Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment

submitted to *Geophys. J. Int.*

Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment

Francesco Serafini[1⋆], Mark Naylor [1], Finn Lindgren[2], Maximilian J. Werner [3], Ian Main[1]

[1] *School of Geosciences, University of Edinburgh,* [2] *School of Mathematics, University of Edinburgh,*

[3] *School of Earth Sciences, University of Bristol*

SUMMARY

Operational earthquake forecasting for risk management and communication during seismic sequences depends on our ability to select an optimal forecasting model. To do this, we need to compare the performance of competing models in prospective experiments, and to rank their performance according to the outcome using a fair, reproducible, and reliable method, usually in a low-probability environment. The Collaboratory for the Study of Earthquake Predictability (CSEP) conducts prospective earthquake forecasting experiments around the globe. In this framework, it is crucial that the metrics employed to rank the competing forecasts are 'proper', meaning that, on average, they prefer the data generating model. We prove that the Parimutuel Gambling score, proposed, and in some cases applied, as a metric for comparing probabilistic seismicity forecasts, is in general improper. In the special case where it is proper, we show it can still be used improperly. We demonstrate the conclusions both analytically and graphically providing a set of simulation based techniques that can be used to assess if a score is proper or not. They only require a data generating model and, at least two forecasts to be compared. We compare the Parimutuel Gambling score's performance with two commonly-used proper scores (the Brier and logarithmic scores) using confidence intervals to account for the uncertainty around the observed score difference. We suggest that using confidence intervals

enables a rigorous approach to distinguish between the predictive skills of candidate forecasts, in addition to their rankings. Our analysis shows that the Parimutuel Gambling score is biased, and the direction of the bias depends on the forecasts taking part in the experiment. Our findings suggest the Parimutuel Gambling score should not be used to distinguishing between multiple competing forecasts, and for care to be taken in the case where only two are being compared.

Key words: Probabilistic forecasting, Statistical methods, Statistical seismology, Earthquake interaction, forecasting, and prediction

1 INTRODUCTION

Probabilistic earthquake forecasts are used to estimate the spatial and/or temporal evolution of seismicity and have potential utility during earthquake sequences, including those following notable earthquakes. For example, they have been applied to forecast (pseudoprospectively) the seismicity that followed the Darfield earthquake and in turn led to the 2011 Christchurch earthquake (Rhoades et al , 2016), and to monitor induced seismicity at Groningen (Bourne et al , 2018). In Italy, earthquake probabilistic forecasts and ground-motion hazard forecasts are produced on a regular basis by the Instituto Nazionale di Geofisica e Vulcanologia (INGV) to inform the Italian government on the risk associated with natural hazard (Marzocchi et al , 2014). INGV is working to use probabilistic forecasts as a basis for modelling important quantities for operational loss forecasting such as the number of evacuated residents, the number of damaged infrastructure, the number of fatalities (Iervolino et al , 2015). A wider uptake requires further demonstrations of the operational utility of the forecasts, and in presence of multiple alternative models, a fair and rigorous method to express a preference for a specific approach is needed. The Collaboratory for the Study of Earthquake Predictability (CSEP, see Jordan 2006; Zechar et al 2010b; Schorlemmer et al 2018) is a global community initiative that seeks to make earthquake research more rigorous and open-science. This is done by comparing forecasts against future data in competition with those from other models through prospective testing in pre-defined testing regions. In this paper, we focus on comparing different forecasts that can be made from such competing models in the light of observed data.

In statistics, a common approach to compare probabilistic forecasts is the use of scoring rules (Gneiting and Raftery , 2007). Scoring rules have been widely applied in many fields of science to

* Corresponding author's email address : francesco.serafini@ed.ac.uk

measure the quality of a forecasting model and to rank competing models based on their consistency with the observed data and the degree of uncertainty around the forecast itself. Much of the underlying methodology and concepts (such as what it means to be a ”good” forecast) have been developed for weather forecasts (Murphy , 1993; Jolliffe and Stephenson , 2003). A positively oriented scoring rule, to be effective, has to be *proper*, which simply means that the highest score is achieved, on average, by the forecasting model ”closer” to the distribution that has generated the observations. Various meaning of ”closer” can be used depending on the context and the use that will be made of the forecasting model under evaluation, thus, a variety of proper scoring rules exists. Proper scoring rules are mathematically appealing for a range of different tasks: they can be used as utility function tailored to the problem at hand, they can be used as loss functions in parametric estimation problems and they can be used to rank competing models based on different aspects of the phenomenon under analysis (Rosen , 1996; Hyvärinen and Dayan , 2005; Hernández-Orallo et al , 2012).

CSEP aims to compare the predictive performance of diverse earthquake forecasts in a rigorous and reproducible fashion. The forecasts themselves are generated by underlying physical, stochastic or hybrid models using a variety of input data such as past seismicity, deformation rates, fault maps, etc (Field et al , 2014; Steacy et al , 2014; Bayliss et al , 2020). The two most widely used types are alarm-based forecasts and probabilistic forecasts. The first class of forecasts is usually expressed as a binary statement (”alarm” or ”not alarm”) based on the value of a precursory alarm function. In contrast, probabilistic forecasts, as intended in past CSEP experiments (Schorlemmer and Gerstenberger , 2007a), provide a distribution for the number of earthquakes. They can be expressed as grid-based forecasts (providing the expected number of events in each space-time-magnitude bin) or as catalogue-based (providing a number of simulated catalogues, Savran et al 2020). The forecasts are variously compared using a suite of community-endorsed tests. Depending upon the forecasts at hand, three common challenges are the need for a reference model, how to handle bins (or regions) for which the forecaster didn’t provide a forecast and, the need to specify a likelihood. The latter has been partially solved by the possibility of considering a pseudo-likelihood (Savran et al , 2020).

Molchan diagrams (Zechar and Jordan , 2008) and the area-skill score (Zechar and Jordan , 2010) do not need a likelihood and can be used to compare both alarm-based and probabilistic forecasts together. However, they need a reference model for assessing the significance of the results. This can be problematic because specifying a credible reference model is a difficult task (Stark , 1997; Luen et al , 2008; Marzocchi and Zechar , 2011). Likelihood-based tests (Schorlemmer et al , 2007; Zechar et al , 2010a; Rhoades et al , 2011; Schneider et al , 2014) allow for pairwise comparison without the need of a reference model, but can only be applied to probabilistic forecasts. Further, methods for grid-based forecasts rely on the Poisson assumption, which has been observed to be not realistic

(Werner and Sornette , 2008). Moreover, pairwise comparison may lead to paradoxical results like model A is preferred to model B which is preferred to model C which is preferred to model A (Zechar et al , 2013). Bayesian methods have been proposed (Marzocchi et al , 2012) but they also rely on the Poisson assumption. Catalogue-based forecasts, can be evaluated using a pseudo-likelihood approach (Savran et al , 2020) which does not rely on the Poisson assumption and enable information gains and likelihood ratios to be used. However, the latter are unbounded and sensitive to low-probability events, meaning that they can be unduly influenced by a few observations (Holliday et al , 2005; Zechar and Zhuang , 2014). Lastly, in past experiments such as the Regional Earthquake Likelihood Models (RELM) (Field , 2007), forecasters did not provide a forecast for all bins, some of them were left as missing value; for the methods outlined above, making a comparison is complex, given that considering only the overlapping space-time-magnitude volume may be too restrictive and introduce unfairness in the evaluation.

Zhuang (2010) and Zechar and Zhuang (2014) tried to overcome the difficulties outlined above by introducing the parimutuel gambling score, which provides a framework to evaluate different types of forecasts, with no need to explicitly specify a reference model or a likelihood, and with the ability to handle missing values in an intuitive way. This approach is based on the idea that alarm-based forecasts could be imagined as gamblers engaged in a game called the seismic roulette, where Nature controls the wheel (Main , 1997; Kossobokov , 2004; Kossobokov , 2006). In this framework, the forecasters are the gamblers, a forecast consists of a collection of probabilities for observable events (bets) like *observing at least one earthquake in a specified space-time-magnitude bin*. Each bin represents a bet and the probability assigned by the forecaster represents the amount of money wagered. The observations consist of binary variables taking value 1 if the event occurs and zero otherwise. The forecaster gets a reward depending on the forecasted probability and the actual observation of an event or not. The forecasts are ranked based on their rewards. In this sense, the parimutuel gambling score is a positively oriented score (the higher, the better) for binary probabilistic forecasts. In this paper, we prove analytically and graphically that the parimutuel gambling score is not proper in general but only in a specific situation and we compare its performance with two proper alternatives: the Brier (Brier , 1950) and the logarithmic (Good , 1952) score.

The parimutuel gambling score has not been used systematically in CSEP, but it has been used to evaluate global forecasts (Taroni et al , 2016) and forecasts for Italy (Taroni et al , 2014, 2018) in situations where the score is improper. In the context of Italy, it has also been used in combination with two other scores to weight different source and ground motion models in the new Italian seismic hazard ensemble model (MPS19, Meletti et al 2021). Furthermore, the parimutuel gambling score was mentioned by Schorlemmer et al (2018) as a new method for evaluating earthquake forecasts

without any warning about possible biases. We use the parimutuel gambling score to illustrate different techniques to assess if a score is proper, both analytically and graphically. We find that the parimutuel gambling score is proper only in a specific situation, and event then, it can be used improperly. This finding is emblematic of how much care should be taken in checking if (and when) a metric is proper.

To fairly compare the performance of the scores in a realistic framework, we use simulated data from a known model and we compare it with alternative models. In doing that, it is crucial to account for the uncertainty in the observed score difference. In fact, properness ensures that, at least on average, the scoring rule provides the correct ranking. However, the score calculated from any finite set of observations could be far from its average and, therefore, we need to account for uncertainty. In this paper, we show how to express a preference towards a model using confidence intervals for the expected score difference. This method introduces the possibility of not expressing a preference. Considering this outcome is potentially useful because it indicates that, for a scoring rule, the forecasts have similar performances, or the data are not enough to distinguish between models, or the bins' dimension is offset (too large or too small).

In summary, the main goal of this article is to present the notion of proper scoring rules for probabilistic forecasts of binary events: why is it crucial for a scoring rule to be proper? How can we verify if a score is proper or not? And, how do different scores penalize the same forecast differently? In Section 2 we define a proper and a strictly proper scoring rule, introduce the Brier and log scores as examples, and give a brief proof of their propriety. We also show how differently forecasts close to zero are penalized by the two scores. In Section 3, we introduce the parimutuel gambling score and analytically explore its improperness in the context of a forecast for a single bin. If a score is proper for single bins, then, the average score of different bins is also a proper score (Gneiting and Raftery , 2007). In Section 4, we generalise to the case where we have multiple bins but with the same probability, *Multiple Bins Single Probability*. This case is equivalent to considering the activity rate in each bin as independent and identically distributed. This is a significant assumption but allows us to calculate analytically the confidence intervals and the probability of expressing a preference for a given model. We generalise further to the case in which we have multiple bins but with a different probability for each bin, *Multiple Bins Multiple Probabilities*. In this case, we do not have analytical results and we are required to use approximate confidence intervals and simulations to calculate the probability of expressing a preference. These simulations are now close to a real forecast scenario. We illustrate this case using simulations from the time-independent 5 year adaptively-smoothed forecast for Italy (Werner et al , 2010). We choose this model because the adaptively-smoothed approach performed well across multiple metrics in the RELM experiment (Zechar et al , 2013) and, as a result,

was incorporated into the California seismic hazard map produced by the third Uniform California Earthquake Rupture Forecast model (UCERF3, Field et al 2014).

2 PROPER SCORES

Scoring rules quantify the quality of probabilistic forecasts, allowing them to be ranked. The quality depends on both the predictive distribution, produced by the model in true prospective mode, and on the subsequent observations. A scoring rule is a function of the forecast and the data measuring two factors: the consistency between predictions and observations and the sharpness of the prediction. Consistency assesses the calibration of the model, how well the forecast and the data agree, and is a joint property of the forecast and the data. Sharpness is a measure of the forecast uncertainty and is a property of the forecast only. Different scoring rules measure the consistency and the sharpness of a forecast differently. As in (Gneiting and Raftery , 2007), we call $S(P|x)$ the score for forecast P given the observation x. In general, we use capital letters for random variables, lowercase letters for scalar quantities such as realizations of a random variable (everything that is not random) and bold letters represents vectors. The only exception is N which represents the number of bins.

Thus, a scoring rule, given a forecast P, is a function of the observation only $S(P|\cdot) : \mathcal{X} \rightarrow [-\infty, \infty]$ where $\mathcal{X}$ is the set of all possible values of x. For consistency, we will use a *positively orientated* convention, where a larger score indicates a better forecast. Assuming that the observations are samples from a random variable $X \in \mathcal{X}$ with true distribution Q, the score $S(P|X)$ is a random variable itself, since it is a function of the random variable X. We define $S^E(P|Q)$ as the expected value of the scoring rule under the true distribution Q:

$$S^E(P|Q) = \mathbb{E}_Q[S(P|X)]. \tag{1}$$

A positively oriented scoring rule S is said to be *proper* if, for any forecast P and any true distribution Q, $S^E(Q|Q) \geq S^E(P|Q)$ holds. It is said to be *strictly proper* if $S^E(Q|Q) = S^E(P|Q)$ if and only if $P = Q$. Propriety is essential, as it incentivises the assessor to be objective and to use the forecast P "closer" to the true distribution Q. Different scoring rules rely on different meanings of closer. Also, proper scores can be used as loss functions in parameter estimation; in fact, since the likelihood assigned by a model to the observations can be seen as a proper scoring rule, the maximum likelihood estimator can be viewed as optimizing a score function (Huber , 1992). Investigating the ability of a score of distinguishing between different instances of the same model (with different parameters values) may bring insight regarding parameters identifiability.

Here, we are interested in scoring rules for binary variables, in which the variable X can be

only 0 or 1, namely $X \in \{0, 1\}$. Grid-based earthquake forecasts divide the region of interest into regular space-time-magnitude bins (e.g. the spatial region is divided in bins of 0.1×0.1 degrees, the magnitude by 0.1 magnitude units, and the time is one 5-year bin), and the forecasters estimate the expected number of earthquakes per bin. In this case, for example, the binary variable might be 0 for empty bins and 1 if at least one event occurs. The forecasts may be ranked based on the average score across different bins (Zechar et al , 2013).

Considering a single bin, for grid-based binary forecasts, where both the forecast P and the true distribution Q are specified by just one number: the probability of X being 1. We call p the probability assigned to the event $X = 1$ by the forecaster, and p^* denotes the true probability. Thus, the expectation is given by

$$S^E(P|Q) = S^E(p|p^*) = p^* S(p|1) + (1 - p^*)S(p|0). \tag{2}$$

A scoring rule of this type is proper if, for any $p \in [0, 1]$ and any $p \in [0, 1]$, we have

$$S^E(Q|Q) \geq S^E(P|Q).$$

The properness of a score ensures that given two models p_1, p_2, the model with the greatest expected score $S^E(p_i|Q)$ is the closest to the true p^*. This notion can be generalized to rank a set of k forecasts $p_1, ..., p_k$ according to their expected scores.

Two of the most widely used strictly proper scoring rules, for binary data, are the Brier (or quadratic) score (Brier , 1950) and the logarithmic score (Good , 1952). These are good candidates for evaluating this class of earthquake forecasts. Here we give the definitions of these two scores, including brief proofs of their propriety.

2.1 Brier Score

The positively oriented Brier score (Brier , 1950) for a categorical variable X (the binary case is obtained considering only two possible outcomes) can be defined by:

$$S_B(P|x) = -\sum_{z \in \mathcal{X}} [p(z) - \mathbb{I}(z = x)]^2, \tag{3}$$

where $\mathcal{X}$ is the set of possible outcomes, $p(z)$ is the forecasted probability of the event $X = z$, and $\mathbb{I}(z = x)$ is an indicator function assuming value 1 if $z = x$ and 0 otherwise. This definition differs from the original only in the sign, since the original Brier score is negatively oriented.

The ordinary Brier score for binary events is the special case $\mathcal{X} = \{0, 1\}$, with $p = p(1)$ and

$1 - p = p(0)$:

$$S_B(p|x) = -[(1-p)-(1-x)]^2 - (p-x)^2 = -2(p-x)^2 \quad = \begin{cases} -2(p-1)^2, & x = 1, \\ -2p^2, & x = 0, \end{cases}$$

which has expectation

$$S_B^E(p|p^*) = -2p^*(p-1)^2 - 2(1-p^*)p^2 \tag{4}$$

under the true event probability p^*. Taking the derivative with respect to p and imposing it equal zero, we find that the value $p = p^*$ uniquely maximizes the function $S_B^E(p|p^*)$ which proves that the Brier score is strictly proper.

2.2 Logarithmic Score

The logarithmic (log) score for binary event forecasts is defined as

$$S_L(P|x) = \ln p_P(x). \tag{5}$$

For $\mathcal{X} = \{0, 1\}$, the expectation is

$$S_L^E(p|p^*) = p^* \ln(p) + (1-p^*)\ln(1-p), \tag{6}$$

which, once differentiated with respect to p and set equal zero to identify the maximum, proves that also the log score is strictly proper.

2.3 Score Comparison

Given an observation x, to express a preference between two forecasts p_1 and p_2, an important quantity is the score difference Δ.

$$\Delta(p_1, p_2, x) = S(p_1|x) - S(p_2|x) = \begin{cases} S(p_1|0) - S(p_2|0) & \text{with prob} \quad 1-p^*, \\ S(p_1|1) - S(p_2|1) & \text{with prob} \quad p^*. \end{cases}$$

For example, in the case of the Brier score we have

$$\Delta_B(p_1, p_2, x) \begin{cases} -2(p_1^2 - p_2^2) & \text{when} \quad x = 0, \\ -2[(1-p_1)^2 - (1-p_2)^2] & \text{when} \quad x = 1, \end{cases} \tag{7}$$

while in the case of the log score

$$\Delta_L(p_1, p_2, x) \begin{cases} \log(\frac{1-p_1}{1-p_2}) & \text{when} \quad x = 0, \\ \log(\frac{p_1}{p_2}) & \text{when} \quad x = 1. \end{cases} \tag{8}$$

In principle, if the expected value of Δ is positive we tend to prefer the first forecast, vice versa if

it is negative. Considering the observation as a Bernoulli random variable $X \sim \text{Ber}(p^*)$, the difference $\Delta(p_1, p_2, X)$ is also a binary random variable, assuming the values $\Delta_0 = \Delta(p_1, p_2, 0)$, $\Delta_1 = \Delta(p_1, p_2, 1)$ with probabilities $1 - p^*$ and p^*. The distribution of $\Delta(p_1, p_2, X)$ is therefore completely determined by the distribution of X:

$$\Delta(p_1, p_2, X) = X\Delta_1 + (1 - X)\Delta_0 = \Delta_0 + X(\Delta_1 - \Delta_0). \tag{9}$$

It follows that the expected value and variance of $\Delta(p_1, p_2, X)$ are determined by the properties of X:

$$\mathbb{E}[\Delta(p_1, p_2, X)] = \Delta_0 + \mathbb{E}[X](\Delta_1 - \Delta_0) = \Delta_0 + p^*(\Delta_1 - \Delta_0) \tag{10}$$

$$\mathbb{V}[\Delta(p_1, p_2, X)] = \mathbb{V}[X](\Delta_1 - \Delta_0)^2 = p^*(1 - p^*)(\Delta_1 - \Delta_0)^2 \tag{11}$$

We can give an alternative definition of the properness based on the random variable $\Delta(p_1, p_2, X)$. In fact, a scoring rule S is said to be proper if $\mathbb{E}_X[\Delta(p, p^*, X)] \leq 0$ when $p \neq p^*$, no forecast have an expected score higher than the data generating model p^*. However, they can achieve the same score. S is strictly proper if $\mathbb{E}_X[\Delta] = 0$ if and only if $p = p^*$, the highest score, on average, is achieved only by the data generating model. The definition implies, also, that proper scoring rules are invariant under linear transformations, in the sense that, a linear transformation of a proper score yields another proper score and the operation does not change the ranking.

Figure 1 reports the expected score difference between a candidate forecast p and the true value $p^* = 0.001$ using the Brier and the log score. The value $p^* = 0.001$ was chosen to be comparable to the estimated probability of having an event with magnitude greater than 5.5 calculated the days before the L'Aquila earthquake in the neighbourhood of where it struck (Fig. 4 in Marzocchi and Lombardi 2009). To enable a visual comparison, the expected Brier score values have been normalized to match the curvature of the log score when $p = p^*$. This is done by multiplying the expected Brier score values by the ratio of the second derivatives of the two expected scores calculated at $p = p^*$. The proper score scale invariance ensures that the ranking obtained using the original and normalized version of the Brier score is unchanged.

Both expected score differences are uniquely maximized at $p = p^*$ which means that the forecast matching the true probability has the highest expected score. This is an easy way to assess if a scoring rule for binary outcomes is proper or not. Furthermore, Figure 1 offers an example of how different scores penalize differently the same forecasts. The log score is asymmetric and takes into account the relative differences between the forecasts (equation 8), and if $p^* \neq 0$ the expected score for $p = \{0, 1\}$ is $-\infty$. The log score is analogous to a likelihood score, and brings the same properties: a model which is correct in all the bins but one for which it provides zero probability will have the worst possible score. The Brier score, instead, considers the absolute difference between forecasts (equation

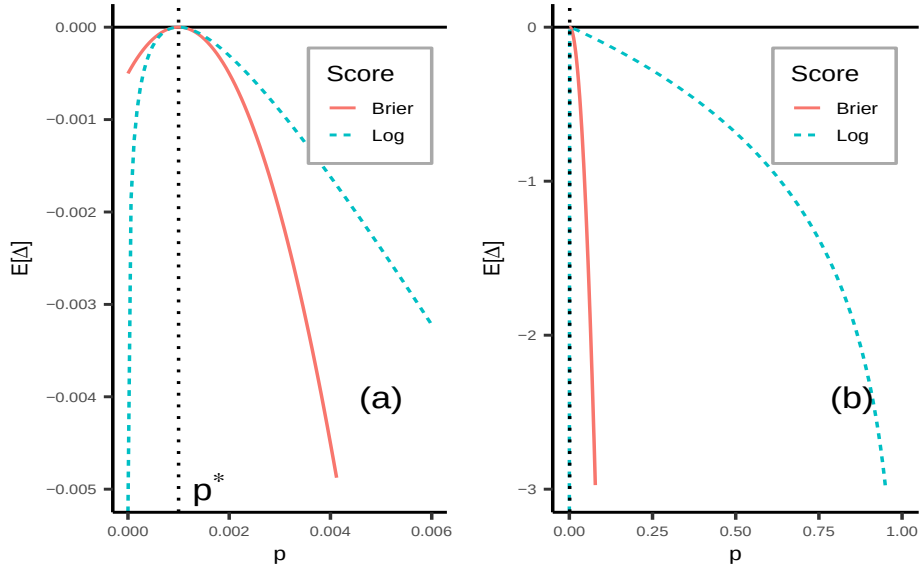

Figure 1. Differences in the scores expected value for a generic value of the forecast p and the optimal forecast $p = p^*$, namely $\mathbb{E}[\Delta] = S^E(p|p^*) - S^E(p^*|p^*)$, in the case $p^* = 0.001$. Panel (a) $p \in (0, 0.006]$, Panel (b) $p \in (0, 1)$. The expected Brier score have been normalised to match the curvature of the log score when $p = p^*$.

7) resulting in a symmetric distribution For example, using the Brier score, a forecast $p = 0$ will be preferred to any forecast in $(2p^*, 1)$, for any $p^* < 1/2$.

The choice of score, and consequently the style of penalty, should reflect the task at hand. Predicting $p = 0, 1$ means that we are absolutely certain about the outcome of X. If the forecasts under evaluation are planned to be used in an alarm based system, for which an alarm is broadcasted if the probability is above or below a certain threshold, being overconfident may put lives at risk and perhaps the log score would be the right choice in this situation. On the other hand, the Brier score may be suitable when such a strict penalty is not desirable (e.g., to calculate the weights of an ensemble model as done by Taroni et al 2018 and Meletti et al 2021). This example illustrates the flexibility of proper scores and how important it is to choose the right one depending on the purposes of the forecast under evaluation.

3 IMPROPER SCORES

Scores which are not proper are called improper. Being improper means that a model may exist with expected score greater than the data generating model. In the specific case of probabilistic forecasts for binary events, a score is improper if it is biased towards models which systematically under/overestimate the true probability p^*. In the context of earthquake forecasting experiments we do not know the true value of p^*. Therefore, it is crucial to use proper scoring rules for which we are sure

that, at least on average, they will prefer the closest model to the data generating one. Improper scoring rules do not have this property, which implies that the smallest or the largest (or any other) forecast, on average, could achieve the highest score. This is in clear contrast with the aim of any forecasting experiment. Below, we demonstrate that the parimutuel gambling score (Zhuang , 2010; Zechar and Zhuang , 2014) is an example of a scoring rule which is proper only in a specific situation and not in general.

3.1 Definition of the parimutuel gambling score

The parimutuel gambling score was designed to rank forecasting models for binary events and was applied to rank earthquake forecasting models in CSEP experiments (Taroni et al , 2018; Zechar and Zhuang , 2010). Initially, it was used to compare models against a reference model (Zhuang , 2010), which is improper. Later, it was generalized to compare models against each other simultaneously (Zechar and Zhuang , 2014), the case with only two players is the special case for which the score is proper, all the others are not. The score is based on a gambling scheme in which the forecasting models play the role of the gamblers and, for each observation, they obtain a reward proportional to the probability assigned by the gambler to the event occurring. In particular, it is a zero-sum game, in the sense that bids and rewards in each bin sum to zero, which makes the parimutuel gambling score relative to one forecast dependent on the other forecasts.

In contract to the Brier and log scores, it is not possible to define the parimutuel gambling score using the form $S(p|x)$ because it needs at least two forecasts to be evaluated and is a function of them all. Given a set of k forecasts $\mathbf{p} = (p_1, ..., p_k)$, we define $S_G(\mathbf{p}|x)$ as the vector such that the i-th component, $S_{G,i}(\mathbf{p}|x)$, is given by the parimutuel gambling score of the i-th forecast, given x has been observed. In the case of the Brier and log score the components of the vector $S(\mathbf{p}|x)$ are defined independently, in the case of the parimutuel gambling score they have to be defined jointly. Let $\bar{p}$ be the average probability involved in the gambling scheme, namely $\bar{p} = \sum_{i=1}^{k} p_i/k$. The parimutuel gambling score relative to the i-th forecast is defined as

$$S_{G,i}(\mathbf{p}|x) = \begin{cases} \frac{p_i}{\bar{p}} - 1, & x = 1, \\ \frac{1-p_i}{1-\bar{p}} - 1, & x = 0. \end{cases}$$

The above expression is a zero-sum game, meaning that $\sum_i S_{G,i}(\mathbf{p}|x) = 0$, therefore the rewards may be positive or negative. Each gambler obtains a positive reward if and only if they assign a greater probability to the observed event than the average gambler involved in the game. Vice versa, the reward is negative if the probability is smaller.

The expected value with respect the true probability p^* is given by

$$\begin{aligned} S^E_{G,i}(\mathbf{p}|p^*) &= p^* \left(\frac{p_i}{\bar{p}} - 1 \right) + (1 - p^*) \left(\frac{1 - p_i}{1 - \bar{p}} - 1 \right), \\ &= \frac{p^* p_i}{\bar{p}} + \frac{(1 - p^*)(1 - p_i)}{1 - \bar{p}} - 1, \\ &= \frac{(p_i - \bar{p})\,(p^* - \bar{p})}{\bar{p}(1 - \bar{p})}. \end{aligned} \tag{12}$$

Equation (12) is the same as equation (5) in (Zechar and Zhuang , 2014). The denominator involves all the probabilities in the game which demonstrates the interdependence with all other forecasts and complicates the study of the derivatives. However, it is still possible to prove that the gambling score is strictly proper when $k = 2$. In this case, $\mathbf{p} = (p_1, p_2)$, and

$$\begin{aligned} 4\bar{p}(1 - \bar{p}) S^E_{G,1}(\mathbf{p}|p^*) &= 4 \left(p_1 - \frac{p_1 + p_2}{2} \right) \left(p^* - \frac{p_1 + p_2}{2} \right), \\ &= (p_1 - p_2)\,(2p^* - p_1 - p_2)\,, \\ &= -\left[(p_1 - p^*) - (p_2 - p^*) \right] \left[(p_1 - p^*) + (p_2 - p^*) \right], \\ &= (p_2 - p^*)^2 - (p_1 - p^*)^2. \end{aligned}$$

The expected reward of the first modeler is non-negative when $|p_1 - p^*| \leq |p_2 - p^*|$, implying that p_1 is favoured over p_2 if it is closer to the true probability p^*. In fact, if $p_2 = p^*$ then, $S^E_{G,1}(\mathbf{p}|p^*) \leq 0$, with the equality verified only for $p_1 = p^*$. Furthermore, the expected gambling score in this case is proportional to the expectation of the corresponding Brier score differences $\Delta_B = S^E_B(p_1|p^*) - S^E_B(p_2|p^*)$, thus, they produce, on average, the same rankings.

3.2 Improper use of proper score

When comparing forecasting models, ensuring that the score is proper may not be sufficient. It also has to be used properly. The gambling score with $k = 2$ offers a nice example of this situation. We have demonstrated that the parimutuel gambling score is proper when $k = 2$, however, the dependence of the score value on all the forecasts involved in the comparison is a source of bias. In fact, $S^E_{G,1}(\mathbf{p}|p^*) \geq S^E_{G,2}(\mathbf{p}|p^*)$ when p_1 is closer to p^* than p_2, however, $p_1 = p^*$ does not maximize $S^E_{G,i}(\mathbf{p}|p^*)$ as shown in Figure 2. This means that the score becomes biased when we rank forecasts based on the score difference against a reference model.

Formally, we are considering pairwise vectors $\mathbf{p}_1 = (p_1, p_0)$, $\mathbf{p}_2 = (p_2, p_0)$, etc., where p_0 is the reference model. For each of these we can estimate pairwise comparison score vectors $S_G(\mathbf{p}_1|x)$, $S_G(\mathbf{p}_2|x)$, and so on. The first component of each vector, namely $S_{G,1}(\mathbf{p}_1|x)$, $S_{G,1}(\mathbf{p}_2|x)$, etc, represents the score of p_1 and, respectively, p_2 against the reference model p_0. At this point, one would be tempted to rank the models based on $S_{G,1}(\mathbf{p}_1|x)$ and $S_{G,1}(\mathbf{p}_2|x)$, and this is the approach taken in

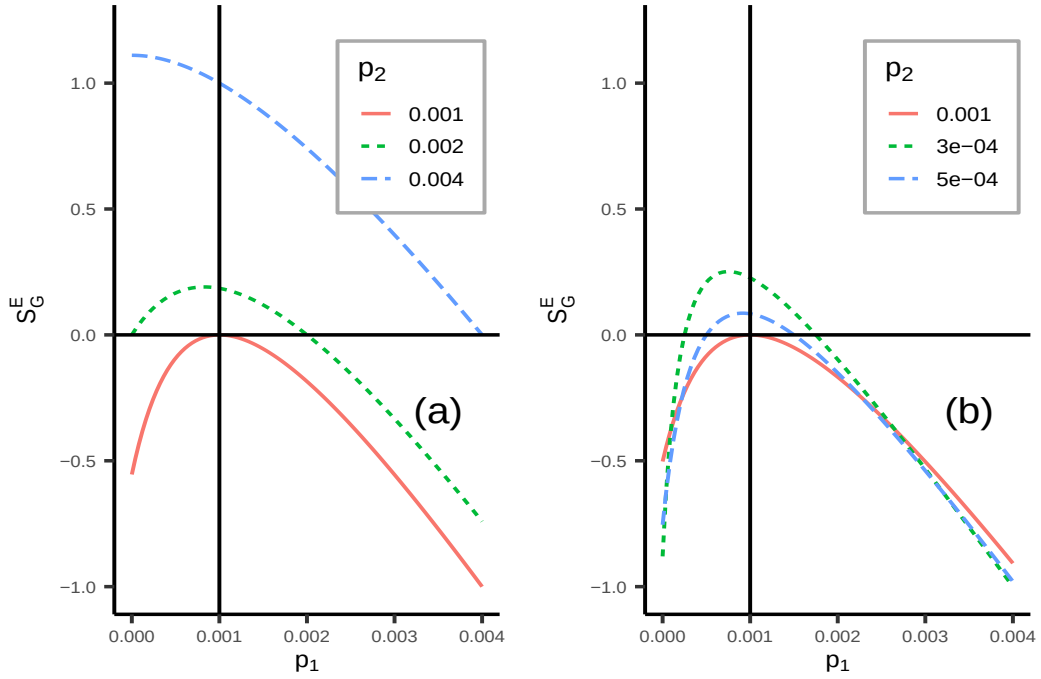

Figure 2. Expected value of the parimutuel gambling score (k = 2), $S^E_{G,1}(\mathbf{p}|p^*)$, varying $p_1 \in (0, 0.004)$, $p_2 = \{p^*, 2p^*, 4p^*\}$ (a) and $p_2 = \{p^*, p^*/2, p^*/4\}$(b). The solid vertical line represents the true probability $p^* = 0.001$. The expected scores have been normalized so that their minimum is equal to -1.

Taroni et al (2014) in which the official national time-independent model (Gruppo di Lavoro , 2004) is used as reference model.

If the parimutuel gambling score is used to rank forecasts based on the score difference relative to a reference model, it will *not* reliably favour the model closest to the true one, and the size of the bias will depend on the choice of the reference model. For example, in Figure 2a, for $p_2 = 0.004$, the gambling score is maximized at $p_1 = 0$. This means that if the reference model is $p_0 = 0.004$, the overconfident forecast $p_1 = 0$ would be favoured by the ranking even if another forecast is perfect, e.g. $p_3 = p^*$. This problem can be particularly relevant in operational seismology where it is common for candidate forecasts to be compared against a reference model which is known to be based on simplistic assumptions (for example a homogeneous Poisson process).

Hereon, the term pairwise gambling score refers to the comparison against a reference model as described in this section, while the term full gambling score will refer to the case where the forecasts compete directly against each other as we describe in the next section. Using this terminology, the full gambling score with $k = 2$ is the only proper score.

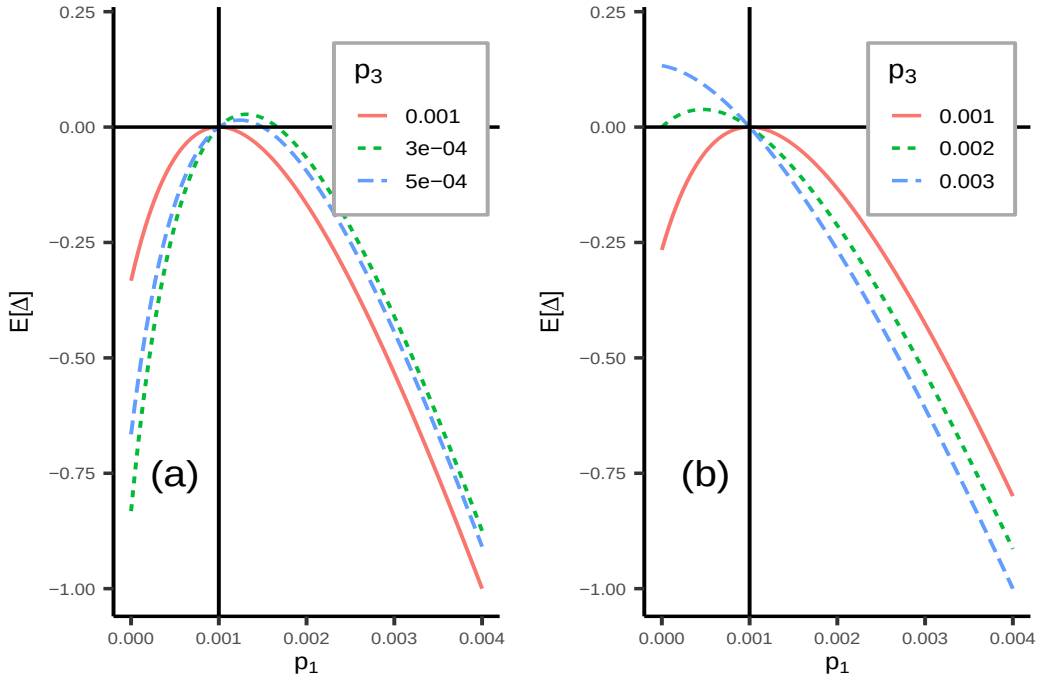

Figure 3. Expected gambling score differences (k = 3) between p_1 and p_2, $S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)$, as a function of $p_1 \in (0, 0.004)$, $p_2 = p^* = 0.001$ (vertical line), and $p_3 \in \{p^*, p^*/2, p^*/3\}$ (a) and $p_3 \in \{p^*, 2p^*, 3p^*\}$ (b) .

3.3 Improperness of the multi-forecast gambling score for $k \geq 3$

The generalized version of the full parimutuel gambling score, as presented in Zechar and Zhuang (2014), for $k \geq 3$ is improper. For example, when $k = 3$ and $p_2 = p^*$, following equation 12 the difference between the expected score for p_1 and p_2 is given by

$$3\bar{p}(1-\bar{p})[S^E_{G,1}(p_1, p^*, p_3|p^*) - S^E_{G,2}(p_1, p^*, p_3|p^*)] = 3(p_1 - p^*)(p^* - \bar{p})$$
$$= (p_1 - p^*)(2p^* - p_1 - p_3),$$

with both sides scaled by the common factor $3\bar{p}(1-\bar{p})$. This means that when $2p^* - p_3 \geq p_1$, the forecast p_1 will have a positive score for any $p_1 \geq p^*$. Any value of $p_1 \in [p^*, 2p^* - p_3]$ will be preferred to p_2 that is equal to p^*. When $2p^* - p_3 \leq p_1$, with the same reasoning, p_1 is preferred over $p_2 = p^*$ in the interval $[2p^* - p_3, p^*]$.

In Figure 3 we consider $k = 3$, $p^* = p_2 = 0.001$ and report the difference between the expected scores of p_1 and p_2, namely $S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)$, for different values of p_3. The expected score difference is not maximize at $p_1 = p^*$, which means that the score is biased, and the "direction" of the bias depends on p_3 being greater than or equal to p^*.

Consider $k > 3$ gamblers who propose probabilities $\mathbf{p} = \{p_1, ..., p_k\}$. It is helpful to consider the

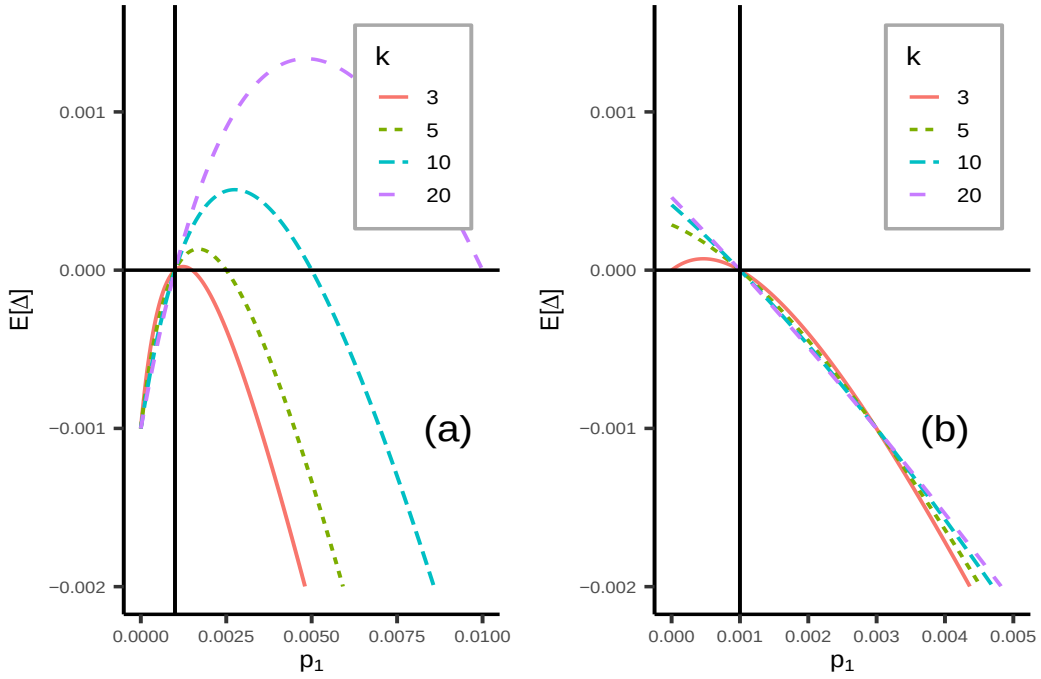

Figure 4. Expected gambling score differences ($k \in \{3, 5, 10, 20\}$) between p_1 and p_2 as a function of p_1 considering $p_2 = p^* = 0.001$ (vertical line). The averageof the forecast probabilities (excluding the first forecast) is the constant $\bar{p}_{-1} = p^*/2$ (a) and $\bar{p}_{-1} = 2p^*$ (b).

vector of probabilities that excludes the first component; we name this $\mathbf{p}_{-1} = \mathbf{p}/\{p_1\}$ and its mean $\bar{p}_{-1}$. Assuming $p_2 = p^*$, thus $\mathbf{p} = (p_1, p^*, ..., p_k)$, we have that

$$k\bar{p}(1-\bar{p})[S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)] = k(p_1 - p^*)(p^* - \bar{p}),$$
$$= (p_1 - p^*)[kp^* - p_1 - (k-1)\bar{p}_{-1}],$$

from which we conclude that the expected score difference is positive when $p_1 \in [kp^* - (k-1)\bar{p}_{-1}, p^*]$ or $p_1 \in [p^*, kp^* - (k-1)\bar{p}_{-1}]$, depending on if $p^* \gtrless \bar{p}_{-1}$. Specifically, when $p^* > \bar{p}_{-1}$ (Figure 3a) the first gambler is encouraged to bet on a $p_1 > p^*$ and vice versa when $p^* < \bar{p}_{-1}$ (Figure 3b). Furthermore, p^* is always an extreme of the interval where the expected score difference is positive. Considering $\bar{p}_{-1}$ as fixed, the length of the interval is an increasing function of the number of gamblers k (Figure 4) which means that the size of the set of forecasts capable of obtaining a score value higher than the data generating model is an increasing function of the number of forecasts involved in the comparison. It is clear that the multi-forecast parimutuel gambling score favours models that are contrary to the average of the other forecasts. This could be particularly dangerous when evaluating the performance of earthquake forecasting models. For example, the trigger for an alarm being broadcast (or not) is often defined when the probability of having an earthquake above a certain magnitude exceeds a specified threshold. Using a model chosen looking at the full parimutuel gambling score

could therefore lead to broadcasting alarms when they are not needed ($p \gg p^*$, 'crying wolf', Figure 4a) or not broadcasting an alarm when needed ($p \ll p^*$, providing 'false reassurance', Figure 4b).

The root of the problems with this score is that the score, relative to a candidate forecast, explicitly depends on the other forecasts. This design brings two problems: (i) the score, even in the special case when it is proper, can be used improperly and (ii) the score is never proper when considering more than two models. The Brier and log scores do not suffer from the same problem since the score of a forecast depends only on the forecast and the observation. Furthermore, the improperness demonstrated here can be expressed in terms that show that the gambling metaphor is part of the problem: If the outcome $x = 0$ is likely (e.g. $p^* = 0.001$) and the majority of the forecasts have too large probabilities, then the expected gain is higher for an overconfident forecast, $p \ll p^*$, since that will give the forecaster a larger share of the total payout.

4 FORECASTING ACROSS MULTIPLE BINS

Until now, we have analysed the expected score for a single bin, here we analyse the ability to express a preference between two forecasts using the average score across multiple bins. We assume to have access only to one observation $x_i \in \{0, 1\}$ per bin . We analyse extensively the case in which the probability of observing $x_i = 1$, is the same for each bin, $p_i^* = p^*$ for any $i = 1, ..., N$. We refer to this as the Multiple Bins Single Probability case; the only quantity of interest is p^* and a forecast is represented by a single value p. Even though, the Multiple Bins Single Probability case is clearly unrealistic in practice, it builds the basic concepts we will then use to explore the Multiple Bins Multiple Probabilities case where the probability of observing $x_i = 1$ is potentially different for each bin.

Considering multiple bins, we observe a realization of the random variable $X_i \sim \text{Ber}(p_i^*)$ for $i = 1, ..., N$. A forecast is given by the vector $\mathbf{p} = (p_1, ..., p_N)$ specifying the probability of $X_i = 1$ for each bin. Following the terminology in the literature regarding Bernoulli random variables, the event $X_i = 1$ is referred to as a *success*. The quantity $X_S = \sum_i X_i$ is therefore referred as the sum of the observations or the number of successes or the number of active bins.

Given an arbitrary scoring rule $S(p|X)$, the average score associated with the forecast $\mathbf{p}$ is given by:

$$S(\mathbf{p}|\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} S(p_i|X_i).$$

The quantity $S(\mathbf{p}|\mathbf{X})$ is a random variable itself, because it is a function of random variables $\mathbf{X}$. To

compare two forecasts $\mathbf{p}_1$ and $\mathbf{p}_2$, we study their score difference:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) = \frac{1}{N}\left(\sum_{i=1}^{N} S(p_{1i}|X_i) - \sum_{i=1}^{N} S(p_{2i}|X_i)\right),$$
$$= \frac{1}{N}\sum_{i=1}^{N} \Delta(p_{1i}, p_{2i}, X_i).$$

The quantity $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})$ is also a random variable as it too depends on the vector of random variables $\mathbf{X}$. If $S(p|X)$ is a proper scoring rule, and if the expected value of the score difference is positive, namely $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})] > 0$, the forecast $\mathbf{p}_1$ is "closer" to the true $\mathbf{p}^*$ than the alternative forecast $\mathbf{p}_2$. The expected value should be considered with respect the distribution of the observations $\mathbf{X}$. However, we do not observe the full distribution - we only observe a sample (i.e. we observe the quantity $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ which is a realization of the random variable $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})$). Even if the expected score difference $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})]$ is positive, which means that we should express a preference for the first forecast, the observed score difference $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ may be negative and lead to the opposite conclusion. To avoid this problem we need to account for the uncertainty around the observed $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ which is the point estimate of the expected score difference $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})]$.

4.1 The distribution of score differences - Multiple Bins Single Probability

For the Multiple Bins Single Probability case, the observation in each bin is a binary random variable $X_i \sim \text{Ber}(p^*)$, $i = 1, ..., N$. Given an arbitrary scoring rule $S(p|X)$ and two candidate forecasts p_1 and p_2, the score difference for the i-th bin is a discrete random variable with distribution:

$$\Delta(p_1, p_2, X_i) = \begin{cases} \Delta_0 = S(p_1|0) - S(p_2|0) & \text{with probability} \quad 1 - p^*, \\ \Delta_1 = S(p_1|1) - S(p_2|1) & \text{with probability} \quad p^*. \end{cases}$$

The forecasts are ranked based on the average score difference across all bins:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \Delta(p_1, p_2, X_i),$$
$$= \frac{1}{N}\sum_{i=1}^{N} \left(\Delta_0 + X_i(\Delta_1 - \Delta_0)\right),$$
$$= \Delta_0 + \frac{X_S}{N}(\Delta_1 - \Delta_0),$$

where, $X_S = \sum_i X_i$ is the sum of all observations or, equivalently, the total number of successes. By definition, X_S is the sum of N (assumed to be) independent and identically distributed Bernoulli trials X_i. Therefore, X_S has a Binomial distribution with size parameter N, the number of bins, and probability parameter p^*. When we observe a sample $x_1, ..., x_N$, the observed score difference is given

by:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x}) = \Delta_0 + \frac{\mathrm{x}_S}{N}(\Delta_1 - \Delta_0),$$

where x_S is a realization of the random variable X_S. The observed score difference depends on the observations only through the quantity x_S/N. Thus, it is enough to study the quantity x_S/N to make inference about the expected value of the score difference. The quantity x_S/N it is said to be *sufficient* (Fisher , 1922) with respect to the expected score difference because it contains all the information provided by the observations $x_1, ..., x_N$ on the parameter of interest (in this case, the expected score difference $\mathbb{E}[\Delta(p_1, p_2, X)]$). For an introduction to statistical inference and the theory behind we refer to Schervish (2012); Hastie et al (2009).

4.2 Confidence Intervals for the Expected Score Difference

A way to account for the uncertainty around the observed score difference is to consider an interval estimate of the expected value of the score difference. Once a sample $\mathbf{x} = x_1, ..., x_N$ has been observed and the confidence interval calculated, if the entire interval lies above zero we express a preference towards p_1, alternatively if it lies below zero we express a preference towards p_2. If the interval contains the value zero we conclude that the observed sample does not contain enough information to express a preference. It is important to consider the latter case as a possible outcome because it is an indication that we need to collect more data or that the forecasts perform similarly (as measured by the score) and provide an additional information than the pure rankings.

We are considering the confidence interval for the expected value of the score difference:

$$\begin{aligned}\mathbb{E}[\Delta(p_1, p_2, \mathbf{X})] &= \Delta_0 + \frac{\mathbb{E}[X_S]}{N}(\Delta_1 - \Delta_0), \\ &= \Delta_0 + p^*(\Delta_1 - \Delta_0). \end{aligned} \tag{13}$$

Having an observation $\mathbf{x} = x_1, ..., x_N$ per bin, the point estimate of $\mathbb{E}[\Delta(p_1, p_2, X)]$ is the observed score difference:

$$\Delta(p_1, p_2, \mathbf{x}) = \Delta_0 + \hat{p}(\Delta_1 - \Delta_0), \tag{14}$$

where $\hat{p} = \mathrm{x}_S/N$ is the observed probability of success. Comparing the equations 13 and 14, the point estimate of the score difference is retrieved plugging in the point estimate of the probability of success $\hat{p}$ in place of p^*. In the same way, to retrieve an interval estimate of the expected score difference is sufficient to retrieve an interval estimate of the probability of success p^*.

Therefore, we need the confidence interval of level α for the true probability p^* given observations $x_1, ..., x_N$ from a Ber(p^*), namely $CI_{p^*}(\alpha) = (\hat{p}_L, \hat{p}_U)$, and plug those values into expression 14 to obtain a confidence interval for $\mathbb{E}[\Delta(p_1, p_2, X)]$, namely $CI_\Delta(\alpha) = (\Delta_L, \Delta_U)$. Various methods have

been found to estimate $\hat{p}_L$ and $\hat{p}_U$, most of them relying on a Gaussian approximation. However, this approximation is not reliable for small sample sizes (number of bins N) and for values of p^* close to zero or one, as in our case (Wallis , 2013).

Hereafter, we use the Clopper-Pearson confidence interval (Clopper and Pearson , 1934). This method is referred to as *exact* because it relies on cumulative binomial probabilities rather than an approximation and is therefore more efficient and accurate than simulation based methods. The confidence interval with level α for $\hat{p}$ is given by:

$$p_L(\alpha) = \text{BetaQ}(\frac{\alpha}{2}; \mathrm{x}_S, N - \mathrm{x}_S + 1),$$
$$p_U(\alpha) = \text{BetaQ}(1 - \frac{\alpha}{2}; \mathrm{x}_S + 1, N - \mathrm{x}_S),$$

where the function $\text{BetaQ}(q; a, b)$ is the q-th quantile of a Beta distribution with parameters a and b. We can construct confidence intervals for $\mathbb{E}[\Delta(p_1, p_2, \mathbf{X})]$ as follows:

$$\Delta_L = \Delta_0 + \hat{p}_L(\Delta_1 - \Delta_0),$$
$$\Delta_U = \Delta_0 + \hat{p}_U(\Delta_1 - \Delta_0).$$

The obtained confidence interval for p^* depends on the data only through the sum of the observations x_S, which is a sufficient statistic for the problem. Similarly, the confidence interval for $\mathbb{E}[\Delta(p_1, p_2, \mathbf{X})]$ depends on the data through the value of the sufficient statistic, x_S.

Figure 5 shows the confidence interval for the score difference as a function of the sum of observations x_S considering two competing forecasts $p_1 = 0.001$, $p_2 = p_1/3$, a reference model for the pairwise gambling score $p_0 = 5p_1$ and $N = 10,000$ bins. Here, we do not need to choose a value for p^*. Indeed, the confidence interval is determined solely by the forecast and observation. The Brier, log and full gambling score(Figure 5a, 5c, 5d) all express a preference for p_1 if we observe $\mathrm{x}_S > 12$, while they express a preference for p_2 when $\mathrm{x}_S < 2$. This result is expected because $p_1 > p_2$, which means that $\mathrm{x}_S > 12$ is much more probable under p_1 than p_2. In fact, the average number of successes using p_1 is $Np_1 = 10$ while $Np_2 = 3.34$. The same reasoning applies when we express a preference for p_2 ($\mathrm{x}_S < 2$).

The pairwise gambling score (Figure 5 (b)), instead, requires $\mathrm{x}_S > 24$ to express a preference for p_1 and $\mathrm{x}_S < 9$ to express a preference for p_2. It is heavily biased toward the forecast closer to zero. In fact, when p_1 is the true probability, the probability of observing $\mathrm{x}_S > 24$ is less than 0.0001 and the probability of observing $\mathrm{x}_S < 9$ is 0.33. Therefore, we are more likely to express a preference for p_2 than for p_1, even when $p_1 = p^*$. This reinforces the problems with employ improper scores introduced in Section 3.

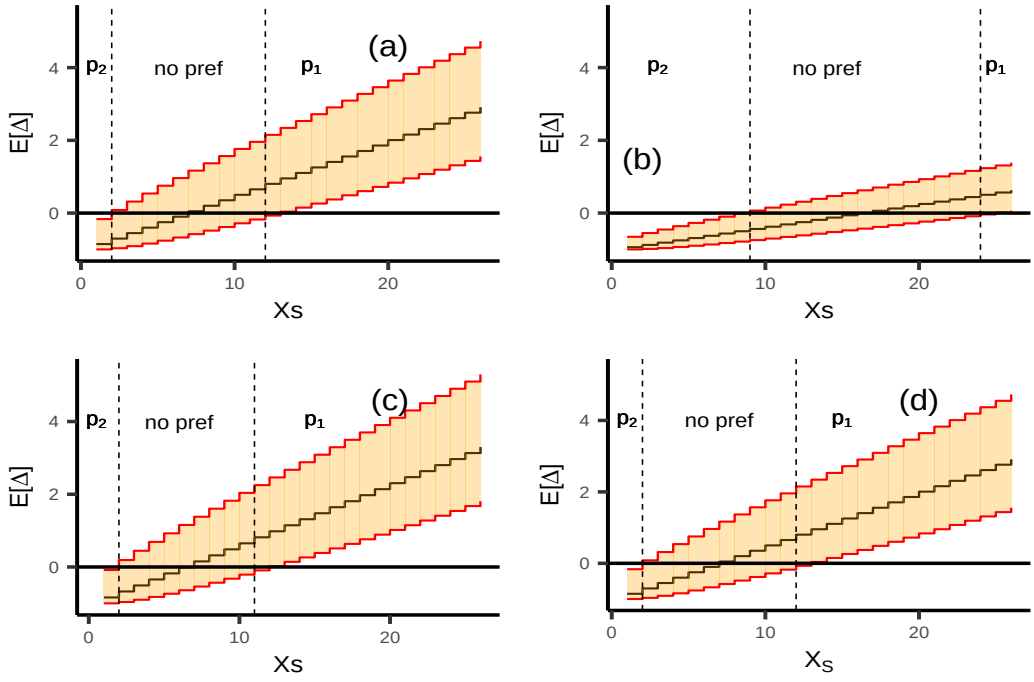

Figure 5. Confidence interval (shaded area) and point estimate (black solid line) for $\mathbb{E}[\Delta]$ as a function of the number of observed successes x_S considering $p_1 = 0.001$, $p_2 = p_1/3$, $p_0 = 5p_1$ and N = 10000. In each plot shows a different score: (a) Brier score; (b) pairwise gambling score; (c) logarithmic score; (d) full gambling score. Black solid line represents the observed score difference while the orange area represents the confidence interval. The black vertical dashed lines represent the interval of values of x_S for which we do not express a preference

4.3 Preference Probabilities

The confidence interval for the expected score difference, $CI_\Delta(\alpha)$, is a function of the competing forecasts, the scoring rule and depends on the data only through the sum of the observations $\mathrm{x}_S = \sum_i x_i$. In particular, there are a range of values (between the dashed lines in Figure 5) of x_S for which we are not able to express a preference. We refer to this interval as $(\mathrm{x}_{min}, \mathrm{x}_{max})$. With respect to the sum of the observations x_S there are only three possible outcomes:

$$\begin{aligned} \mathrm{x}_S < \mathrm{x}_{min} &\longrightarrow \text{preference for } p_2, \\ \mathrm{x}_{max} \leq \mathrm{x}_S \leq \mathrm{x}_{max} &\longrightarrow \text{no preference}, \\ \mathrm{x}_S > \mathrm{x}_{max} &\longrightarrow \text{preference for } p_1. \end{aligned}$$

The values x_{min} and x_{max} are determined solely by p_1, p_2, the number of bins N and the scoring rule. Table 1 reports the values of x_{min} and x_{max} for the scoring rules depicted in Figure 5. These values can be used to compute the preference probabilities once a value for p^* is assumed. Indeed,

Table 1. Multiple Bins Single Probability case: table reporting the values x_{min} and x_{max} for the Brier, log, pairwise gambling (PG) and full gambling (FG) score. The reported values refers to the case where $N = 10,000$, $p_1 = 0.001$, $p_2 = p_1/3$, and do not depend on p^*.

Score	x_{min}	x_{max}
Brier	2	12
Log	2	11
PG	9	24
FG	2	12

in the Multiple Bins Single Probability case, the distribution of X_S is a Binomial distribution, $X_S \sim \mathrm{Bin}(N, p^*)$. Table 2 reports the probabilities of i) no preference; ii) Preference for p_1; iii) Preference for p_2. The probabilities are calculated considering alternatively p^* equal to p_1 (first half of the table) or p_2 (second half of the table).

The similarity among the values x_{min} and x_{max} for the proper scores lead to similar preference probabilities. The proper scores always assign the greatest probability to the case in which we are not able to express a preference, however, when $p^* = p_1$ it is unlikely to express a preference for p_2. Vice versa when $p^* = p_2$. There is a slightly difference between the Brier and the log score coming from the different penalty applied to forecasts close to zero. The log score penalises more heavily forecasts close to zero and, in fact, when $p_1 = p^*$ chances to express a preference for p_1 are higher than using the Brier score. The full gambling score for p_1 against p_2 is proportional to the Brier score difference between p_1 and p_2 (see Section 3.2) and thus, their preference probabilities coincide. Considering the pairwise gambling score the probability of expressing a preference for p_1 is always very close to zero, even when $p^* = p_1$. This shows again that it is possible to find a combination of p_1, p_2 and p_0 such that the model providing the smallest forecast obtains the highest reward with probability over 0.9, even when the other forecast is equal to the true p^*.

Figure 6 shows the preference probabilities as a function of $p^* \in (10^{-6}, 10^{-2})$ which is the range of values of the 5-year adaptively-smoothed forecast for Italy (aggregating over the magnitude bins) used later to illustrate the Multiple Bins Multiple Probabilities case. The Brier score behaves as expected. The probability of expressing a preference for p_2 increases as p^* goes to zero, which is what we expect given $p_2 < p_1$. On the other hand, the probability of preferring p_1 increases when p^* increases, because $p_1 > p_2$. Finally, the probability of not being able to express a preference is higher when $p_2 < p^* < p_1$ (Figure 6a). The pairwise gambling score, instead, does not behave as expected. The probability of preferring p_1 is almost zero in the range of values of p^* considered in the example.

Table 2. Multiple Bins Single Probability case: table reporting for each score (row) the probabilities of expressing (or not) a preference using either the Brier, log-, pairwise gambling (PG) or full gambling (FG) score. The probabilities are calculated considering $N = 10,000$, $p_1 = 0.001$, $p_2 = p_1/3$ and considering two cases: $p^* = p_1$ and $p^* = p_2$.

Score	No pref	Pref p_1	Pref p_2
$p^* = p_1$			
Brier	0.7912	0.2083	0.0005
Log	0.6963	0.3032	0.0005
PG	0.6672	0.0000	0.3327
FG	0.7912	0.2083	0.0005
$p^* = p_2$			
Brier	0.8454	0.0000	0.1545
Log	0.8453	0.0000	0.1545
PG	0.0073	0.2083	0.9927
FG	0.8454	0.0000	0.1545

The two most probable outcomes are: expressing a preference for p_2 or not expressing a preference at all.

4.4 Probability of expressing a preference

It is interesting to study how the probability for each case changes as a function of p^* for different numbers of bins N and different ratios between p_1 and p_2. To do that it is useful to focus only on two possible outcomes: expressing a preference and not expressing a preference. The probability of expressing a preference is given by the probability of observing a sample such that the sum of the observations x_S is greater than x_{max} or smaller than x_{min}. We refer to this probability as β which is given by

$$\beta = 1 - \Pr[\mathrm{x}_{min} \leq X_S \leq \mathrm{x}_{max}].$$

This probability depends on the scoring rule, the forecasts p_1 and p_2,the number of bins N and the true probability p^*. We study β as a function of p^* for different numbers of bins. In this artificial case, to increase the number of bins we are considering additional bins with the same probability, we

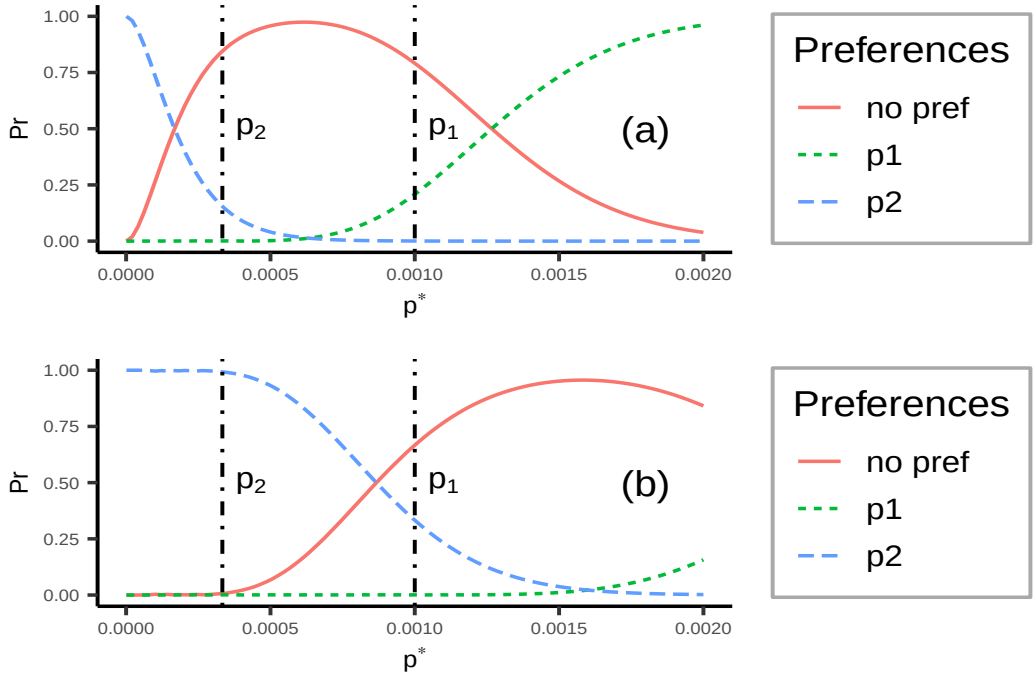

Figure 6. Multiple Bins Single Probability case: each plot shows the probability of each possible outcome (solid lines no preference, dotted lines preference for p_1, dashed lines preference for p_2) as a function of p^* using the Brier score (a) and the pairwise gambling score (b) considering $p_1 = 0.001$ and $p_2 = p_1/3$ (vertical lines), $p_0 = 5p_1$ and $N = 10,000$. The true probability p^* varies in $(10^{-6}, 2 \cdot 10^{-2})$ which is a realistic range of values in Italy.

are explicitly not splitting any bin; this is analogous to increase the data at hand applying the model to a larger spatio-temporal region.

Figure 7a considers only the Brier score. The region of p^* presenting low values for β shrinks when the number of bins increase which simply means that the more data we have, the more chances of expressing a preference. Moreover, β is at the minimum when $p^* \in (p_2, p_1)$, which is reasonable because if the distances $|p^* - p_1|$ and $|p^* - p_2|$ are similar the probability of no preference should be high. The $N = 2000$ can be explained considering $p_1 = p^*$. In this case, the expected sum of observations is $Np_1 = 2$ and it is more probable to observe $X_S < 2$ than $X_S > 2$. Given that $Np_2 < Np_1$, the probability of not expressing a preference is high.

Figure 7 presents the probability β as function of p^* for different scores with a fixed number of bins $N = 5000$ (b) and $N = 20000$ (c). For $N = 5000$, the proper scores (Brier, log and full gambling score) present the same values of β for any value of p^*. For $N = 20000$, the proper scores start to behave differently. The Brier and full gambling score still coincide, while the log score is slightly different. Specifically, the log score presents higher β values when $p^* = p_1$, and lower when $p^* = p_2$.

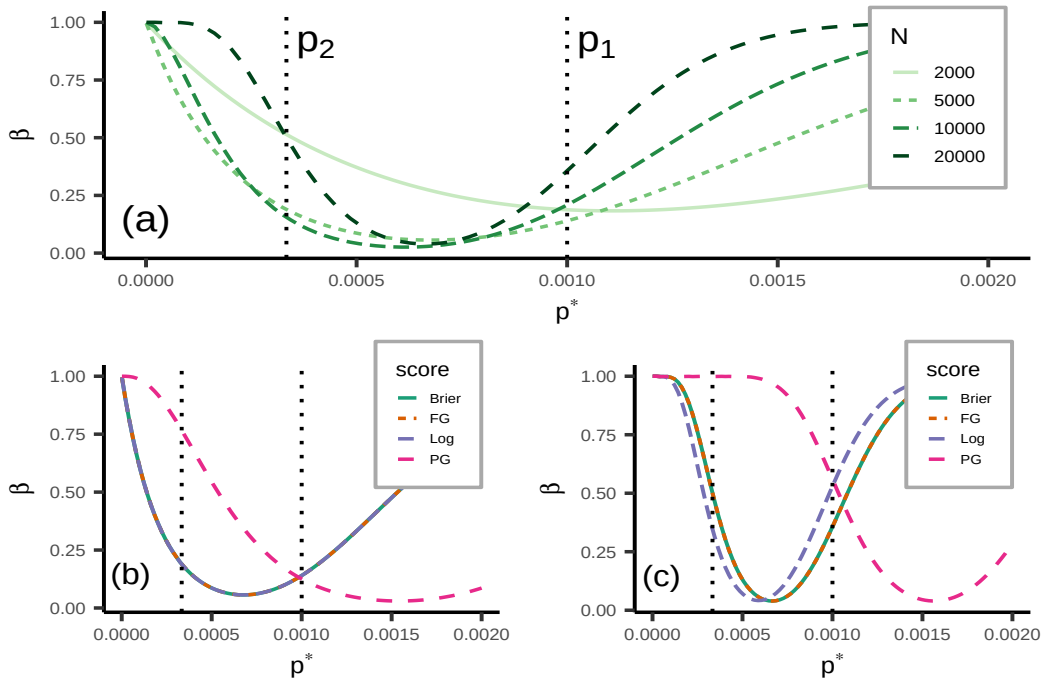

Figure 7. Multiple Bins Single Probability case: (a) Brier score preference probability as a function of p^* for different numbers of bins $N \in \{2000, 5000, 10000, 20000\}$. (b-c) Probability of expressing a preference as a function of p^*. Colors represent the different scores: Brier, log , pairwise gambling (PG), and full gambling (FG) score. The Brier and FG scores coincide. The number of bins is fixed to N = 5000 (b) and N = 20000 (c). We set $p_1 = 0.001$, $p_2 = p_1/3$ (vertical lines), $p_0 = 5p_1$, and $p^* \in (10^{-6}, 2^{-3})$ which is a realistic range of values in Italy.

This depends on the different penalties applied to forecasts close to zero. The log score presents greater chances of expressing a preference for p_1 when $p^* = p_1$ because the other forecast p_2 is smaller than p_1 and, therefore, penalized. On the other hand, when $p^* = p_2$ the log score presents smaller β values than the Brier score.

In contrast to the proper scores, the pairwise gambling score reaches its minimum β value for $p^* > p_1$. Here, the pairwise gambling score tends to express a preference for the smaller forecast even when the other one is closer to p^*. This leads to higher values of β when $p^* \in (p_2, p_1)$ because the pairwise gambling score will likely express a preference for p_2. Only when $p^* > p_1$ the probability of no preference grows and the value of β decreases accordingly.

Given that p_1 and p_2 are scalars, we can consider β as a function of the ratio between p_1 and p_2, $\omega = p_2/p_1$, for a fixed p^*. In principle, we expect that β is an increasing function of ω. We assume that the first forecast and the true probability are identical $p_1 = p^* = 0.001$. The reference model for the pairwise gambling score is $p_0 = 5p_1$ and we consider different numbers of bins $N \in$

$\{2000, 5000, 10000, 20000\}$. The ratio $\omega = p_2/p_1$ varies in the interval $(0.1, 4)$. We expect low β values when ω is around one (similar forecast) and high β values otherwise.

Figure 8a shows that, as expected, for $N > 2000$, β has its minimum when $\omega = 1$. Considering ω as fixed, β is an increasing function of the number of bins. Figure 8b-c compares the β values relative to different scores for a fixed number of bins, $N = 5000$ (b) and $N = 20000$ (c). The Brier and full gambling score coincide, whilst the log score presents slightly different β values. As before, this is due to the different penalties applied to forecasts close to zero.

The pairwise gambling score is not consistent with the trends in the proper scores. Using this score and considering $N = 20000$ (Figure 8c), the probability β is consistently greater than 0.5 for the considered values of ω. Considering that $p_1 = p^*$, the quantity ω is also the ratio between p_2 and p^*. This implies that regardless of ω, we will erroneously express a preference for p_2 with a probability above 0.5.

Importantly, these sanity checks of a proposed scoring procedure can be done before looking at the observations. It is possible to check if forecasts can, in principle, be distinguished in light of the amount of expected data. We recommend the use of such exploitative figures when introducing a new scoring rule whose performance have not been tested. If the proposed scoring rule does not behave acceptably in this simple scenario, it is unlikely that it would behave acceptably in a real application.

4.5 Score difference distribution - Multiple Bins Multiple Probabilities

The Multiple Bins Multiple Probabilities case generalizes the Multiple Bins Single Probability case, and is much more similar to a real earthquake forecasting experiment. For example, the forecasts involved in the first CSEP experiments (Field , 2007; Schorlemmer and Gerstenberger , 2007b; Zechar et al , 2013; Michael and Werner , 2018) were mostly grid-based forecasts providing for each space-time-magnitude bin, the expected number of earthquakes. Then, the number of events in each bin is modelled using a Poisson distribution with intensity equal to the number of events provided by the forecasts and the probability of observing at least one event is calculated accordingly. In this scenario, we do not have analytical results for the score difference distribution and we need to recur to simulations.

We now want to specify a true model which has more realistic probabilities. Since we do not actually know these in reality, we choose to work with one of the CSEP models that was submitted to the 2010 Italy experiment (Taroni et al , 2018). We choose to simulate synthetic data from the 5-year adaptively-smoothed forecast for Italy (Werner et al , 2010) and explore the ability of the scoring rules to discriminate between linearly scaled versions of this true model. This means that we are considering only one time bin of size 5 years, while the space-magnitude domain is divided in multiple

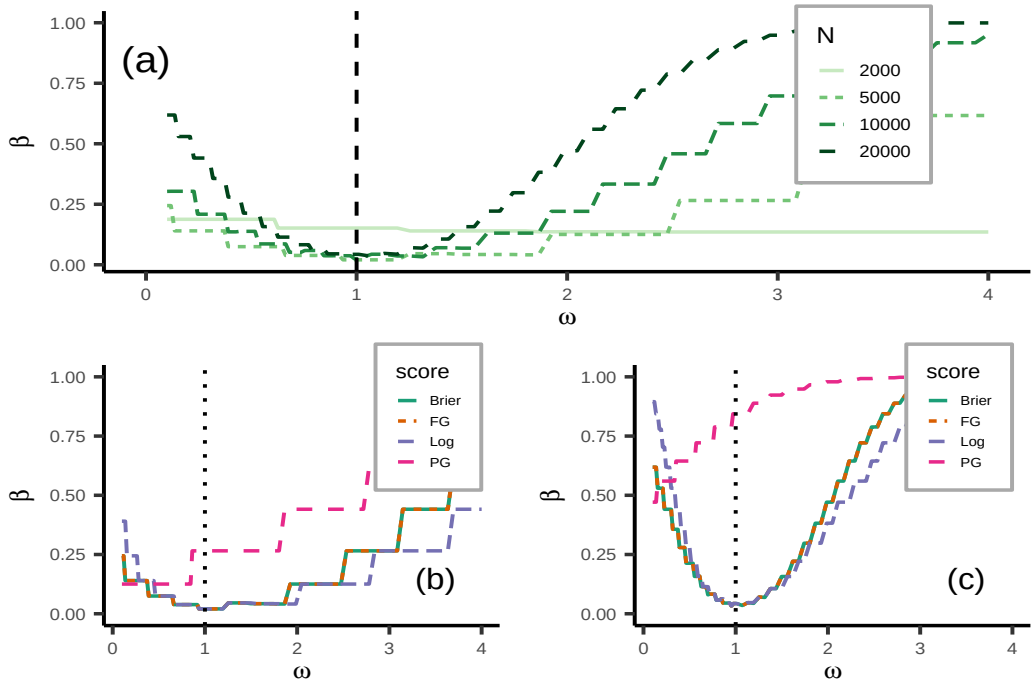

Figure 8. Multiple Bins Single Probability case: (a) Probability of expressing a preference using the Brier score as a function of $\omega = p_2/p_1 \in (0.1, 4)$ for different numbers of bins $N \in \{2000, 5000, 10000, 20000\}$. We set $p_1 = p^* = 0.001$, and the reference model is $p_0 = 5p_1$. (Bottom) Probability of expressing a preference as a function of ω. Colors represent the different scores: Brier, log, pairwise gambling (PG), and full gambling (FG) score. The number of bins is fixed to N = 5000 (b) and N = 20000 (c).

regular bins. The spatial domain is represented by the coloured area in Figure 9 and it is divided in 0.1×0.1 longitude-latitude bins. The magnitude domain ranges from 4.95 to 9.05 magnitude units and is divided in bins of length 0.1. The forecast is relative to the period from January 1, 2010, to December 31, 2014.

The adaptively-smoothed forecast provides the expected number of earthquakes in each space-magnitude bin. For each bin, to calculate the probability of observing at least one earthquake, in accordance with the methodology in the 2010 Italy CSEP forecast experiment, we consider a Poisson distribution for the number of events with intensity given by the predicted number of events. Assuming independence in the magnitude bins, we can aggregate the probabilities over magnitude bins and, for each space bin, obtain the probability of observing at least an earthquake in the period of interest with magnitude greater, or equal, to 4.95. Figure 9 shows the forecasted log-probability for each spatial bin used as data generating model.

The Italian adaptively-smoothed forecast reported in Figure 9 is the vector of true probabilities $\mathbf{p}^* = p_1^*, ..., p_N^*$, where $N = 8993$. As in the previous sections, we compare two forecasts $\mathbf{p}_1 = \mathbf{p}^*$

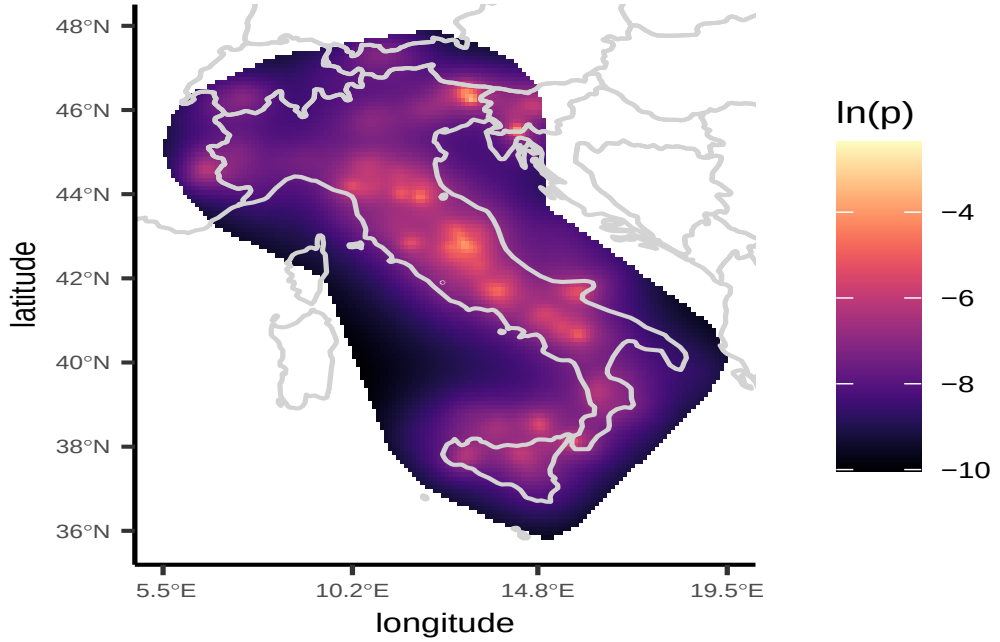

Figure 9. 5-year adaptively-smoothed forecast for Italy (Werner et al (2010)). The figure shows for each spatial bin the natural logarithm of the probability of observing at least one earthquake at or above magnitude 4.95 in the period from January 1, 2010, to December 31, 2014.

and $\mathbf{p}_2 = \omega\mathbf{p}^*$. We will be ignoring the spatial configuration. The average bin score difference is given by

$$\begin{aligned}\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) &= \frac{1}{N}\sum_{i=1}^{N}(\Delta_{0,i} + X_i(\Delta_{1,i} - \Delta_{0,i})),\\ &= \bar{\Delta}_0 + \frac{1}{N}\sum_{i=1}^{N} X_i(\Delta_{1,i} - \Delta_{0,i}),\end{aligned}$$

where, $\Delta_{0,i} = \Delta(p_{1i}, p_{2i}, 0)$ and $\Delta_{1,i} = \Delta(p_{1i}, p_{2i}, 1)$ are, respectively, the score difference in the i-th bin in case we observe $X_i = 0$ (no earthquake at or above magnitude 4.95 during the 5 years) or $X_i = 1$ (at least one earthquake above magnitude 4.95 during the 5 years). The quantity $\bar{\Delta}_0$ is the average $\Delta_{0,i}$. The observations $X_i \sim \text{Ber}(p_i^*)$ follow a Bernoulli distribution, each bin has a potentially different parameter $p_i^* \neq p_j^*$ for any $i \neq j$. The expected value of the score difference is given by

$$\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})] = \bar{\Delta}_0 + \frac{1}{N}\sum_{i=1}^{N} p_i^*(\Delta_{1,i} - \Delta_{0,i}).$$

Table 3. Expected score difference considering $\mathbf{p}^*$ equal to the 5 year Italy adaptively-smoothed forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$ and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$. The scores considered are: the Brier score, the log score, the pairwise gambling (PG) score and the full gambling (FG) score.

Score	$\mathbb{E}[\Delta]$
Brier	0.0000026
Log	0.0003137
PG	-0.0000900
FG	0.0002422

Given that we are considering $\mathbf{p}_1 = \mathbf{p}^*$ and $\mathbf{p}_2 = \omega\mathbf{p}^*$, the expected score difference is non-negative if a proper scoring rule is used while it could be negative if the scoring rule is improper. Specifically, we show that it is possible to find a reference model $\mathbf{p}_0$ such that, if used in combination with the parimutuel gambling score to rank the forecasts, the expected score difference is negative. As before, the Brier, log and full gambling score are used for comparison. In Table 3 we report the expected score differences considering different scores. As expected, they are all positive except for the pairwise gambling score.

In Figure 10 is showed the expected score difference as a function of the forecasts ratio $\omega \in [10^{-3}, 4]$. The results are similar to the ones reported in Figure 1 and 2. The Brier, log and full gambling scores behave suitably, while the pairwise gambling score does not. The Brier and full gambling score are bounded and they prefer a forecast $\mathbf{p} = 10^{-3}\mathbf{p}^*$ to $\mathbf{p}' = 4\mathbf{p}^*$. Indeed, in Figure 10 the left hand side is greater than the right hand side. That is because the penalty is based on the absolute difference between a forecast and the data generating model, therefore, a forecast $\mathbf{p} = 10^{-3}\mathbf{p}^*$ is preferred to $\mathbf{p}' = 4\mathbf{p}^*$, because $\|10^{-3}\mathbf{p}^* - \mathbf{p}^*\| \leq \|4\mathbf{p}^* - \mathbf{p}^*\|$. On the other hand, the log score is unbounded and is based on the relative difference. With the log score, a forecast $\mathbf{p}' = 4\mathbf{p}^*$ is preferred to $\mathbf{p} = 10^{-3}\mathbf{p}^*$ because $\|\mathbf{p}^*/10^{-3}\mathbf{p}^*\| > \|\mathbf{p}^*/4\mathbf{p}^*\|$. The pairwise gambling score, instead, is heavily biased towards zero.

We can extend the comparison by considering $k = 3$ forecasts. In this case, we consider the reference model $\mathbf{p}_0 = 5\mathbf{p}^*$ as third competitor. Figure 11, for each scoring rule, shows the expected score differences $\mathbb{E}[\Delta(\mathbf{p}_2, \mathbf{p}_1, \mathbf{X})]$ (dashed blue) and $\mathbb{E}[\Delta(\mathbf{p}_2, \mathbf{p}_0, \mathbf{X})]$ (solid red), representing the expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$, and the expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_0$. Given that $\mathbf{p}_1$ is equal to the true probabilities, the score differences have to be negative for any value of $\omega \neq 1$ in order for the scoring rule to be effective. Indeed, this is the case for the Brier and log score (Figure 11 (a), (c)). On the other hand, both the pairwise and full gambling score (Figure 11 (b), (d))

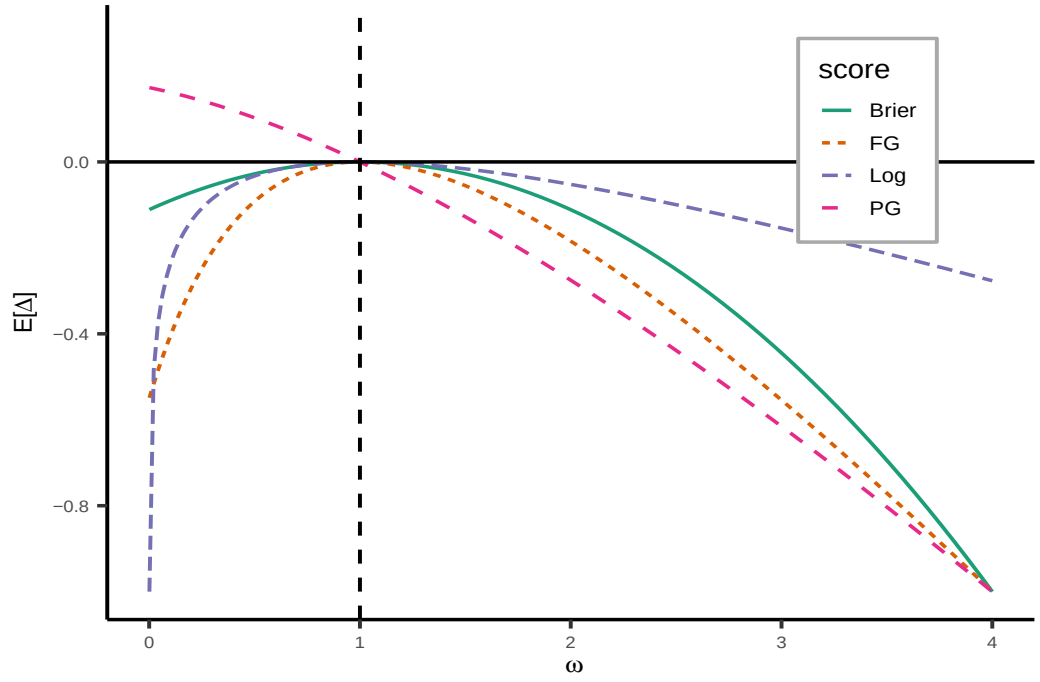

Figure 10. Expected score difference between $\mathbf{p}_1$ and $\mathbf{p}_2$ as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 4)$. We set $\mathbf{p}^*$ equal to the 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$.

are improper and prefer $\mathbf{p}_2$ over $\mathbf{p}_1$ when $\omega \in (0, 1)$. Moreover, all the scores prefer $\mathbf{p}_0$ to $\mathbf{p}_2$ when $\omega > 5$. However, the log score prefers $\mathbf{p}_0$ to $\mathbf{p}_2$ also when ω approaches zero. This shows, again, how different scoring rules apply different penalties to the forecasts.

We note that the pairwise and full gambling score present almost the same expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$. This is because both scoring procedures implicitly assume a reference model given by the average forecast. If the average forecast in a bin is greater than p_i^*, a forecaster will obtain a positive reward each time they submit a value smaller than $\mathbf{p}_i^*$ and $X_i = 0$ occurs. Therefore, given that we are in a low probability environment for which $\Pr[X_i = 0] > 0.99$, the smallest forecast is likely to be preferred. The bias depends on the relationship between the reference model and the true probabilities. In the gambling metaphore, the reference model plays the role of the house (or banker) which determines the returns, and against which all forecasts are competing. Considering equation 12, if $p^* < p_0$, the player has a positive reward forecasting $p < p_0$. If the number of forecasts is large enough that changing a forecast does not affect significantly the average, then, the smaller the forecast the higher the reward, and the forecaster is encouraged by the score to provide $p = 0$. The same reasoning applies if $p^* > p_0$.

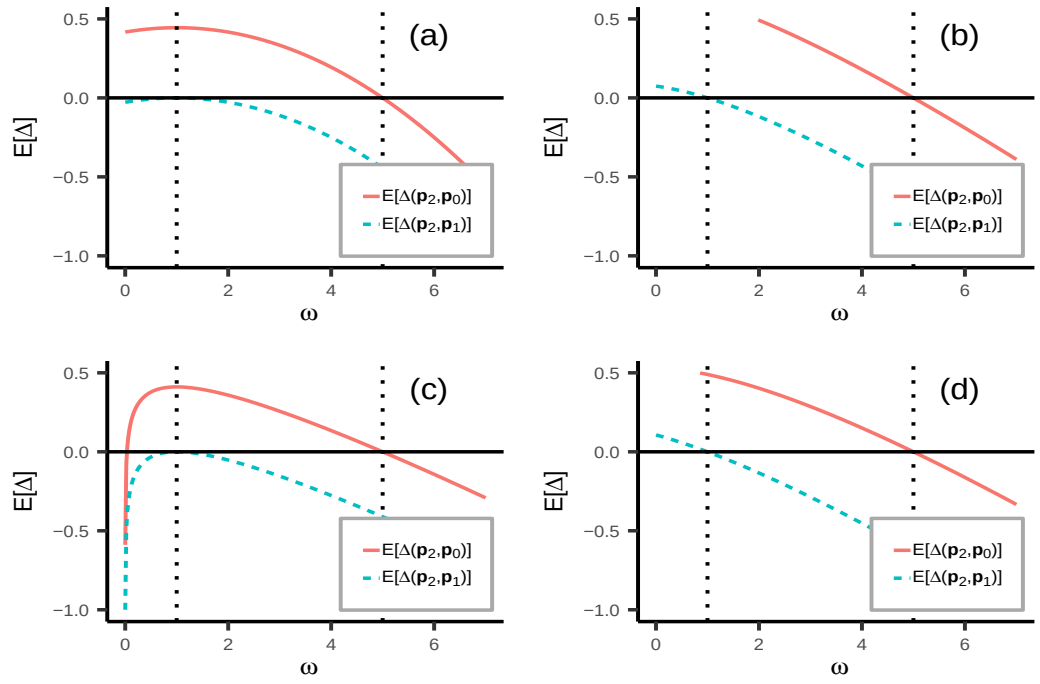

Figure 11. Expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$ (blue dashed) and $\mathbf{p}_2$ and $\mathbf{p}_0$ (red solid) as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 7)$. We set $\mathbf{p}^*$ equal to 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$ and $\mathbf{p}_0 = 5\mathbf{p}^*$. Vertical lines represent $\omega = 1$ and $\omega = 5$. We consider the Brier (a), pairwise gambling (b), log (c), and full gambling (d) scores.

4.6 Confidence Interval and preference probabilities - Multiple Bins Multiple Probabilities

Also in the Multiple Bins Multiple Probabilities case it is crucial to account for the uncertainty around the observed score difference. The binomial formulation used before to retrieve confidence intervals no longer holds, and we need an alternative methodology. One approach to calculate confidence intervals for the expected score difference relies on a Gaussian approximation of the score difference distribution (Rhoades et al , 2011). The score difference in each bin, $\Delta(p_{1i}, p_{2i}, X_i)$ for $i = 1, ..., N$, are assumed to be independent draws from a Gaussian distribution with expected value $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, X)]$ and variance σ^2. When we observe a sample $\mathbf{x} = x_1, ..., x_N$, the point estimate of the expected score difference is the observed score difference $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ and the $(1-\alpha)\%$ confidence interval is given by:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x}) \pm t_{1-\alpha/2, N-1} \frac{s}{\sqrt{N}},$$

where s^2 is an estimate of the variance σ^2 and $t_{1-\alpha/2,N-1}$ is the $1 - \alpha/2$ percentile of a t-student distribution with $N - 1$ degrees of freedom. The reliability of such interval estimates is determined by the accuracy of the Gaussian approximation, which, in turns, depends on the amount of data (the

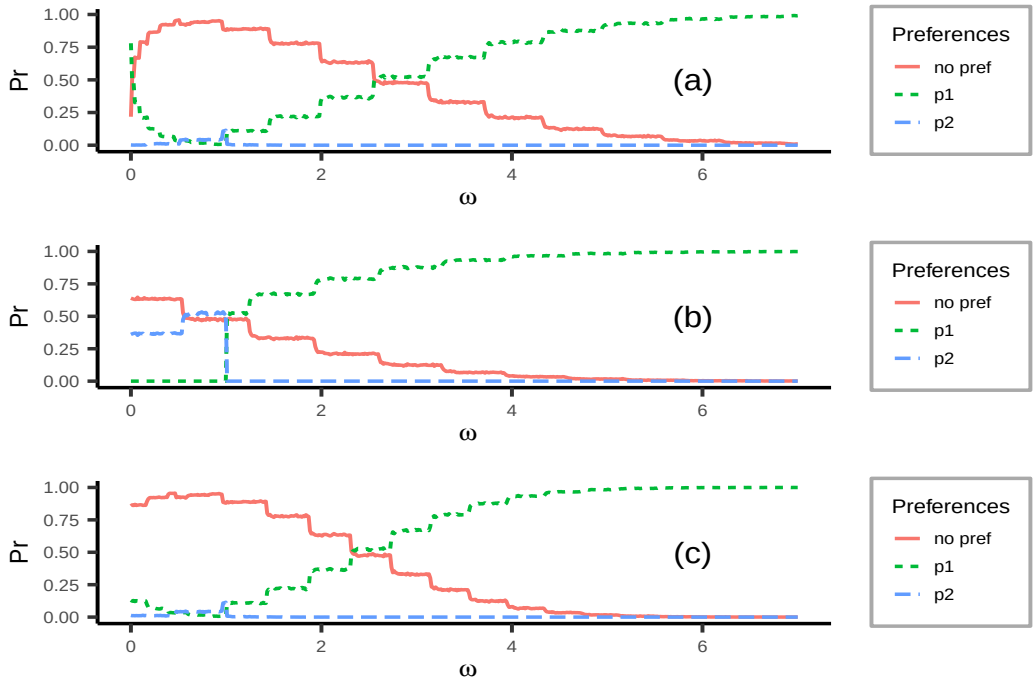

Figure 12. Multiple Bins Multiple Probabilities case: each plot shows the probability of each possible outcome (solid no preference, dotted preference for $\mathbf{p}_1$, dashed preference for $\mathbf{p}_2$) as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 7)$. The log (a), the pairwise gambling (b) and the full gambling (c) scores are considered. We set $\mathbf{p}^*$ equal to 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$.

more the better) and on the correlation between the score difference in each bin (the more the worst). We analyse the reliability of this approximation in Appendix A: Reliability of the gaussian confidence intervals and conclude that it can be used with the log, pairwise gambling and full gambling score but not with the Brier score.

Figure 12 shows the evolution of the preference probabilities varying the forecasts ratio, $\omega = \mathbf{p}_2/\mathbf{p}_1$. It is quite similar to Figure 6 and the same problems with the pairwise gambling score are evident; i.e. it favours forecast smaller than the true probability when the average forecast is greater than the latter. On the other hand, the log score probability of preferring $\mathbf{p}_1$ increases rapidly when $\omega \to 0$, while the full gambling score is not able to distinguish between $\mathbf{p}_1$ and $\mathbf{p}_2$ for $\omega < 2.5$. The latter remark suggests a potential problem with the use of the full gambling score given that, in real forecasting experiments, the competing forecasts tends to be quite similar, in which case there is an high probability of no preference.

This concludes our analysis on the use of proper scoring rules to rank earthquake forecasting models.

5 DISCUSSION

The parimutuel gambling score was introduced as a general scoring rule to compare, within a unified framework, earthquake forecasts of different kinds (e.g. alarm-based forecast and probabilistic forecast). It overcomes two limitations common to other forecast comparison techniques: i) the need to define a reference model, and ii) to allow forecasts defined on different space-time-magnitude regions to be compared. We showed that the parimutuel gambling score is proper only when two forecasts are compared directly against each other. In the other cases (multi-forecast comparison and comparison against a reference model), the parimutuel gambling score is improper. Consequently, we discourage its use in multi-model comparisons such as CSEP and encourage researchers and practitioners to re-consider rankings obtained using this score.

Specifically, the parimutuel gambling score tries to avoid the need to pre-define a reference model by using the average forecast. Therefore, for each bin, a positive reward means that the model is *better* than the average forecast, vice versa if the reward is negative. This allows to produce a map of the parimutuel gambling rewards from which to infer the bins where the forecast is better than the average forecast, and the bins where it is not. Since the parimutuel gambling score is proper only when $k = 2$, any map obtained by computing the comparisons for $k > 2$ may be biased. This difficulties may be circumvented by using any proper scoring rule that allows for multi-forecast comparison. In fact, maps of this kind may be produced by reporting the score difference between a forecast and the average one. Furthermore, given that proper scores are scale invariant, we can re-scale the score values to be between -1 and 1. In this way, we can visualize which bins has a positive or negative contribution to the average score difference.

The need to compare forecasts defined on different set of bins comes from the design of the forecasting experiment. In the RELM experiment (Zechar et al , 2013) modelers were allowed to choose a subset of bins to include in their forecast, referred to as masking. Modellers involved in the RELM experiment provided forecasts with very different masks; some issued forecasts for the entire California region (Bird and Liu , 2007; Helmstetter et al , 2007; Holliday et al , 2007), some for only Southern California (Ward , 2007; Shen et al , 2007; Kagan et al , 2007), while others used irregular masks (Ebel et al , 2007). The parimutuel gambling score addressed these differences using the gambling metaphor. Each forecaster is a gambler which plays a certain number of rounds (bets) corresponding to the bins. A forecaster does not have to make a forecast for every bin - they can just sit out this round. The forecasters are ranked by their total reward (i.e. the sum of the rewards for each bin). We argue that this solution is still problematic. First, the parimutuel gambling score needs at least two forecasts for each bin to be computed. If only one forecaster *plays* in a bin we can not calculate the parimutuel gambling score for that bin. Second, consider two bins for which different sets

of forecasters provided a forecast; in this situation the models are rewarded with respect to different odds. This becomes problematic when we attempt to interpret the observed result because in each bin the reference model is given by a potentially different combination of models.

The Brier score can also be used to assess masked forecasts. The maximum Brier score value obtainable by a forecast is zero and it is achieved by the *perfect* forecast which assumes $p = 1$ when $x = 1$ and $p = 0$ when $x = 0$. Any other forecast obtains a negative Brier score. Therefore, the Brier score of a forecast can be seen as the Brier score difference between the perfect forecast and the forecast under evaluation. Given two models, the one with the highest average Brier score is the one *closest* (on average) to the perfect forecast. If the models provide forecasts on two different sets of bins, B_1 and B_2, we can still compare the forecasts in terms of their average Brier score. Suppose the first model achieves an higher average Brier score, we can conclude that, on average, the first model in B_1 is closer to the perfect forecast than the second model in B_2. We can make the above comparison also when B_1 and B_2 have zero bins in common because the Brier score requires only a forecast and the observation to be calculated.

In this paper, we have used confidence intervals to asses the statistical significance of observed score differences. Analytically determining these confidence intervals may be too complex and a basic approximative Gaussian approach may fail, as highlighted by the Brier score example. Problems of this type come from the fact that the score differences per bin are treated as independent and identically distributed. This assumption is false, especially when considering space-time bins which depend on each other both in time and space due to clustering of earthquakes. A possible solution to relax the independence assumption is to consider a Diebold-Mariano test (Diebold and Mariano , 2002) on the score differences which takes into account the correlation structure of the score differences sequence.

6 CONCLUSIONS

The parimutuel gambling score, commonly applied to compare earthquake forecasts, is improper when the number of forecasts being tested is greater than two. In the special case of two competing forecasts, the score is proper, and can return results similar to alternate proper scoring methods, but even then it can be used improperly. In the common testing scenario of multiple forecasts being compared simultaneously, or when multiple forecasts are compared against a reference model, the parimutuel gambling score provides a biased assessment of the skill of a forecast when it is tested against a given outcome. This is fundamentally a problem of the gambling analogy itself; the betting strategy of maximizing the expected reward (score) does not have to be consistent with the data generating model (in the case where this is known) and, therefore, gamblers (modellers) are not encouraged to provide forecasts resembling the data generating model. This is because the score for a given forecast is dependent

on all the forecasts taking part in the competition, not just on the observed data; one can therefore change the ranking of two models by changing one of the other models in the pool. This introduces the undesirable property that one can potentially game the system to prefer a specific model. Further, if we only have access to the forecasts and the data, it is impossible to know if the parimutuel gambling score results will be biased or not. Moreover, the only case in which they are correct is when one of the competing forecasts is the data generating model, which is highly unlikely. These findings are sufficiently clear for us to discourage the use of the parimutuel gambling score in distinguishing between multiple competing forecasts, and for care to be taken even in the case where only two are being compared.

We recommend that alternative scores that do not suffer from these shortcomings should be used instead to assess the skill of prospective earthquake forecasts in a formal testing environment. The Brier and log scores are both proper, and require no new information beyond what was used to calculate the parimutuel gambling score, so switching existing analyses to a proper score should be simple to implement. We recommend testing for properness when introducing new scoring rules, either analytically or via simulations using a known model to generate testing data.

ACKNOWLEDGMENTS

This research was supported by the European Union H2020 program (No 821115, Real-time earthquake rIsk reduction for a reSilient Europe RISE, `http://www.rise-eu.org/home/`). This research was also supported by the Southern California Earthquake Center (Contribution No. 11791). SCEC is funded by NSF Cooperative Agreement EAR-1600087 USGS Cooperative Agreement G17AC00047. All the code to produce the present results is written in the R programming language. We have used the package `ggplot2` (Wickham , 2016), the package `rnaturalearth` (South , 2017) for the Italy contour map presented in Figure 9 and the package `bayesianETAS` (Ross , 2016) for the Maximum Likelihood estimate used in Section 5. We thank Kirsty Bayliss for her useful feedback and constructive discussions which led to the final version of this work.

DATA AVAILABILITY

The code used to generate all the results presented in this article can be found at `https://github.com/Serra314/Serra314.github.io/tree/master/Ranking_earthquake_forecast`. There, it is also possible to find slides summarizing the article results and the TeX file, with all the figures, to generate the article itself.

References

Bayliss K, Naylor M, Illian J, et al (2020) Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using inlabru. Journal of Geophysical Research: Solid Earth 125(11):e2020JB020,226

Bird P, Liu Z (2007) Seismic hazard inferred from tectonics: California. Seismological Research Letters 78(1):37–48

Bourne S, Oates S, Van Elk J (2018) The exponential rise of induced seismicity with increasing stress levels in the groningen gas field and its implications for controlling seismic risk. Geophysical Journal International 213(3):1693–1700

Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly weather review 78(1):1–3

Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4):404–413

Diebold FX, Mariano RS (2002) Comparing predictive accuracy. Journal of Business & economic statistics 20(1):134–144

Ebel JE, Chambers DW, Kafka AL, et al (2007) Non-poissonian earthquake clustering and the hidden markov model as bases for earthquake forecasting in california. Seismological Research Letters 78(1):57–65

Field EH (2007) Overview of the working group for the development of regional earthquake likelihood models (relm). Seismological Research Letters 78(1):7–16

Field EH, Arrowsmith RJ, Biasi GP, et al (2014) Uniform california earthquake rupture forecast, version 3 (ucerf3)the time-independent model. Bulletin of the Seismological Society of America 104(3):1122–1180

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character 222(594-604):309–368

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102(477):359–378

Good IJ (1952) Rational decisions. Journal of the Royal Statistical Society, Ser B pp 107–114

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

Helmstetter A, Kagan YY, Jackson DD (2007) High-resolution time-independent grid-based forecast for m $\geq$ 5 earthquakes in california. Seismological Research Letters 78(1):78–86

Hernández-Orallo J, Flach P, Ferri Ramírez C (2012) A unified view of performance metrics: Trans-

lating threshold choice into expected classification loss. Journal of Machine Learning Research 13:2813–2869

Holliday JR, Nanjo KZ, Tiampo KF, et al (2005) Earthquake forecasting and its verification. arXiv preprint cond-mat/0508476

Holliday JR, Chen Cc, Tiampo KF, et al (2007) A relm earthquake forecast based on pattern informatics. Seismological Research Letters 78(1):87–93

Huber PJ (1992) Robust estimation of a location parameter. In: Breakthroughs in statistics. Springer, p 492–518

Hyvärinen A, Dayan P (2005) Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research 6(4)

Iervolino I, Chioccarelli E, Giorgio M, et al (2015) Operational (short-term) earthquake loss forecasting in italy. Bulletin of the Seismological Society of America 105(4):2286–2298

Jolliffe I, Stephenson D (2003) Forecast verification. chichester, england and hoboken

Jordan TH (2006) Earthquake predictability, brick by brick. Seismological Research Letters 77(1):3–6

Kagan YY, Jackson DD, Rong Y (2007) A testable five-year forecast of moderate and large earthquakes in southern california based on smoothed seismicity. Seismological Research Letters 78(1):94–98

Kossobokov (2004) Earthquake prediction: basics, achievements, perspectives. Acta Geodaetica et Geophysica Hungarica 39(2-3):205–221

Kossobokov (2006) Testing earthquake prediction methods: The west pacific short-term forecast of earthquakes with magnitude mwhr $\geq$ 5.8. Tectonophysics 413(1-2):25–31

Gruppo di Lavoro M (2004) Redazione della mappa di pericolosità sismica prevista dallordinanza pcm 3274 del 20 marzo 2003. Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma 5

Luen B, Stark PB, et al (2008) Testing earthquake predictions. In: Probability and Statistics: Essays in Honor of David A. Freedman. Institute of Mathematical Statistics, p 302–315

Main I (1997) Long odds on prediction. Nature 385(6611):19–20

Marzocchi W, Lombardi AM (2009) Real-time forecasting following a damaging earthquake. Geophysical Research Letters 36(21)

Marzocchi W, Zechar JD (2011) Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model. Seismological Research Letters 82(3):442–448

Marzocchi W, Zechar JD, Jordan TH (2012) Bayesian forecast evaluation and ensemble earthquake forecasting. Bulletin of the Seismological Society of America 102(6):2574–2584

Marzocchi W, Lombardi AM, Casarotti E (2014) The establishment of an operational earthquake forecasting system in italy. Seismological Research Letters 85(5):961–969

Meletti, C., Marzocchi, W., D'Amico, V., Lanzano, G., Luzi, L., Martinelli, F., Pace, B., Rovida, A., Taroni, M. & Visini, F. The new Italian seismic hazard model (MPS19). *Annals Of Geophysics*. (2021)

Michael AJ, Werner MJ (2018) Preface to the focus section on the collaboratory for the study of earthquake predictability (csep): New results and future directions. Seismological Research Letters 89(4):1226–1228

Murphy AH (1993) What is a good forecast? an essay on the nature of goodness in weather forecasting. Weather and forecasting 8(2):281–293

Rhoades D, Liukis M, Christophersen A, et al (2016) Retrospective tests of hybrid operational earthquake forecasting models for canterbury. Geophysical Journal International 204(1):440–456

Rhoades DA, Schorlemmer D, Gerstenberger MC, et al (2011) Efficient testing of earthquake forecasting models. Acta Geophysica 59(4):728–747

Rosen DB (1996) How good were those probability predictions? the expected recommendation loss (erl) scoring rule. In: Maximum Entropy and Bayesian Methods. Springer, p 401–408

Ross G (2016) Bayesian estimation of the etas model for earthquake occurrences. Preprint

Savran WH, Werner MJ, Marzocchi W, et al (2020) Pseudoprospective evaluation of ucerf3-etas forecasts during the 2019 ridgecrest sequence. Bulletin of the Seismological Society of America 110(4):1799–1817

Schervish MJ (2012) Theory of statistics. Springer Science & Business Media

Schneider M, Clements R, Rhoades D, et al (2014) Likelihood-and residual-based evaluation of medium-term earthquake forecast models for california. Geophysical Journal International 198(3):1307–1318

Schorlemmer D, Gerstenberger M (2007a) Relm testing center. Seismological Research Letters 78(1):30–36

Schorlemmer D, Gerstenberger M (2007b) Relm testing center. Seismological Research Letters 78(1):30–36

Schorlemmer D, Gerstenberger M, Wiemer S, et al (2007) Earthquake likelihood model testing. Seismological Research Letters 78(1):17–29

Schorlemmer D, Werner MJ, Marzocchi W, et al (2018) The collaboratory for the study of earthquake predictability: achievements and priorities. Seismological Research Letters 89(4):1305–1313

Shen ZK, Jackson DD, Kagan YY (2007) Implications of geodetic strain rate for future earthquakes, with a five-year forecast of m5 earthquakes in southern california. Seismological Research Letters

78(1):116–120

South A (2017) Rnaturalearth: world map data from natural earth. R package version 01 0

Stark PB (1997) Earthquake prediction: the null hypothesis. Geophysical Journal International 131(3):495–499

Steacy S, Gerstenberger M, Williams C, et al (2014) A new hybrid coulomb/statistical model for forecasting aftershock rates. Geophysical Journal International 196(2):918–923

Taroni M, Zechar J, Marzocchi W (2014) Assessing annual global m 6+ seismicity forecasts. Geophysical Journal International 196(1):422–431

Taroni, M., Marzocchi, W. & Roselli, P. Assessing alarm-based CNearthquake predictions in Italy. *Annals Of Geophysics*. **59**, S0648-S0648 (2016)

Taroni M, Marzocchi W, Schorlemmer D, et al (2018) Prospective csep evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for italy. Seismological Research Letters 89(4):1251–1261

Wallis S (2013) Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. Journal of Quantitative Linguistics 20(3):178–208

Ward SN (2007) Methods for evaluating earthquake potential and likelihood in and around california. Seismological Research Letters 78(1):121–133

Werner MJ, Sornette D (2008) Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments. Journal of Geophysical Research: Solid Earth 113(B8)

Werner MJ, Helmstetter A, Jackson DD, et al (2010) Adaptively smoothed seismicity earthquake forecasts for italy. arXiv preprint arXiv:10034374

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, `https://ggplot2.tidyverse.org`

Zechar JD, Jordan TH (2008) Testing alarm-based earthquake predictions. Geophysical Journal International 172(2):715–724

Zechar JD, Jordan TH (2010) The area skill score statistic for evaluating earthquake predictability experiments. In: Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II. Springer, p 39–52

Zechar JD, Zhuang J (2010) Risk and return: evaluating reverse tracing of precursors earthquake predictions. Geophysical Journal International 182(3):1319–1326

Zechar JD, Zhuang J (2014) A parimutuel gambling perspective to compare probabilistic seismicity forecasts. Geophysical Journal International 199(1):60–68

Zechar JD, Gerstenberger MC, Rhoades DA (2010a) Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts. Bulletin of the Seismological Society of America 100(3):1184–1195

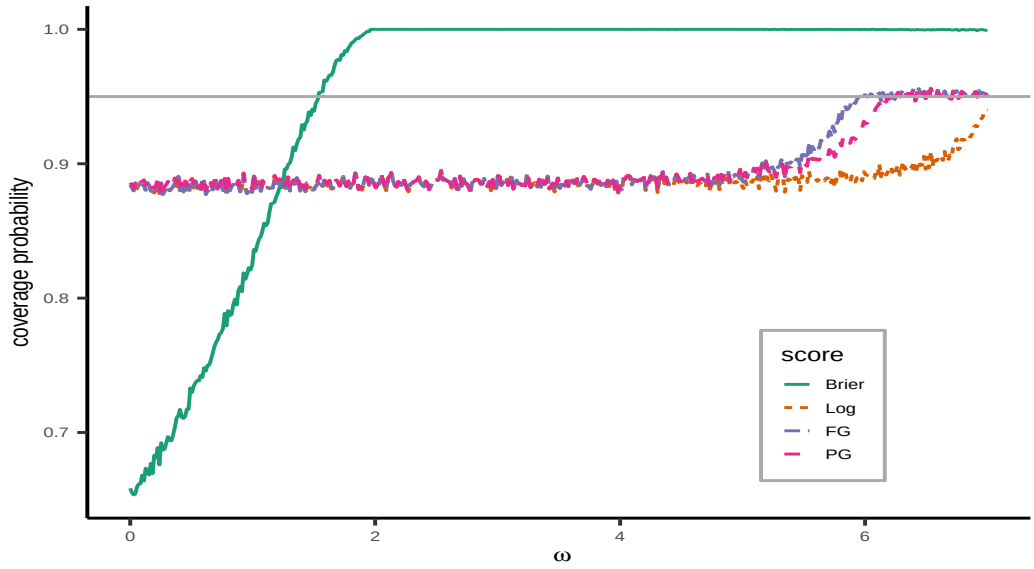

Figure A1. Coverage probability for the Gaussian approximated confidence interval for the expected score difference between $\mathbf{p}_1$ and $\mathbf{p}_2$ as a function of $\omega = \mathbf{p}_2/\mathbf{p}_1$ for $\omega \in (10^{-3}, 7)$. We consider the Brier, log, pairwise gambling (PG) and full gambling (FG) scores. We set $\mathbf{p}^*$ equal to the 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$ and $\mathbf{p}_0 = 5\mathbf{p}^*$. Horizontal line represents the target coverage probability (0.95)

Zechar JD, Schorlemmer D, Liukis M, et al (2010b) The collaboratory for the study of earthquake predictability perspective on computational earthquake science. Concurrency and Computation: Practice and Experience 22(12):1836–1847

Zechar JD, Schorlemmer D, Werner MJ, et al (2013) Regional earthquake likelihood models i: First-order results. Bulletin of the Seismological Society of America 103(2A):787–798

Zhuang J (2010) Gambling scores for earthquake predictions and forecasts. Geophysical Journal International 181(1):382–390

APPENDIX A: RELIABILITY OF THE GAUSSIAN CONFIDENCE INTERVALS

We assess the reliability of the Gaussian approximation to calculate confidence intervals for the expected score differences using simulated replicates from $\mathbf{p}^*$ equal to the 5-year adaptively-smoothed Italy forecast. We set the first forecasts $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and the reference model $\mathbf{p}_0 = 5\mathbf{p}^*$. The method is reliable if the probability that the approximated confidence interval of level α contains the true expected value (coverage probability) is close to $1 - \alpha$. To estimate the coverage probability, for a set of values of $\omega \in (10^{-3}, 7)$, we simulated $\mathbf{x} = x_1, ..., x_N$ 10,000 times and calculated the approximate 95% confidence intervals. The coverage probability is given by the fraction of times in which the confidence intervals contains the true expected score difference.

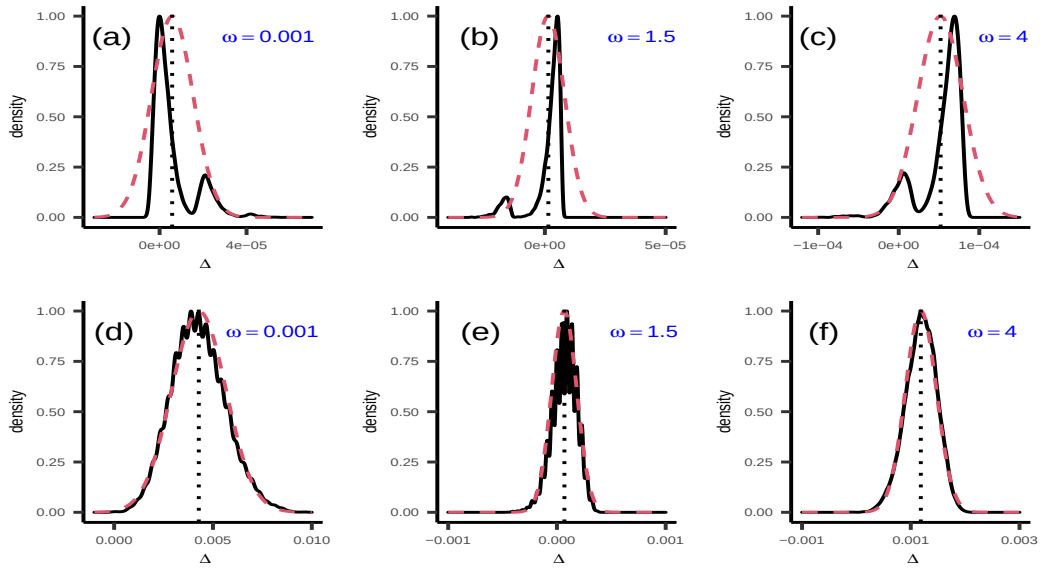

Figure A2. Average score difference distribution for the Brier score (a-c) and log score (d-f). We set $\mathbf{p}^*$ equal to the 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$. We consider $\omega = 0.001$ (a,d); $\omega = 1.5$ (b,e); $\omega = 4$ (c,f). Black solid line represents the empirical distribution obtained from 10000 simulations. Red dashed line represents the corresponding Gaussian approximation. Vertical dotted line represents the true value of the expected score difference.

Figure A1 shows that the approximation is not reliable for the Brier score for which produces confidence intervals which are too small (coverage below 0.95) or too wide (coverage above 0.95) depending on the value of ω. This is because the Brier score differences have a asymmetrical bimodal distribution as shown in Figure A2a-c and therefore are not normally distributed. Furthermore, Figure A1 shows also that the coverage for the log, pairwise gambling and full gambling score is usually below 0.95 and reaches this value only for $\omega > 6$. This is because, taking the log score as example, the distribution of the score differences becomes smoother as ω grows (Figure A2d-f). The distributions of the pairwise and full gambling score resemble the log score one.

From this analysis we conclude that it is possible to use the gaussianly approximated confidence intervals for the log, pairwise gambling and full gambling because the coverage probability is always between 0.88 and 0.96. In this example, the reliability of the approximation depends on how much different the forecasts are.