



Verification of probabilistic forecasts: comparing proper scoring rules

Thordis L. Thorarinsdottir and Nina Schuhen
Norwegian Computing Center, Oslo, Norway

In ensemble weather forecasting, verification methods are required to diagnose both the need for statistical post-processing, and the effectiveness of the post-processing methods in producing calibrated and accurate forecasts. Scoring rules assess the accuracy of probabilistic forecasts by assigning a numerical penalty to each instance of a forecast and it is imperative for them to be proper in order to prevent hedging strategies and to identify the most skillful forecast model.

A plethora of proper scoring rules, assessing different aspects of the predictive distribution, has been proposed in the literature. We will demonstrate a coherent approach to using proper scores, as well as rank or PIT histograms in practice, while highlighting some of the difficulties that might arise when working with limited data sets. In particular, we show how different scoring rules rank forecasts according to their skill, and how the ranking may differ with the amount of available data, suggesting that forecast ranks based on limited data sets should be interpreted with care.