershock forecasts during the 2017–2019 western iran sequence, *Geophys. J. Int.,* **232**(1), 147–161.

Asim, K.M., Idris, A., Iqbal, T. & Martínez-Álvarez, F., 2018. Earthquake prediction model using support vector regressor and hybrid neural networks, *PloS One,* **13**(7), e0199004. https://doi.org/10.1371/journal.pone.0199004

Asim, K.M., Schorlemmer, D., Hainzl, S., Iturrieta, P., Savran, W.H., Bayona, J.A. & Werner, M.J., 2023. Multi-resolution grids in earthquake forecasting: the Quadtree approach, *Bull. seism. Soc. Am. (Under review)*.**113 (1),** 333–347. https://doi.org/10.1785/0120220028

Bayliss, K., Naylor, M., Kamranzad, F. & Main, I., 2022. Pseudo-prospective testing of 5-year earthquake forecasts for California using inlabru, *Natural Hazards and Earth System Sciences Discussions,* **22**(10), EGU publications, 3231–3246. https://doi.org/10.5194/nhess-22-3231-2022

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

Bayona, J.A., Savran, W.H., Rhoades, D.A. & Werner, M., 2022. Prospective evaluation of multiplicative hybrid earthquake forecasting models in California, *Geophys. J. Int.,* **229**(3), 1736–1753. https://doi.org/10.1093/gji/ggac018

Bezeau, S. & Graves, R., 2001. Statistical power and effect sizes of clinical neuropsychology research, *J. Clin. Exp. Neuropsychol.,* **23**(3), 399–406.

Bird, P. & Kreemer, C., 2015. Revised tectonic forecast of global shallow seismicity based on version 2.1 of the global strain rate map, *Bull. seism. Soc. Am.,* **105**(1), 152–166.

Bird, P., Jackson, D.D., Kagan, Y.Y., Kreemer, C. & Stein, R., 2015. GEAR1: A global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity, *Bull. seism. Soc. Am.,* **105**(5), 2538–2554.

Bray, A. & Schoenberg, F.P., 2013. Assessment of point process models for earthquake forecasting, *Stat. Sci.,* **28**(4), 510–520.

Button, K.S., Ioannidis, J., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S. & Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience, *Nature Rev. Neurosci.,* **14**(5), 365–376.

Ebrahimian, H., Jalayer, F., Maleki Asayesh, B., Hainzl, S. & Zafarani, H., 2022. Improvements to seismicity forecasting based on a bayesian spatio-temporal etas model, *Sci. Rep.,* **12**(1), 1–27.

Ekström, G., Nettles, M. & Dziewoński, A., 2012. The global CMT project 2004–2010: Centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200,** 1–9.

Helmstetter, A., Kagan, Y.Y. & Jackson, D.D., 2007. High-resolution time-independent grid-based forecast for m 5 earthquakes in California, *Seism. Res. Lett.,* **78**(1), 78–86.

Jordan, T.H., 2006. Earthquake predictability, brick by brick, *Seism. Res. Lett.,* **77**(1), 3–6.

Kagan, Y. & Jackson, D., 2010. Earthquake forecasting in diverse tectonic zones of the globe, *Pure Appl. Geophys.,* **167**(6), 709–719.

Kagan, Y.Y., 2007. Simplified algorithms for calculating double-couple rotation, *Geophys. J. Int.,* **171**(1), 411–418.

Kagan, Y.Y. & Jackson, D.D., 2011. Global earthquake forecasts, *Geophys. J. Int.,* **184**(2), 759–776.

Lehmann, E.L., Romano, J.P. & Casella, G., 2005. *Testing Statistical hypotheses,* Vol. **3,** Springer. ISBN: 978-0-387-27605-2.

Lombardi, A.M., Cocco, M. & Marzocchi, W., 2010. On the increase of background seismicity rate during the 1997–1998 umbria-marche, central Italy, sequence: apparent variation or fluid-driven triggering? on the increase of background seismicity rate during the 1997–1998 Umbria-Marche sequence, *Bull. seism. Soc. Am.,* **100**(3), 1138–1152.

Maleki Asayesh, B., Hamzeloo, H. & Zafarani, H., 2019. Coulomb stress changes due to main earthquakes in southeast iran during 1981 to 2011, *J. Seismol.,* **23**(1), 135–150.

Mancini, S., Segou, M., Werner, M. & Cattania, C., 2019. Improving physics-based aftershock forecasts during the 2016–2017 Central Italy Earthquake Cascade, *J. geophys. Res.: Solid Earth,* **124**(8), 8626–8643.

Martínez-Álvarez, F., Reyes, J., Morales-Esteban, A. & Rubio-Escudero, C., 2013. Determining the best set of seismicity indicators to predict earthquakes. two case studies: Chile and the iberian peninsula, *Knowledge-Based Syst.,* **50,** 198–210.

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

Mignan, A. & Broccardo, M., 2020. Neural network applications in earthquake prediction (1994–2019): Meta-analytic and statistical insights on their limitations, *Seismol. Res. Lett.,* **91**(4), 2330–2342.

Morales-Esteban, A., Martínez-Álvarez, F., Troncoso, A., Justo, J. & Rubio-Escudero, C., 2010. Pattern recognition to forecast seismic time series, *Expert Syst. Appl.,* **37**(12), 8333–8342.

Raybaut, P., 2009. Spyder-documentation, Available online at: pythonhosted.org.

Rhoades, D.A., J Rastin, S. & Christophersen, A., 2020. The effect of catalogue lead time on medium-term earthquake forecasting with application to New Zealand Data, *Entropy,* **22**(11), 1264. https://doi.org/10.3390/e22111264

Savran, W.H., Werner, M.J., Marzocchi, W., Rhoades, D.A., Jackson, D.D., Milner, K., Field, E. & Michael, A., 2020. Pseudoprospective evaluation of UCERF3-ETAS forecasts during the 2019 Ridgecrest sequence, *Bull. seism. Soc. Am.,* **110**(4), 1799–1817.

Savran, W.H. *et al.*, 2022a. pyCSEP: a Python toolkit for earthquake forecast developers, *Seismol. Soc. Am.,* **93**(5), 2858–2870.

Savran, W.H., Werner, M.J., Schorlemmer, D. & Maechling, P.J., 2022b. pyCSEP: a Python toolkit for earthquake forecast developers, *J. Open Source Software,* **7**(69), 3658 doi:10.21105/joss.03658.

Schorlemmer, D. & Gerstenberger, M., 2007. Relm testing center, *Seismol. Res. Lett.,* **78**(1), 30–36.

Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D. & Rhoades, D., 2007. Earthquake likelihood model testing, *Seismol. Res. Lett.,* **78**(1), 17–29.

Schorlemmer, D., Christophersen, A., Rovida, A., Mele, F., Stucchi, M. & Marzocchi, W., 2010. Setting up an earthquake forecast experiment in Italy, *Annals Geophys.*. doi: 10.4401/ag-4844

Schorlemmer, D. *et al.*, 2018. The Collaboratory for the Study of Earthquake Predictability: achievements and priorities, *Seismol. Res. Lett.,* **89**(4), 1305–1313.

Sharma, S., Hainzl, S., Zöeller, G. & Holschneider, M., 2020. Is Coulomb stress the best choice for aftershock forecasting?, *J. geophys. Res.: Solid Earth,* **125**(9), e2020JB019553. https://doi.org/10.1029/2020JB019553

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

Tareen, A. D.K., Asim, K.M., Kearfott, K.J., Rafique, M., Nadeem, M. S.A., Iqbal, T. & Rahman, S.U., 2019. Automated anomalous behaviour detection in soil radon gas prior to earthquakes using computational intelligence techniques, *J. Environ. Radioact.,* **203,** 48–54.

Tariq, M.A., Shah, M., Hernández-Pajares, M. & Iqbal, T., 2019. Pre-earthquake ionospheric anomalies before three major earthquakes by GPS-TEC and GIM-TEC data during 2015–2017, *Adv. Space Res.,* **63**(7), 2088–2099.

Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M.J., Wiemer, S., Zechar, J.D., Heiniger, L. & Euchner, F., 2018. Prospective CSEP Evaluation of 1-Day, 3-Month, and 5-yr Earthquake Forecasts for Italy, *Seismol. Res. Lett.,* **89**(4), 1251–1261.

Tsuruoka, H., Hirata, N., Schorlemmer, D., Euchner, F., Nanjo, K.Z. & Jordan, T.H., 2012. CSEP Testing Center and the first results of the earthquake forecast testing experiment in Japan, *Earth Planets Space,* **64**(8), 661–671.

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

Zechar, J.D., Gerstenberger, M.C. & Rhoades, D.A., 2010. Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts, *Bull. seism. Soc. Am.,* **100**(3), 1184–1195.

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

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1. Global Uniform forecast model created for 0.1° × 0.1° spatial grid, where the forecast rate of the cells is proportional to the area of the cell, along with the 651 earthquake of M_w5.95+ observed in 2014–2019.

Figure S2. Flowchart for computing the statistical power of S-test. The S-test should be able to identify two seismicity models (Λ_1, Λ_2) as inconsistent which are different from each other. '*N*' is the number of simulations carried out to compute the power, '*i*' and '*count*' are initialized as 0 at the start.

Figure S3. Frequency-magnitude distribution of global CMT catalog from 1976 to 2013. The dashed vertical line indicates the cut-off magnitude used in our analysis.

Figure S4. GEAR1 forecast model (colour-coded) aggregated on grid L6 with 4096 spatial cells along with the 651 earthquakes of M_w5.95+ observed in 2014-2019 (points).

Figure S5. GEAR1 forecast model aggregated on grid N50L11 with 1780 spatial cells along with the 651 earthquakes of M_w5.95+ observed in 2014–2019 (points).

Figure S6. Earthquake forecast evaluation using Poisson S-test for different aggregations of the global forecast models shown as non-normalized confidence intervals.

Figure S7. Earthquake forecast evaluation using Binary S-test for different aggregations of the global forecast models shown as non-normalized confidence intervals.

Figure S8. The statistical power of the S-test with TEAM as data-generating model to reject the uniform forecast model for single-resolution grids indicates that the power of the S-test increases for the grids with lower resolution and a larger test sample size.

Figure S9. The statistical power of the S-test with TEAM as data-generating model to reject the uniform model forecasts for multi-resolution grids indicates that with multi-resolution grids, statistically powerful testing can be performed with as few as four or eight earthquakes in the observed catalog.

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