Ranking earthquake forecasts: On the use of proper scoring rules to discriminate forecasts

Francesco Serafini¹, Mark Naylor¹, Finn Lindgren¹, and Maximilian Werner²
¹University of Edinburgh, School of Geosciences, United Kingdom of Great Britain – England, Scotland, Wales (s1898281@ed.ac.uk)
²University of Bristol, School of Earth Sciences, United Kingdom of Great Britain

Recent years have seen a growth in the diversity of probabilistic earthquake forecasts as well as the advent of them being applied operationally. The growth of their use demands a deeper look at our ability to rank their performance within a transparent and unified framework. Programs such as the Collaboratory Study for Earthquake Predictability (CSEP) have been at the forefront of this effort. Scores are quantitative measures of how well a dataset can be explained by a candidate forecast and allow forecasts to be ranked. A positively oriented score is said to be proper when, on average, the highest score is achieved by the closest model to the data generating one. Different meanings of closest lead to different proper scoring rules. Here, we prove that the Parimutuel Gambling score, used to evaluate the results of the 2009 Italy CSEP experiment, is generally not proper, and even for the special case where it is proper, it can still be used improperly. We show in detail the possible consequences of using this score for forecast evaluation. Moreover, we show that other well-established scores can be applied to existing studies to calculate new rankings with no requirement for extra information. We extend the analysis to show how much data are required, in principle, to distinguish candidate forecasts and therefore how likely it is to express a preference towards a forecast. This introduces the possibility of survey design with regard to the duration and spatial discretisation of earthquake forecasts. Our findings may contribute to more rigorous statements about the ability to distinguish between the predictive skills of candidate forecasts in addition to simple rankings.