Eur. Phys. J. Special Topics **230**, 451–471 (2021)

https://doi.org/10.1140/epjst/e2020-000260-2

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Democratizing earthquake predictability research: introducing the RichterX platform

Yavor Kamer[1,a], Shyam Nandan[1,2], Guy Ouillon[3], Stefan Hiemer[1], and Didier Sornette[4,5]

[1] RichterX.com, Mittelweg 8, Langen 63225, Germany
[2] Windeggstrasse, 5, 8953 Dietikon, Zurich, Switzerland
[3] Lithophyse, 4 rue de l'Ancien Sénat, 06300 Nice, France
[4] ETH Zurich, Department of Management, Technology and Economics, Scheuchzerstrasse 7, 8092 Zurich, Switzerland
[5] Institute of Risk Analysis, Prediction and Management (Risks-X), Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology (SUSTech), Shenzhen, P.R. China

Received 5 October 2020 / Accepted 7 October 2020
Published online 19 January 2021

Abstract. Predictability of earthquakes has been vigorously debated in the last decades with the dominant -albeit contested -view being that earthquakes are inherently unpredictable. The absence of a framework to rigorously evaluate earthquake predictions has led to prediction efforts being viewed with scepticism. Consequently, funding for earthquake prediction has dried out and the community has shifted its focus towards earthquake forecasting. The field has benefited from collaborative efforts to organize prospective earthquake forecasting contests by introducing protocols, model formats and rigorous tests. However, these regulations have also created a barrier to entry. Methods that do not share the assumptions of the testing protocols, or whose outputs are not compatible with the contest format, can not be accommodated. In addition, the results of the contests are communicated via a suite of consistency and pair-wise tests that are often difficult to interpret for those not well versed in statistical inference. Due to these limiting factors, while scientific output in earthquake seismology has been on the rise, participation in such earthquake forecasting contests has remained rather limited. In order to revive earthquake predictability research and encourage wide-scale participation, here we introduce a global earthquake prediction platform by the name RichterX. The platform allows for testing of any earthquake prediction in a user-defined magnitude, space, time window anywhere on the globe. Predictions are assigned a reference probability based on a rigorously tested real-time global statistical earthquake forecasting model. In this way, we are able to accommodate methods issuing alarm based predictions as well as probabilistic earthquake forecasting models. We formulate two metrics to evaluate the participants' predictive skill and demonstrate their consistency through synthetic tests.

[a] e-mail: yaver.kamer@gmail.com

1 Introduction

Earthquake prediction is a hard problem, which has remained an elusive holy grail of seismology. Unfortunately, the current incentive structures are pushing researchers away from hard problems where results are rarely positive. Negative results are less likely to lead to a publication or a citation. While the utility of these quantities is being put into question [8], they are still widely used as performance criteria in academia.

To avoid negative results and public reactions associated with failed earthquake predictions, the seismological community has mainly shifted its focus to descriptive case-studies, long-term probabilistic hazard analysis, and probabilistic forecasting experiments. However, by not engaging the prediction problem, we have effectively left it to be exploited by less reputable actors. These actors often emerge during times of crisis, spreading disinformation leading to public anxiety. As a result, it has become common to view any sort of prediction effort with suspicion and often negative prejudice, forgetting that the scientific principle requires hypotheses to be tested rather than disregarded due to prior held beliefs. It can be argued that many of these prediction claims are not formulated as falsifiable hypotheses, yet it is our duty as scientists to assist those interested by providing guidelines, protocols, and platforms facilitating the scientific method.

To help revive the earthquake prediction effort and bring scientific rigor to the field, here we propose a general platform to facilitate the process of issuing and evaluating earthquake predictions. The platform is general as it allows for the testing of both deterministic alarm based predictions and probabilistic forecasts. The common metrics proposed to evaluate the respective skills of these two different model classes will put methods relying on different theories, physical mechanisms and datasets on the same footing. In this way, we aim to achieve larger participation, to facilitate the inclusion of different methods from various fields, and to foster collaborative learning through regular feedback.

The paper is structured as follows. First, we introduce the general requirements that a public earthquake prediction platform must satisfy and briefly explain how RichterX addresses these. Second, we describe the implementation of our global real-time earthquake forecasting model that the RichterX platform uses to inform the public of short term earthquake probabilities, and that is also taken as a reference to evaluate all submitted predictions. Next, we introduce two complementary performance metrics that allow us to assess the predictive performance of the participants. Finally, we conduct synthetic contests with known ground truth and candidate models to test the consistency of the proposed metrics.

2 Characteristics of the earthquake prediction platform RichterX

Participation: Considering that earthquakes have a huge global impact, previous and current forecasting experiments have reached out to only a small number of participants [41,42]. Our platform aims to attract broader participation, not only from the relatively small seismology community in academia but also other scientific disciplines including fields like machine learning, pattern recognition, data mining, remote sensing, etc., active in information technologies and engineering applications. The platform also encourages and rewards public participation to increase public awareness of the earthquake hazard, motivate prediction efforts, and, more importantly, allow citizens to participate in a scientific challenge. Previous prediction experiments have been criticized in this regard because they have treated the public as mere subjects of a scientific experiment and sometimes as means to higher ends (i.e increased

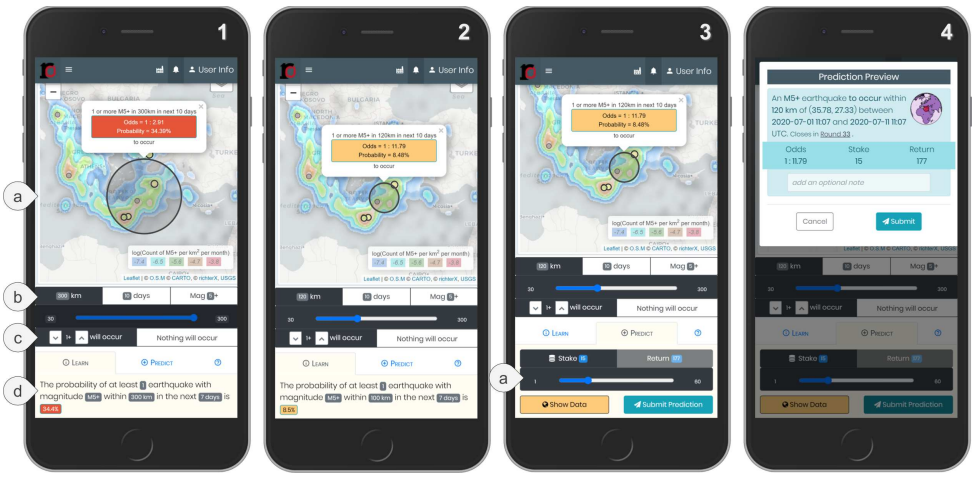

Fig. 1. The RichterX platform accessible at www.richterX.com, as viewed on a mobile phone. (1) Forecast screen (a) Map colors indicate the monthly M5 earthquake count; black circle represents the target area of the prediction; pop-up message reports the probability according to the RichterX model; (b) Three tabs with a slider for adjusting the radius, time duration and minimum magnitude of the prediction; (c) Toggle button to switch between probabilities *to-occur* and *not-to-occur*; the number of events to-occur can be specified via the up/down arrows; (d) Summary of the RichterX forecast in human-readable format. (2) Forecast at the same location with radius reduced from 300 km to 100 km. (3) Forecast screen with the *Predict* toggle on: (a) Slider for setting the prediction stake. (4) Prediction preview screen showing a summary and the round at which the prediction closes.

public awareness) [43]. Participating in a global earthquake prediction contest will allow the public to gain hands-on insight, internalize the current achievements and difficulty of the problem. To achieve this, we have built a minimal graphical user interface, compatible with desktop devices as well as most mobile phones, allowing anyone to participate (see Fig. 1). We also provide an application programming interface (API), allowing more sophisticated participants to submit predictions using a computer algorithm. Moreover, we see the language barrier as one of the main factors hindering participation. We will, therefore, make the platform and the relevant publications available in multiple languages.

Privacy: The negative connotation associated with failed predictions is an important factor hampering prediction efforts. The RichterX platform provides the participants with the option to anonymize their identity, allowing them to focus on the scientific question instead of worrying about the possible loss of reputation.

Transparency: Results and conclusions of any forecasting or prediction contest must be accessible to the general public. The results of previous forecasting contests such as CSEP and RELM have been published, but these papers are often behind paywalls. The CSEP public website containing results of several models and tests, although rather technical and not very intuitive for the general public, has since gone offline. In our view, transparency and ease of access to contest results, reinforce responsibility and accountability.

Assuming that science is conducted to enhance public utility, the public is entitled to know of its progression, which entails not only successes but also failures. Thus, we are committed to making the results openly available to the public. In addition to results about each earthquake (whether it was predicted or not), metrics regarding the overall performance of each participant are updated on an hourly basis and in

the form of regularly issued public reports. The provision of this information will counter false prediction allegations, serve as a verifiable track record, and allow the public to distinguish between one-time guesses and skilled predictions.

Global coverage: Earthquakes do not occur randomly in space, but cluster on tectonic features such as subduction zones, continental plate boundaries, volcanic regions and intraplate faults. These active features span across the whole globe and produce large earthquakes continuously. Previous forecasting experiments have focused mainly on regions with very good instrumental coverage, available only in a small number of countries (USA, Japan, New Zealand, Iceland, Italy, etc) [42]. Our goal is to foster a worldwide earthquake prediction effort by providing the community and the public with a global reference earthquake forecasting model. With the help of such a reference model, our platform will be able to accommodate any regional model by evaluating it against the same global baseline, putting regional added value in a global perspective.

Real-time updates: Temporal clustering is another main feature of earthquake occurrence: the probability that an earthquake will occur in a given space window can vary greatly in time. Thus, if a prediction is to be evaluated according to a reference model probability, such a reference model should be updated in near real-time as soon as a new event occurs. Together with global coverage, this requirement poses serious computational demands that have hindered the implementation of such models. Recent advances in the field of statistical earthquake modeling [17,24,33,34] have allowed us to undertake this challenge. Having secured the computational and hosting capabilities, the RichterX platform is able to provide the global community with worldwide earthquake forecasts updated on an hourly basis.

Active learning: The reference model provided on the RichterX platform aims to reflect the current state-of-art in statistical seismology. Hence it is not set in stone but is subject to further improvements as the participants, through successful predictions, effectively highlight regions and time frames where the model performance is lacking. In this way, the participants serve as referees continuously peer-reviewing the reference model, which thereby is permanently improving, providing the community with a higher bar to surpass.

Feedback mechanism: The goal of our prediction experiment is to provide the participants with meaningful feedbacks, allowing them to test their hypotheses, models, assumptions, and auxiliary data. Through repeated iteration of submitting predictions, testing, and receiving feedback, we expect the participants to improve their prediction performance. For the public observers, the results should be presented transparently and succinctly, allowing for an intuitive comparison of the participants' performances. Therefore, we have developed a skill assessment scheme that is both easy to understand for the public and powerful in distinguishing between different predictive skills. It is important to note that the participants may lose interest if the experiment takes too long to deliver results. The provided feedback may also lose its relevance if not provided in a timely fashion. The RichterX platform issues results on a bi-weekly basis and cumulative metrics spanning longer durations. In contrast, consider that previous earthquake forecasting experiments by CSEP were carried out for 5 years [27], with preliminary results being released only after 2.5 years [41].

Incentives: We hope that the opportunity to easily test and compare different hypotheses and models on a global scale would provide enough stimulus for the academic community. At the same time, it is important to recognize that science can be costly. Apart from the devoted time, many published studies are behind paywalls, data processing requires expensive hardware, and some data sources can be subject to fees. Thus, we believe amateur scientists, students, and the general public can

be incentivized to participate by providing rewards, with "scientific microfundings" similar to microcredits in the business field. These can be monetary or in the form of technical equipment or journal subscriptions. Some studies have raised concerns that improper use of monetary rewards can reduce the intrinsic motivation of the participants [28]. However, recent studies have shown that financial rewards have a positive effect on engagement and satisfaction [4,12]. The delivery of such monetary rewards is now much easier due to the popularization of crypto-currencies [10,26]. These recent developments allow us to financially transact with the successful participants without requiring a bank account, which almost half of the world's population does not have access to [5].

Social responsibility: It is essential to recognize that earthquake prediction is not only a scientific goal but also a topic that has the potential to affect the lives of many people, especially those living in seismically active regions. The contest participants should be aware that the events that they are trying to predict are not just numbers on a screen, but actual catastrophes causing human suffering. We believe that providing a mechanism for expressing solidarity with the victims can help raise this awareness. To this end, the RichterX platform encourages the participants to donate their rewards to charitable organizations such as GiveWell, Humanity Road [44] and UNICEF [11], which take part in global earthquake relief efforts. Recent studies indicate that the use of decentralized ledger technologies can improve transparency and accountability in humanitarian operations [7]. Therefore, all donations on RichterX are made using cryptocurrencies and recorded on the blockchain, allowing for anyone to verify the amount and destination independently. In this way, we hope to prevent a possible detachment between a community that engages with earthquakes from a scientific perspective and people who suffer their physical consequences.

3 A global, real-time reference earthquake forecasting model

3.1 Introduction

The characteristics summarized in the previous section have emerged due to the experience gained from previous earthquake prediction and forecasting experiments. In his address to the Seismological Society of America in 1976, Clarence Allen proposed that an earthquake prediction should be assigned a reference probability indicating how likely it is to occur by chance [2]. Indeed, the development of a model that can assign a probability for any time window anywhere in the world has been one of the main hurdles. There have been several accomplishments in the modeling of global seismicity. Those efforts began with models based on smoothing locations of observed large earthquakes [22], progressing to combining past seismicity with strain rates estimates [3,21]. Recently, the Global Earthquake Model working group led a collaborative effort to harmonize many regional models [39]. Although these models are important milestones, they model seismicity as a memoryless, stationary process. As a result, they do not capture the time-dependent aspect of earthquake occurrence. The choice of treating earthquakes as a stationary process likely is motivated by the risk assessment practices in the civil engineering and insurance industry. Yet, we believe that, as the seismology community reassesses its assumptions and develops more realistic models, the industry will, in turn, adapt to these changes.

Based on empirical laws derived from observed seismicity, the Epidemic Type Aftershock Sequence (ETAS) model was introduced to enhance stationary statistical earthquake models by accounting for the time-space clustering of seismicity [38]. Retrospective and prospective studies show that statistical models outperform models derived from physical concepts such as rate-and-state, stress transfer, seismic gaps,

or characteristic earthquakes [6,23,48]. The recent developments in ETAS modeling have not only made a global scale application possible, but they have also highlighted the importance of abolishing assumptions about the distribution of simulated events [34,35]. Details about the model development, testing, and prospects can be found in the accompanying paper [36]. Here we describe the real-time online implementation and operation of the model in the context of the platform.

3.2 Data

The RichterX platform employs a dedicated server, the so-called "grabber", that periodically connects to web-based earthquake data center feeds. The grabber compares our current local database with the remote host for the addition of new or the deletion of old events. If any change is detected, the grabber synchronizes our current event catalog with the remote database. We are obtaining data from multiple global agencies, such as the GFZ Geofon [14] and INGV Early-Est [29], but our primary data source is the USGS ComCat feed [45]. Our data polling frequency is usually around once every few minutes but can be increased automatically during elevated seismic activity.

3.3 Model selection, calibration, and forward simulations

We have developed multiple candidate models that are different variations of the ETAS model or use different input datasets. We use pseudo-prospective testing to select among these models. See details of the experiment and competing models in the accompanying paper [36]. In this procedure, only data recorded before a certain time is considered and divided into sets of training and validation; competing models are trained on the training data, and their forecasting performances are compared on the validation set. The performances are averaged by moving forward in time and repeating the tests. The top-ranking model, and its final parameter set optimized over the whole dataset, is deployed online on the platform servers. These servers use the real-time earthquake event data provided by the grabber as input and conduct forward simulations on an hourly basis. The result of these forward simulations is a collection of synthetic event datasets that represent a spatio-temporal projection of how global seismicity will evolve.

The ETAS model is stochastic, i.e., its output samples statistical distributions, and therefore multiple forward simulations are needed to obtain an accurate representation of the underlying probabilities. Each such simulation produces a global synthetic catalog containing location, time, and magnitudes of events for the next 30 days. The total number of events produced at each real-time update can reach several millions. These simulated events are uploaded onto our online database, where they can be queried via the web-based user interface on www.richterX.com. Using this interface, the participants can select any point on the globe, define a circular region, a time window, and a magnitude range that they are interested in (see Fig. 1). The full distribution of the simulated events within the user-specified time-space-magnitude range is then used to calculate the probability of earthquake occurrence. In essence, this probability corresponds to the number of simulations having events satisfying the user-defined criteria divided by the total number of simulations.

To cope with the computational demands of these simulations, we have scheduled several servers to run periodically in a staggered fashion. In this way, we can assure that the model forecasts are updated within less than an hour after each earthquake. The servers are also distributed in different locations to add redundancy in case of service interruptions.

4 How RichterX works

Earthquake predictions and forecasts have been issued and studied for decades. However, there is still confusion about their definition and proper formulation. We believe it is essential to be strict about terminology. Science advances by accumulating evidence in support or against hypotheses, and vague statements can become a missed opportunity for testing and obtaining such evidence.

4.1 Earthquake forecast and earthquake prediction

We define an earthquake forecast as the statement of a probability that a minimum number of events will occur within a specific time-space-magnitude window. Therefore, a statement cannot be regarded as an earthquake forecast if either one of these four parameters is omitted. For instance, the operational aftershock forecasts issued by the USGS do not specify a space window for the issued probabilities (Field et al., 2014; USGS, 2019), and therefore cannot be tested. Similarly, any ambiguity in the parameters also renders the statement untestable. For instance, the statement "The probability (that the San Andreas fault) will rupture in the next 30 years is thought to lie somewhere between 35% and 70%" [20] does not satisfy the forecast definition because both rupture size and occurrence probability are ambiguous. Unfortunately, this is a common malpractice, and public officials often communicate probabilities by giving a range [9]. The range is usually due to several models or different scenarios leading to different probabilities. Using different approaches and assumptions is to be encouraged; however, the resulting probability should be communicated as a single value. There exist various techniques on how models can be weighed and ensembled according to their predictive performances [13].

We define earthquake prediction as the statement that a minimum number of earthquakes will, or will not occur in a specific time-space-magnitude window. Under our definition, an earthquake prediction always results in a binary outcome: it is either true or false. This definition is more general than its commonly used predecessors [18,20,49] because it considers the negative statement, that an earthquake will not occur, as an equally valid earthquake prediction. By construction, if the probability of an earthquake to occur in a space-time-magnitude window is P, the probability of an earthquake not to occur is $1\text{-}P$. While P is often small, it can exceed 0.5 immediately after large earthquakes or during seismic swarms. In such cases, a prediction of no occurrence carries more information potential, as it refers to a more unlikely outcome. In this way, negative predictions can serve as a feedback mechanism that counters overestimated earthquake forecast probabilities.

Once an earthquake prediction is issued, it is considered to be in a pending (i.e. open) state. The "*To-occur*" predictions, which predict the occurrence of an event or events, are closed as true if the number of predicted events is observed in the predefined space-time-magnitude window, or as false if otherwise. The "*Not-to-occur*" predictions, which predict that no event will occur, are closed as true if there are no events in their predefined space-time-magnitude windows, or as false if at least one such event occurs.

The definitions of earthquake forecast and earthquake predictions are similar, as they both refer to a constrained space-time-magnitude window. They differ in that the former conveys the expected outcome with a real number while the latter uses a binary digit. In that sense, regardless of the observed outcome, a forecast carries more information compared to a prediction. Forecasts are also more straightforward to evaluate; any set of independent earthquake forecasts (i.e having non-overlapping space-time windows) can be evaluated based on the sum of their log-likelihood, which

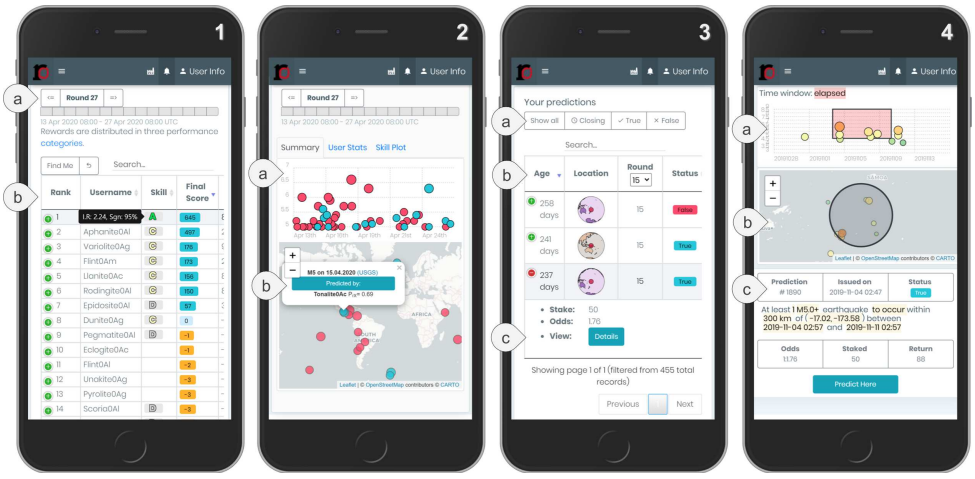

Fig. 2. (1) Ranks screen showing the scores for a selected round: (a) round beginning and end dates; (b) table showing anonymized user names, skill class and current rX score, see Section 3.3 for details. (2) M5+ target events colored as blue for *predicted* and red for *not-predicted*: (a) magnitude vs time plot; (b) spatial distribution of the events. (3) Results screen for a participant: (a) filter criteria; (b) results table list of prediction locations, round number and status; (c) expandable row with further details. (4) Prediction details screen: (a) magnitude-time plot highlighting the elapsed portion of prediction window with red; (b) spatial distribution of events around the prediction circle; (c) prediction statement with space-time-magnitude and event number details.

is analogous to their joint likelihood:

$$LL = \sum_{i=1}^{N} \log\left(O_i P_i + (1 - O_i)\,(1 - P_i)\right) \tag{1}$$

where N is the total number of forecasts, P_i denotes the probability of each forecast and O_i represents the outcome as 1 (true) or 0 (false). The larger the sum, the higher the joint likelihood and hence the more skillful a forecast set is. The performance evaluation of prediction sets is covered in the Performance Assessment section.

4.2 Rules and regulations

The goal of the RichterX platform is to organize a prediction contest that provides timely feedback, skill assessment, and incentives for participation. Therefore, we have tried to devise a system that fosters collaborative competition and rewards skill while maintaining fairness. To attract broader participation, we tried to make the contest's regulations intuitive and straightforward without sacrificing statistical rigor. Here, we will present these rules and the reasons behind them.

4.2.1 Limited number of predictions

Each participant is allowed to place a maximum of 100 predictions every 24 hours. This prediction budget is expressed in units of so-called "earthquake coins" (EQC). It is recharged continuously in real-time, such that after ∼15 minutes, the participant

accumulates 1 EQC and can submit another prediction. The accumulated budget cannot exceed 100 EQC. Hence if participants want to submit more predictions, they have to wait. In this way, we hope to encourage the participant to engage with the platform regularly and follow the evolution of seismicity and think thoroughly before using it to submit predictions. We expect the participants to perceive their limited prediction budget as valuable, since it is scarce.

4.2.2 One user – one account

Each participant is allowed to have only one account on the platform. Since we are providing monetary rewards as an incentive for public participation, users could increase their chance of getting a reward by creating multiple accounts and placing random predictions. We have addressed this by requiring each user to validate their account via a mobile phone application, i.e., a chatbot. The bot runs on the messaging platform Telegram and verifies the user by requiring them to enter a secret code. If the code is correct, the user is matched with their unique Telegram ID, which requires a valid mobile phone number. All reward-related operations are verified through this unique ID.

It is important to note that policies limiting participation rate and preventing multiple accounts are common in online courses and contests such as Kaggle [37,47]. However, previous earthquake forecasting competitions conducted by CSEP, and also its upcoming second phase CSEP2 [40], do not impose such policies. As a result, participants who submit several versions of a model can increase their chance of obtaining a good score, creating a disadvantage for participants who submit only a single model.

4.2.3 Submitting earthquake predictions

The user interface provided on the RichterX platform allows the participants to query our global reference model and obtain its forecasts within the following ranges: time duration from 1 to 30 days, a circular region with radius from 30 to 300 km, lower magnitude limit from M5+ to M9.9+ and a number of events from 1+ to 9+. Once these parameters are set, the platform will report a probability of occurrence P (or non-occurrence $1-P$). The participant can then submit a prediction assigned with this model probability. This probability is used to assess the participant's prediction skill by accounting for the outcome of their closed predictions.

In addition to the time, space, magnitude, and number parameters, the user can also specify a so-called "stake" for each prediction. The stake acts as a multiplier allowing the participants to submit the same prediction several times, provided that it is within their prediction budget (EQCs). Therefore the stake can be thought of as a proxy for the confidence attributed to a prediction.

The reference model updates automatically on an hourly basis. Thus, when a new earthquake occurs, the region in its vicinity becomes unavailable for submitting predictions. Once the new earthquake is incorporated as an input and the model has been updated, the region becomes available for the submission of new predictions. This allows us to fairly compare users and our reference model, as both parties are fed with the same amount of information. The radius of the blocked area (R_b) scales as a function of the event magnitude according to the empirical magnitude-surface rupture length scaling [46] given in the following equation.

$$R_b = 10 + 10^{-3.55+0.74M}(\text{km}) \tag{2}$$

This assures that the projection of the fault rupture, where most aftershocks are expected to occur, remains within the restricted region regardless of the rupture direction. The additional 10 km in the R_b term accounts for the typical global location uncertainty.

4.2.4 Evaluation of earthquake predictions

Target events are all $M \geq 5$ events, as reported by the USGS ComCat [45]. Predictions are evaluated on a bi-weekly round basis. A time frame of only 14 days may seem too short, yet our target region is the whole Earth rather than a specific locality. To put this in perspective, the first regional forecasting experiment, the Regional Earthquake Likelihood Models (RELM), was limited to the state of California, USA, took place during a 5 year period of 2006–2010 and had a total of 31 target events [27]. This corresponds to roughly half of the global bi-weekly M5+ event count (mean 63, median 58 since 1980).

5 Performance assessment metrics

5.1 Conditions for proper metrics

In the case of probabilistic forecasts, a scoring rule is said to be proper if it incentivizes the forecaster to convey their actual estimated probability [19]. In other words, a proper scoring rule does not affect the probability issued by the forecaster. An improper scoring rule, however, can be exploited by modifying one's forecast in a certain way specific to the scoring rule. For example, if a scoring rule does not penalize false positives, then participants can gain an advantage by issuing more alarms than they usually would have. Deterministic predictions do not convey the information of probability; thus, the definition of properness given above becomes irrelevant [19]. Yet it is useful to consider a more general definition: a proper scoring rule, in the context of a contest, aligns the goals of the organizers and the incentives of the participants.

One goal of the RichterX platform is to encourage broad public and academic participation from various fields of expertise. Therefore, we need a scoring rule that is statistically rigorous, easy to understand, and applicable on short time scales. The scoring rule should also ensure that the public participants are rewarded proportionally to their predictive skills, as opposed to a winner-takes-all approach, while incentivizing their regular participation. Another important goal is to provide the scientific community with a generalized platform where different models and hypotheses (be it alarm based or probabilistic) can be evaluated to determine performance and provide feedback to researchers. For this second goal, the scoring rule needs to be exclusively focused on skill and be generally applicable. To achieve both goals, we have chosen to implement two scoring strategies that complement each other. These are the RichterX score and the information ratio score. In the following section, we will describe how these two scores are implemented and used jointly in the competition.

5.2 RichterX Score (rX)

The definition of the rX score is straightforward; each submitted prediction counts as a negative score equal to the prediction stake (s). If a prediction comes true, the

value of the stake(s) multiplied by the odds ($1/p$) is added to the score; if a prediction fails, the score remains at $-s$:

$$\Delta R = \begin{cases} s\left(\frac{1}{p}\right) - s & \text{if true} \\ -s & \text{if false} \end{cases} \tag{3}$$

This can be rewritten as

$$\Delta R = O\left(\frac{s - sp}{p}\right) + (O - 1)\, s \tag{4}$$

where $O = 1$ if the prediction is true and $O = 0$ if false.

Our goal is to incentivize the participants to challenge our model and highlight regions or time frames where it can be improved; thus, we want to reward those who perform better than our reference model. The expected gain from any prediction, according to the model, can be written as the probability-weighted sum of wins and losses:

$$E\left[R\right] = p\left(s\left(\frac{1}{p}\right) - s\right) + (1 - p)\,(-s) = 0. \tag{5}$$

The expected gain of a participant is thus zero (positive scores indicating a better performance than our model). The significance of a positive score (i.e. the probability for a participant to improve on it by chance assuming that the model is correct) has to be estimated for each participant. The latter performance estimator will become more reliable as a participant accumulates independent submitted predictions.

At the end of each bi-weekly contest round, each participant's score is calculated using all their N predictions closed during the round. Thus, summing expression (4) over all N predictions yields:

$$R = \sum_{i=1}^{N} O_i \left(\frac{S_i - S_i P_i}{P_i}\right) + (O_i - 1)\, S_i \tag{6}$$

where S_i is the stake, P_i is the prediction probability given by the reference model, and O_i is a binary variable representing the prediction results as 1 (true) or 0 (false). Monetary reward is distributed proportionally among all participants with positive scores.

The scores are reset to 0 at the beginning of each round to encourage new participants to join the competition. However, resetting the scores each round introduces a problem. Since the only participation cost is the invested time, participants who see that their negative scores are reset at the beginning of each round are incentivized to make low-probability/high reward predictions, especially towards the end of the round. If a few such predictions come true, the user can get a positive score, and if they end up with a negative score, the participants would just have to wait for the next round for their scores to be reset, and then they can try again. This would be problematic because the participants can start treating the experiment as a game of chance with no penalty for false predictions, rather than a contest of skill.

To counter this, we apply a carry-over function that introduces a memory effect for negative scores: if a participant has a negative score not less than -100 at the end of the round, they carry-over 10% of this negative score to the next round as a penalty. The carry-over percentage increases proportionally with the amount of the negative score and caps of at 90% as given in the following equation:

$$\Delta C_t = \begin{cases} \max\left\{|\Delta R_{t-1}|/1000, 0.9\right\} \Delta R_{t-1} & \text{if } \Delta R_{t-1} < -100 \\ 0.1 \Delta R_{t-1} & \text{if } 0 > \Delta R_{t-1} \geq -100 \end{cases} \tag{7}$$

Hence, a participant with a score of −200 would carry over a penalty of −40, while a user with −1000 would carry over −900 to the next round. In this way, participants are incentivized to obtain positive scores consistently, instead of intermittently. Nevertheless, since predictions can be submitted at any time, a participant may stop submitting new predictions as soon as they have reached a positive score. This problem, which has already been discussed by [19], is somewhat alleviated by distributing the reward proportionally to the participant's score with respect to the combined total of all other positive participants. Therefore, there is always an incentive to continue participating as other users become positive and start claiming larger portions of the fixed overall reward.

Another aspect of the rX score is that it is a function of the prediction stake. Two participants with the same predictions but different stakes will get different scores. Assuming they possess some information gain, regular participants will be able to submit the same prediction repeatedly, thereby increase their stake, and obtain higher scores compared to those who follow a different strategy, testing their predictions regardless of their returns. This makes sense in the context of a competition where the participants provide added value by testing our reference model through their predictions. Yet, as we are not necessarily interested in the optimization of staking strategies, there is also a need to assess the predictive skill of each participant regardless of their staking weights. For this purpose, we employ a second metric.

5.3 Information ratio score (IR)

To assess the predictive skill of a participant, we need to answer the following two questions: Firstly, how much better is the participant's performance compared to the reference model, and secondly, is this performance significant. To answer the first question, we calculate a metric called the "information ratio" (IR):

$$\mathrm{IR} = \frac{\frac{1}{N}\sum_{i=1}^{N} O_i}{\frac{1}{N}\sum_{i=1}^{N} P_i} \tag{8}$$

IR is essentially the participant's success rate (fraction of true predictions among all predictions) divided by the reference model probability averaged over all predictions (i.e., the model's expected success rate) of the participant. This formulation implies that there is an upper bound of $\mathrm{IR} = 1/\min(P_i)$ and incentivizes the participants to achieve higher success rates in regions and time frames for which the model gives low probabilities. Assuming that the reference model is true, the expected IR value for any set of predictions would tend to 1.

To answer the question of whether a participant's IR is statistically significant, we employ Monte Carlo sampling to build an IR distribution given their set of submitted predictions. This distribution is independent of the actual prediction outcomes as we sample the model probability of each prediction P_i to generate several possible outcomes O'_i

$$\begin{aligned} x_i &\in U(0,1) \\ O'_i &= \begin{cases} 1 & \text{if } x_i < P \\ 0 & \text{if } x_i \geq P_i \end{cases} \end{aligned} \tag{9}$$

where $U(a,b)$ is the uniform distribution within bounds a and b. We then calculate the IR_m of each outcome set according to equation (8), where m denotes the index of

the Monte Carlo sample. This forms the null-distribution that is used to benchmark the actual IR value of the participant. The ratio of the sampled model IR values that are above or equal to the participant's value (α) can then be interpreted as the probability of observing an IR at least as high as the participant's, i.e., the p-value under the null hypothesis that the reference model is true:

$$\begin{aligned} g_m &= \begin{cases} 1 & \text{if } \mathrm{IR}_m \geq \mathrm{IR}_u \\ 0 & \text{if } \mathrm{IR}_m < \mathrm{IR}_u \end{cases} \\ \alpha &= \frac{1}{M} \sum_{m=1}^{M} g_m \end{aligned} \tag{10}$$

where IR_u is the participant's information ratio, and M is the number of Monte Carlo samples used to sample the distribution. If $\alpha \leq 0.05$, then the participant is considered to be significantly better than the reference model.

5.4 Accounting for overlapping predictions

In all previous equations, we have assumed that the submitted predictions are independent, i.e., that they do not overlap in space and time. This assumption simplifies the derivations of expected success rate, allowing for probabilities of independent predictions to be averaged, and also makes it easier to calculate significance levels by sampling each prediction independently during the Monte Carlo procedure. However, the participants are free to submit predictions at anytime, anywhere on the globe. Since predictions are submitted as circles with a maximum radius, participants who want to cover larger areas completely will have to submit several overlapping predictions. We also see that some participants re-issue predictions at the same locations when an earthquake does not occur (assuming some local stress accumulation or a characteristic period) or when it occurs (expecting aftershocks). Updating a hypothesis as new information becomes available is the hallmark of the scientific method. In the ideal case, if a precursory signal becomes gradually more prominent as an earthquake approaches, one can expect overlapping predictions with narrower space-time windows to be issued. Therefore instead of constraining the participants by forbidding overlapping predictions, we prefer to deal with such predictions.

The question of evaluating overlapping predictions has been investigated previously by Harte and Vere-Jones, Harte et al. [15,16], who introduced the entropy score as a pseudo-likelihood to evaluate M8 predictions, which are expressed as a set of overlapping circles [25]. The entropy score is rather complicated and "awkward", as the authors put it, thus we have refrained from using it as we would like to keep the performance criteria as intuitive as possible for the general public. The Molchan diagram, which accounts for the total time-space volume covered by prediction alarms, can also be employed to deal with predictions overlapping in space and time [31,32]. It is worth noticing that Molchan and Romashkova [30] successfully adopted their methodology to the M8 predictions using specific features, such as constant large circle sizes and large magnitudes, to assess its predictive skill. This is rather different from our application, which involves evaluating and comparing different sets of predictions that can each be a mix of to-occur and not-to-occur, with varying circle sizes.

For the particular case of the RichterX prediction contest, the rX score is additive and already incorporates the concept of "stake" that has the same effect as re-issuing the same prediction; thus, it does not require any modification. However, the overlapping predictions constitute a problem for the IR score and its significance α. This can be seen with a simple example of two non-overlapping to-occur predictions that

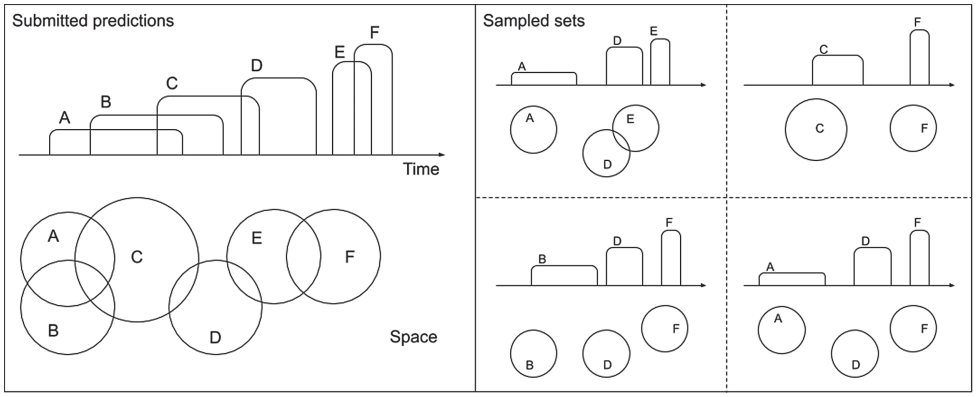

Fig. 3. *Left*: a set of overlapping predictions showing time and space domain. *Right*: a sample of 4 sets containing only non-overlapping predictions obtained by the selective sampling procedure described in the text.

require two earthquakes to come true. In comparison, two identically overlapping predictions would come true with a single event. Intuitively, it follows that true independent predictions are "worth" more in terms of significance than overlapping ones. To take into account the presence of overlapping predictions, we employ a sampling approach, whereby we begin with the full set of overlapping and non-overlapping predictions of each participant and, by selective sampling, create sets consisting only of non-overlapping predictions (see Fig. 3). The IR metric and the associated α values are calculated for each of these sampled sets, and the resulting averages are assigned as the participant's skill and significance. The selective sampling of each participant's predictions is performed in the following steps:

1. Considering all closed predictions in a given round, we calculate the distance between the prediction centers for all predictions that overlap in the time domain.
2. If the distance between the centers of two predictions that overlap in time is less than the sum of their radii, then these predictions are labeled as "overlapping". Predictions that do not overlap with any other prediction are labeled as "non-overlapping".
3. After all predictions are labeled, the overlapping predictions are put in the "candidate" set. We begin by randomly selecting one of these candidates and remove all the predictions that overlap it (both in space and time).
4. We put the selected prediction in the "selection" set and repeat the procedure by randomly selecting one of the predictions in the candidate set. We repeat this until the candidate set is exhausted.
5. We then add all the non-overlapping predictions to the selection set. This set constitutes a sample set of independent prediction that we then use to calculate the IR score and α values as described above. We calculate an average value for both metrics by repeating this sampling procedure several times.

Based on the significance threshold ($\alpha \leq 0.05$) combined with the IR metric, we categorize the participants into the following skill classes: (A) significant participants with IR ≥ 2 and at least 5 independent predictions; (B) significant participants with IR ≥ 1.33 and at least 5 independent predictions; (C) participants with IR> 1 but who fail to satisfy either the significance, prediction number or IR criteria to become an A or B; (D) all participants with IR< 1. It can be argued that requesting a minimum number of predictions may affect the participants' behavior; some might start placing predictions that they would not have placed just to reach the limit. We

concede that the contest regulations will affect participant behavior in one way or another and deem such effects admissible as long as they do not hinder the goals of the competition. Participants who achieve skill classes of A or B are rewarded additionally to the reward distributed proportionally to the rX score. By distributing rewards according to two different but complementary performance metrics, we hope to make exploiting a single metric less enticing and to incentivize demonstrating actual skill. The rX score is relatively easier to calculate since the score of each new prediction is simply added to the current balance. However, the skill classes based on the IR score are more difficult to calculate because each new prediction affects the average prediction probability and estimating significance requires numerical simulation. We acknowledge that such statistical concepts can be intimidating for the general public and hinder participation. Therefore, we have implemented a recommendation algorithm that uses the currently closed predictions of each participant to suggest an additional number of true predictions with a probability sufficient to achieve skill classes B or A. The flowchart of the recommendation algorithm is given in Figure 4. In essence, the algorithm estimates what is the minimum number and highest reference model probability of additional true predictions that would satisfy both the significance and the IR criteria. If the participant has achieved skill class B, the algorithm would recommend predictions for achieving skill class A, while for classes C and D the recommendation would aim at B. In principle, similar recommendations can be calculated not necessarily for the minimum but for any number of predictions; the minimum probabilities would increase as the number of predictions increases. Figure 5 shows the outputs of the recommendation system based on the closed predictions of two different participants.

6 Synthetic consistency tests

We proposed the two score metrics introduced in the previous section to assess the predictive skills of individual participants as well as probabilistic forecasting models that can be sampled with deterministic predictions through an application programming interface. Fairness in reward distribution and reputation based contest is an essential factor that motivates participants. Moreover, from a scientific point of view, it is crucial to establish that the proposed metrics are powerful enough to discriminate between good and bad models such that research can be focused in more promising directions.

To test the consistency of the proposed metrics, we conduct a simplified synthetic ranking test. The test consists of three main components: (1) the ground truth model that generates the events; (2) several competing models that issue predictions trying to predict the generated events; (3) a reference model that is used as the basis of prediction probabilities entering in the rX and IR metrics. The synthetic prediction contest is carried out by all of the competing models issuing N_p predictions based on their expectations and the reference model probability. The outcome of the submitted predictions is dictated by the ground truth model. The scores are then calculated using the outcomes and the reference model probabilities assigned to the predictions submitted by the candidate models. The synthetic test is carried out in these steps:

1. The ground truth model is defined as a 1D probability vector with N_p elements $T = U(0.01, 0.99)$
2. Outcomes, occurrence, or no-occurrence, are generated by sampling each of the individual probabilities in the T vector to create an outcome vector O as per equation (9)
3. A set of m progressively worse candidate models C_i is created by perturbing the ground truth model by adding uniform random noise with increasing amplitude.

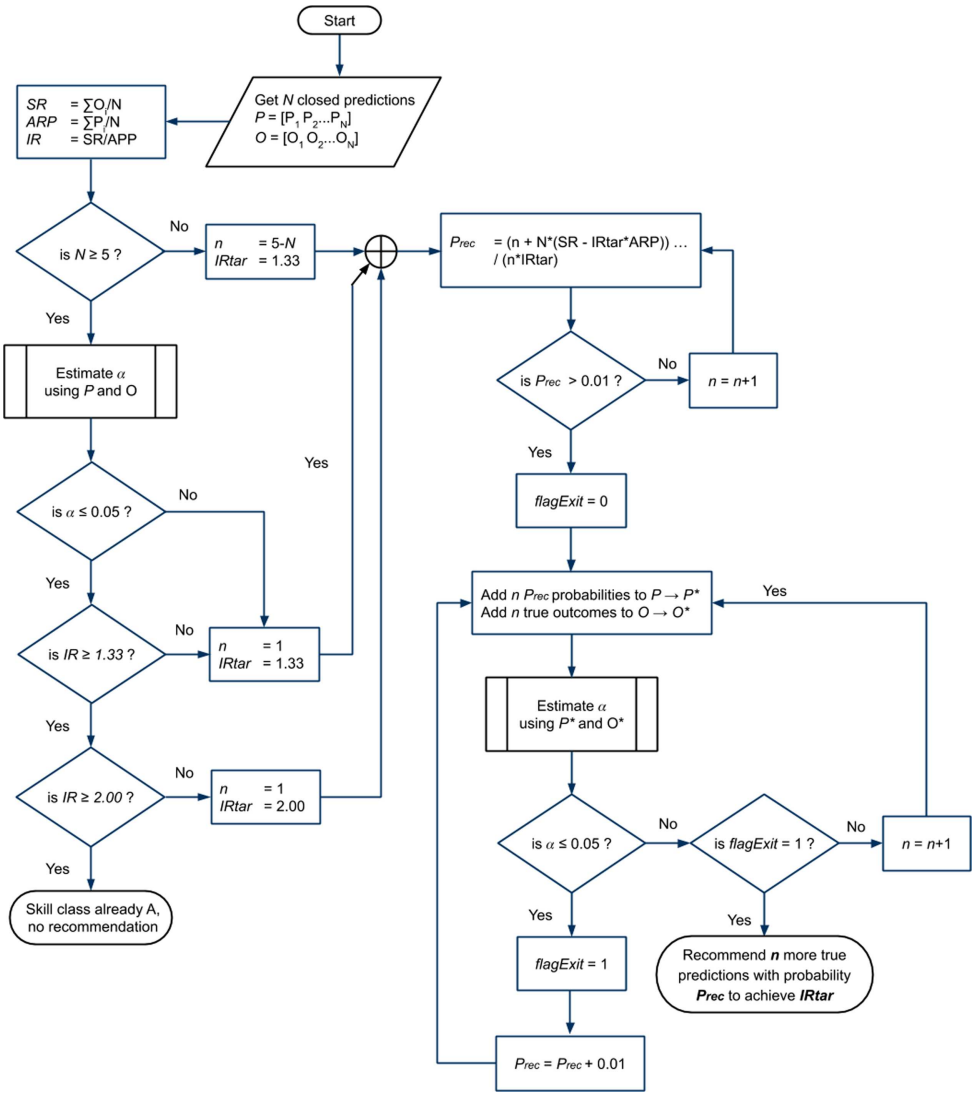

Fig. 4. Flow chart of the recommendation algorithm estimating the probability and number of true predictions needed to achieve a higher skill class. *SR*: success rate, *ARP*: average RichterX probability, *IRtar*: target information ratio.

The perturbed probabilities are capped to remain within the [0.01, 0.99] interval

$$x_i \in U(0,1)$$
$$C_i = \max\left(\min\left(T + \frac{i(x-0.5)}{m}, 0.99\right), 0.01\right) \tag{11}$$

4. For each of the N_p predictions, a candidate model indexed j decides to issue a *to-occur* or *not-to-occur* prediction by choosing the prediction type with the maximal expected return.

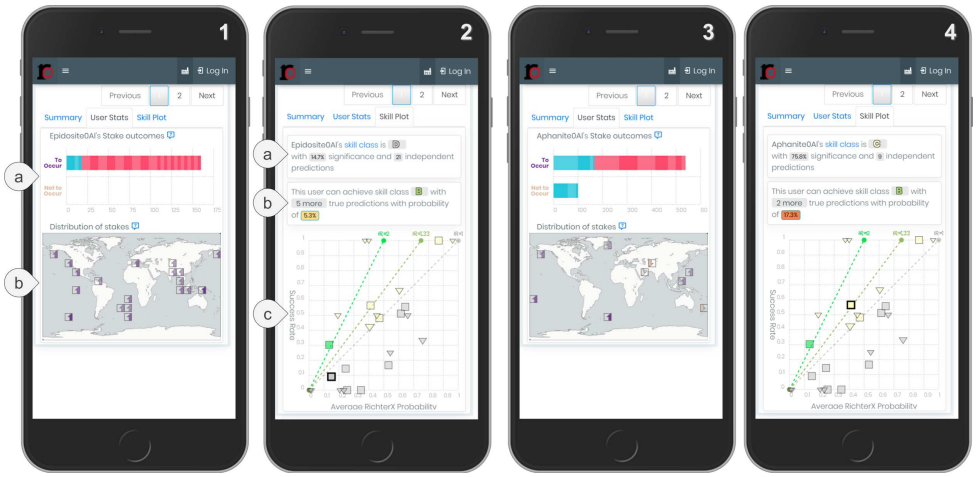

Fig. 5. (1) Summary of prediction outcomes for a participant: (a) Two bar charts indicating outcome of *to-occur* and *not-to-occur* predictions as false (red) or true (blue), the length of each bar scales with the prediction stake; (b) Map showing the location of the *to-occur* and *not-to-occur* predictions as purple up or beige down arrows. (2) Skill assessment plot for a participant: (a) Current skill class, from A to D, significance value $(1-\alpha)$ and number of independent predictions; (b) Recommendation containing the number of true predictions with a given probability needed to achieve a higher skill class; (c) Success rate vs average RichterX probability. Participants with less than 5 independent predictions are shown as triangles, others as squares. Colors indicate the skill class; A bright green, B green, C yellow, D gray. The selected participant is indicated by a symbol with thicker edges. (3) Same as (1) but for a different participant. (4) Same as (2) but for a different participant.

$$
\begin{aligned}
E[\mathrm{Occ}] &= C_i(j)\left(\frac{1-R(j)}{R(j)}\right) - (1-C_i(j)) \\
E[\mathrm{Noc}] &= (1-C_i(j))\left(\frac{R(j)}{1-R(j)}\right) - C_i(j)
\end{aligned}
\tag{12}
$$

where $E[\mathrm{Occ}]$ and $E[\mathrm{Noc}]$ denote a candidate model's expected return for a *to-occur* and *not-to-occur* prediction respectively.

5. All the issued predictions are assigned as true or false according to the outcome vector O, and each candidate model receives rX and IR scores.

We expect the consistency to improve with an increasing number of predictions N_p. Thus, we conducted the synthetic test for $N_p = [100, 1000, 5000]$. Figure 6 shows the results of the synthetic tests for the case when the reference model is chosen as the median ($C_i = 250$) and when the reference model is chosen as the worse ($C_i = 500$), respectively. We can see that, as the number of predictions increases, the fluctuation in both score metrics decreases, highlighting a linear relationship with the true rank. The skill of the reference model relative to the candidate models also plays an important role in interpreting the consistency results.

Since we created the candidate models by adding an increasing amount of noise to the ground truth model we also know the true ranking. We can study the scoring consistency by comparing the ranking obtained by each metric to the true ranking via the Kendall's rank correlation coefficient τ [1] in the following equation, where p_c and p_d are concordant (i.e having the same relative order) and discordant pairs, and n is the number of elements being ranked:

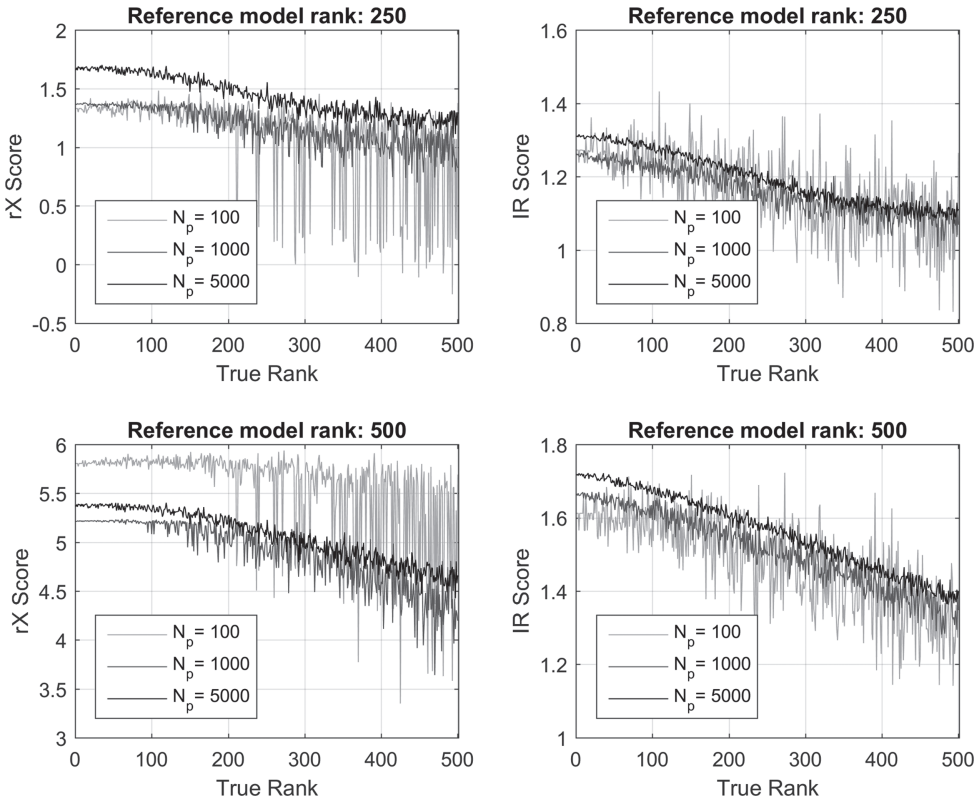

Fig. 6. The rX (left column) and IR (right column) scores for the 500 competing models resulting from N_p independent predictions. Increasing N_p values are shown in darker shades. Top row plots the results when the reference model is chosen as rank 250, i.e the average model. Bottom row corresponds to the reference chosen as rank 500, i.e. the worst model.

$$\tau = \frac{2\,(p_c - p_d)}{n(n-1)} \tag{13}$$

The coefficient τ is equal to 1 when the two rankings are exactly the same; -1 when they are the reverse of each other and values close to zero when they are unrelated. Figure 7 plots the τ coefficients as a function of increasing number of predictions, for different reference model choices. We observe that the IR score is more powerful in retrieving the original ranking, resulting in consistently higher τ values, both at small and large number of predictions and regardless of the choice of the reference model

Another important observation is that, when the reference model is chosen as the best model (i.e very close to the generating one), the ranking becomes inconsistent. Remember from equation (5), that, if the reference model is the generating model, we expect the rX score for any set of predictions to have an expected value of zero. Similarly, for the IR metric, the expected value would be 1 (see Eq. (9)). Figure 8 confirms this by plotting the individual model scores when the reference model is chosen as the best. For any number of predictions, the score values fluctuate around 0 and 1 for rX and IR respectively, explaining the near zero rank correlation values observed in Figure 7.

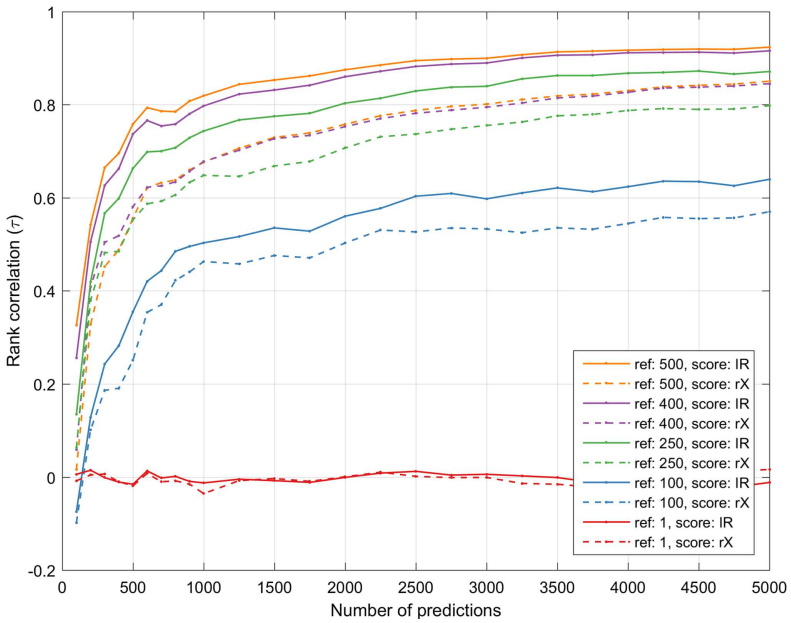

Fig. 7. Rank correlation between true rank and inferred ranked of 500 models as a function of increasing number of predictions. Results based on IR and rX score are shown as solid and dashed lines respectively. Different color are used to show the results when the true rank of the reference model is chosen as 1(best), 100, 250, 400 or 500 (worst).

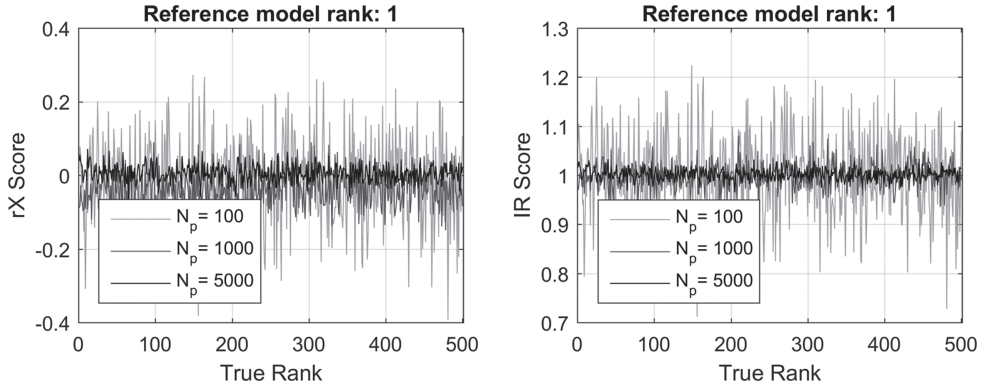

Fig. 8. The same as Figure 6 when the reference model is chosen as rank 1, i.e the best model.

It is important to note that the complete inability to distinguish between worse models will occur only when the reference model is very close to the truth. In the context of a prediction contest, this would mean that we have reached our final goal and that further research cannot provide added value. In reality, we know that our current models have a lot of room for improvement. We have demonstrated that, when this is the case, the proposed metrics are able to rank both models better and worse than the reference (see Figs. 6 and 7). Nevertheless, if there are concerns regarding the ranking of models that are much worse than the RichterX reference model, these can be easily addressed by invoking a very weak reference model, such as smoothing past seismicity assuming Poissonian rates.

7 Conclusion

The RichterX platform aims to rekindle the earthquake prediction effort by organizing a prediction contest invoking large scale participation both from the public and from different fields of academia. To facilitate this contest, we have implemented a real-time global earthquake forecasting model that can estimate the short term earthquake occurrence probabilities anywhere in the world. On one hand, this platform makes the contest highly accessible, allowing anyone with just a mobile phone to submit a prediction. On the other hand, it allows the public to query earthquake occurrence probabilities in real-time for any specific region, which becomes vital information, especially after the large mainshocks. In this way, with a single platform we hope to achieve three main goals: (1) Inform the public about short-term earthquake probabilities anywhere on the globe in real time; (2) Serve as a public record empowering the media and public officials to counter claims of earthquake prediction after the fact; (3) Allow researchers from various fields to easily participate in an earthquake prediction contest and challenge state-of-the-art global statistical seismology models.

Wide scale participation has the potential to bring forward and allow for the testing of various data sources that may or may not have precursory information. Current earthquake forecasting contests, which rely on systematic reporting of earthquake rates for large regions in predefined space-time resolutions, are not suitable for the testing of intermittent observations, such as earthquake lights, groundwater chemistry, electromagnetic and thermal anomalies, and so on. The RichterX platform can easily accommodate alarm based predictions based on such data sources. In addition, through synthetic ranking tests, we have shown that the proposed performance metrics can distinguish between probabilistic models that are better or worse than the reference model, and retrieve the true performance ranking.

References

1. H. Abdi, in *Encyclopedia of Measurement and Statistics* (Sage, Thousand Oaks, CA, 2007), p. 508–510
2. C.R. Allen, Bull. Seismol. Soc. Am. **66**, 2069 (1976)
3. P. Bird, D.D. Jackson, Y.Y. Kagan, C. Kreemer, R.S. Stein, Bull. Seismol. Soc. Am. **105**, 2538 (2015)
4. F. Cappa, J. Laut, M. Porfiri, L. Giustiniano, Comput. Human Behav. **89**, 246 (2018)
5. A. Chaia, A. Dalal, T. Goland, M.J. Gonzalez, J. Morduch, R. Schiff, *Half the world is unbanked: financial access initiative framing note* (Financial Access Initiative, New York, 2009)
6. R. Console, M. Murru, F. Catalli, G. Falcone, Seismol. Res. Lett. **78**, 49 (2007)
7. G. Coppi, L. Fast, *Blockchain and distributed ledger technologies in the humanitarian sector* (Hpg commissioned report, London, 2019), http://hdl.handle.net/10419/193658
8. M.A. Edwards, S. Roy, *Academic research in the 21st Century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition* (2017), https://www.liebertpub.com/doi/abs/10.1089/ees.2016.0223
9. Erdstösse im Wallis, *Zahlreiche Erdstösse schrecken Menschen im Wallis auf* (2019), https://www.tagesanzeiger.ch/panorama/vermischtes/naechtliches-erdbeben-erschuettert-das-wallis/story/13668757
10. A. Extance, Nature **526**, 21 (2015)
11. C. Fabian, Innov. Technol. Governance Globalization **12**, 30 (2018)
12. D. Fiorillo, Ann. Public Cooperative Econ. **82**, 139 (2011)

13. D. Fletcher, *Model averaging* (Springer, 2019)
14. GEOFON, Deutsches GeoForschungszZntrum GFZ (1993)
15. D. Harte, D. Vere-Jones, Pure Appl. Geophys. **162**, 1229 (2005)
16. D. Harte, D.F. Lp, M. Wreede, D. Vere-Jones, Q. Wang, New Zealand J. Geol. Geophys. **50**, 117 (2007)
17. S. Hiemer, Y. Kamer, Seismol. Res. Lett. **87**, 327 (2016)
18. D.D. Jackson, Proc. Nat. Acad. Sci. USA **93**, 3772 (1996)
19. I.T. Jolliffe, Meteorol. Appl. **15**, 25 (2008)
20. T.H. Jordan, Seismol. Res. Lett. **77**, 3 (2006)
21. Y.Y. Kagan, *Worldwide Earthquake Forecasts* (2017)
22. Y.Y. Kagan, D.D. Jackson, Geophys. J. Int. **143**, 438 (2000)
23. Y.Y. Kagan, D.D. Jackson, R.J. Geller, Seismol. Res. Lett. **83**, 951 (2012)
24. Y.Kamer, S. Hiemer, J. Geophys. Res. Solid Earth **120**, 5191 (2015)
25. V.I. Keilis-Borok, V.G. Kossobokov, Phys. Earth Planet Inter. **61**, 73 (1990)
26. Y.M. Kow, First Monday **22** (2017)
27. Y.-T.T. Lee, D.L. Turcotte, J.R. Holliday, M.K. Sachs, J.B. Rundle, C.-C.C. Chen, K.F. Tiampo, Proc. Nat. Acad. Sci. USA **108**, 16533 (2011)
28. M.R. Lepper, D. Greene, *The hidden costs of reward: New perspectives on the psychology of human motivation* (Lawrence Erlbaum, Oxford, England, 1978)
29. A. Lomax, A. Michelini, Pure Appl. Geophy. **170**, 1385 (2013)
30. G. Molchan, L. Romashkova, [arXiv: 1005.3175](2010)
31. G.M. Molchan, Phys. Earth Planet. Inter. **61**, 84 (1990)
32. G.M. Molchan, Tectonophysics **193**, 267 (1991)
33. S. Nandan, G. Ouillon, S. Wiemer, D. Sornette, J. Geophys. Res. Solid Earth **122**, 5118 (2017)
34. S. Nandan, G. Ouillon, D. Sornette, S. Wiemer, Seismol. Res. Lett. **90**, 1650 (2019)
35. S. Nandan, G. Ouillon, D. Sornette, S. Wiemer, J. Geophys. Res. Solid Earth **124**, 8404 (2019)
36. S. Nandan, Y. Kamer, G. Ouillon, S. Hiemer, D. Sornette, Eur. Phys. J. Special Topics **230**, 425 (2021)
37. C.G. Northcutt, A.D. Ho, I.L. Chuang, Comput. Edu. **100**, 71 (2016)
38. Y. Ogata, J. Am. Stat. Assoc. **83**, 9 (1988)
39. M. Pagani, J. Garcia, D. Monelli, G. Weatherill, A. Smolka, Ann. Geophys. **58** (2015), https://www.annalsofgeophysics.eu/index.php/annals/article/view/6677
40. W. Savran, P. Maechling, M. Werner, D. Schorlemmer, D. Rhoades, W. Marzocchi, J. Yu, T. Jordan, *The Collaboratory for the Study of Earthquake Predictability Version 2 (CSEP2): Testing Forecasts that Generate Synthetic Earthquake Catalogs* (EGUGA, 2019), p. 12445
41. D. Schorlemmer, J.D. Zechar, M.J. Werner, E.H. Field, D.D. Jackson, T.H. Jordan, Pure Appl. Geophys. **167**, 859 (2010)
42. D. Schorlemmer, M.J. Werner, W. Marzocchi, T.H. Jordan, Y. Ogata, D.D. Jackson, S. Mak, D.A. Rhoades, M.C. Gerstenberger, N. Hirata, M. Liukis, P.J. Maechling, A. Strader, M. Taroni, S. Wiemer, J.D. Zechar, J. Zhuang, Seismol. Res. Lett. **89**, 1305 (2018)
43. A. Sol, H. Turan, Sci. Eng. Ethics **10**, 655 (2004)
44. K. Starbird, L. Palen, Working & sustaining the virtual disaster desk, in *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, New York, USA, 2013* (ACM Press, New York, USA, 2013)
45. U.S. Geological Survey Earthquake Hazards Program, *Advanced National Seismic System (ANSS) comprehensive catalog of earthquake events and products* (2017)
46. D.L. Wells, K.J. Coppersmith, Bull. Seismol. Soc. Am. **84**, 974 (1994)
47. J. Whitehill, *Climbing the kaggle leaderboard by exploiting the log-loss oracle*, Technical report (2018)
48. J. Woessner, S. Hainzl, W. Marzocchi, M.J. Werner, A.M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M.C. Gerstenberger, S. Wiemer, J. Geophys. Res. **116**, 1 (2011)
49. H.O. Wood, B. Gutenberg, *Earthquake Prediction* (1935)