Skill Scores for Probabilistic Climate and Weather Prediction

T. Maynard, E B Suckling, and L A Smith

London School of Economics, Centre for the Analysis of Time Series, United Kingdom (e.suckling@lse.ac.uk)

A wide variety of skill scores are in use for the evaluation of probability forecasts. While the importance of using proper scores is well recognised, researchers often face requests to present results under a variety of different scores. Is there any sense in which considering many "different" skill scores makes a case more (or less) persuasive? Which set of scores makes the most persuasive case?

A number of scores are considered, and their strengths and weaknesses contrasted, including the most commonly used metrics such as the proper linear score, continous ranked probability score and logartithmic score, amongst others. The role of moving from ensembles of point forecasts to continuous probability distributions is discussed. Illustrations with meteorological forecasts are considered, as are mathematical experiments where the size of the forecast-outcome archive can be made arbitrarily large. The aim is to restrict the number of skill scores considered, reduce the use of misleading scores, identify independent evidence and ensure that the set of scores published are specified before the results are in.

I J Good's logarithmic score is known to be the only local, proper score for continuous variables. Proposed criticisms of locality are considered and rejected and it is argued that proper, non-local scores tend to include a term that is independent of the probability mass associated with the outcome. Thus they reward the forecast for some other property of the distribution function itself, having for example a particular shape; it is conjectured that this may be a general result. This is depricated, except in cases where the forecaster is being rewarded for something other than the relation between the forecast and reality, or whose distribution functions are unavoidably constrained by some particular structure. Remaining challenges to the evaluation of probability forecasts are noted. In partucular the effect of small samples and the implications of uncertinaty in the outcome.