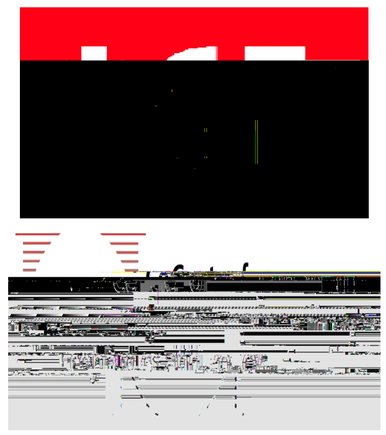

Robust Measure of Predictive Skill and Ensemble Design

Hailiang Du[1], Falk Niehoerster[1] and Leonard A. Smith[1,2]

[1] Centre for the Analysis of Time Series, London School of Economics,
[2] Pembroke College, Oxford
Email: *lenny@maths.ox.ac.uk*, *h.l.du@lse.ac.uk*

Abstract

This poster addresses issues in the interpreting of probability forecasts based on multi-model ensemble simulations. Probabilistic skill in ENSEMBLES seasonal forecasts for Nino 3.4 is demonstrated. True cross-validation is shown to be important given the small sample size available in seasonal forecasting. The sources of apparent (RMS) skill in distributions based on multi-model simulations is discussed, and it is demonstrated that the inclusion of "zero-skill" models in the long range can improve RMS scores. This casts some doubt on one common justification for the claim that all models should be included in forming an operational PDF. RMS "skill" is shown to be misleading. Results using a proper skill score show the multi-model ensembles do not significantly outperform a single model ensemble for Nino 3.4.

Evaluating ENSEMBLES with a Proper Score

The performance of forecast distributions can be evaluated with the "log p score" (Ignorance Score [2]), defined by:

$$\mathbf{S}(\mathbf{p}(\mathbf{y}), \mathbf{Y}) = -\mathbf{log}(\mathbf{p}(\mathbf{Y})), \qquad (1)$$

where $\mathbf{Y}$ is the verification and $\mathbf{p}$ is forecast probabilistic density function. Ignorance is the only proper local score for continuous variables [1,3]. In practice, given K forecast-outcome pairs $(\mathbf{p_t}, \mathbf{Y_t}, \mathbf{t} = 1, ..., \mathbf{K})$, the empirical average Ignorance skill score is: