Skill of Ensemble Seasonal Probabilistic Forecast

Hailiang Du[1], Falk Niehoerster[1], Roman Binter[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

The skill of probability forecasts of the temperature at Nino 3.4 based upon the ENSEMBLES seasonal simulations is considered and contrasted with those of the DEMETER simulations. This poster addresses the problem of interpreting probability forecasts based on these multi-model ensemble simulations; the distributions considered are formed by kernel dressing the ensemble and blending with the climatology. 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, casting some doubt on the common justification for the claim that all models should be included in forming an operational PDF. It is argued that the rational response varies with lead time

1 From Simulation to a PDF

A given ensemble of simulations is translated into a probability distribution function by a combination of kernel dressing and blending with climatology [4]. Given an N member ensemble at time t, $X_t = [x_t^1, ..., x_t^N]$, and treating ensemble members under the same model as exchangeable, kernel dressing defines the model-based component of the density as:

$$p(y : X, \sigma) = \frac{1}{N\sigma} \sum_i^N K\left(\frac{y - x^i - u}{\sigma}\right), \quad (1)$$

where K is a kernel. Here we take

$$K(\zeta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\zeta^2), \quad (2)$$

where y is a random variable corresponding to the density function p. In this case each ensemble member contributes a Gaussian kernel centred at $x^i + u$, where u is an offset accounting for systematical bias. The kernel width, σ, is simply the standard deviation of the Gaussian kernel.

For any finite ensemble, the verification may lie far from the ensemble members even if the verification is selected from the same distribution as the ensemble itself. Blending the most relevant climatological distribution of the system with the model-based distribution yields a probability forecast usually superior to that obtained without blending. The eventual forecast distribution is then:

$$p(\cdot) = \alpha p_m(\cdot) + (1 - \alpha) p_c(\cdot) \quad (3)$$

where p_m is the density function generated by dressing the ensemble and p_c is the estimate of climatological density.

To produce the forecast distribution requires estimation of the kernel width σ the shifting parameter u and the weight α assigned to the model. We fit these three parameters simultaneously by optimising the Ignorance score, introduced below, by leave one out cross validation [1].

2 Contrasting ENSEMBLES & DEMETER

The performance of forecast distributions is evaluated primarily using the "log p score" (Ignorance Score [2]). The Ignorance Score is defined by:

$$S(p(y), Y) = -log(p(Y)), \quad (4)$$

where Y is the verification. Ignorance is the only proper local score for continuous variables [1,3]. In practice, given K forecast-verification pairs $(p_t, Y_t, t = 1, ..., K)$, the empirical average Ignorance skill score is:

$$S_{Emp}(p(y), Y) = \frac{1}{K} \sum_{i=1}^{K} -log(p_i(Y_i)) \quad (5)$$

We evaluate the ENSEMBLES & DEMETER seasonal models [5] by their empirical Ignorance score. From Fig 1. In general, both IFS(ECMWF) model and HadGem2(UKMO) model tend to outperform other models in the ENSEMBLES project. ECHAM5(INGV) model seems doing very well in the DEMETER project. By looking at the relative Ignorance between ENSMEBLE and DEMETER model outputs, it seems except the ECHAM5(INGV) model, all other three models have made improvement in terms of Ignorance.

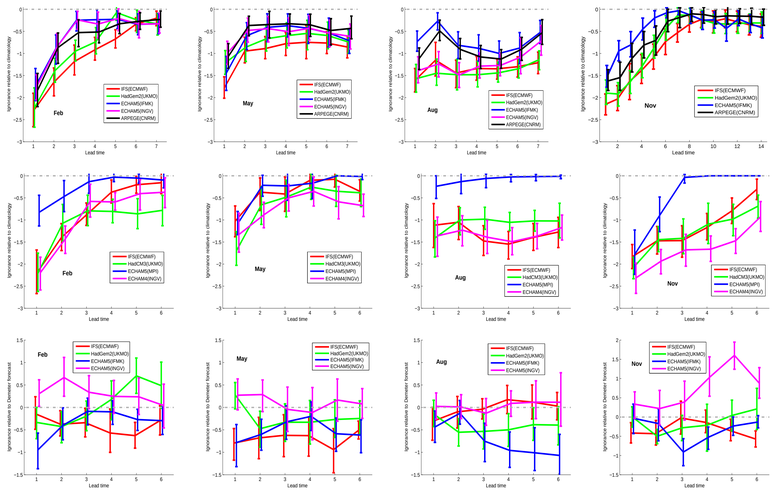

Fig 1: Ignorance score of each model forecast of SST in the nino3.4 region as a function of lead time. The uncertain bars are the 90 percent bootstrap re-sampling bounds, calculated from 512 bootstrap re-samples. Figures in the first row represent Ignorance of each model from ENSEMBLES project relative to climatology, each picture corresponding each launch date; the second row represents Ignorance of each model from DEMETER project relative to climatology; the third row shows the Ignorance score of ENSEMBLES forecast relative to corresponding DEMETER model forecast (ECHAM5(IFMK) is compared with ECHAM5(MPI)).

3 The meaning of the (ensemble) mean and value of large ensembles

It is often said that the ensemble mean outperforms the best model. The right panel in Fig. 2 shows that at large lead times merely decreasing the variance of the IFS(ECMWF) forecast improves the RMS skill. In this case including zero skill forecasts (with zero mean error) would appear to improve the score! While at short lead times (where the ensemble has more significant skill) decreasing the variance increases the RMS error. This casts doubt on the utility of RMS error measures. The left panel in Fig 2. suggests that multi-model ensemble really does contribute skill.

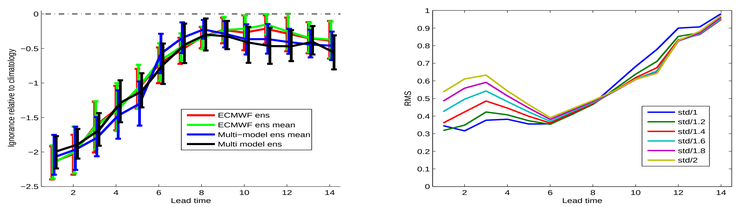

Fig 2: a) Ignorance score of i) IFS(ECMWF) ensemble ii) IFS(ECMWF) ensemble mean iii) Multi-model (Including the four models in Fig 1) ensemble mean iv) Multi-model ensemble, forecast for Nov launch, relative to climatology. b) RMS error for the forecast using IFS(ECMWF) ensemble mean with their variance shrunk.

The seasonal ensembles within the ENSEMBLES project each consist of nine members, decreasing the ensemble degrades forecast skill, as shown in Fig. 3 where skill of two member (red) and four member (green) ensembles are shown relative to the full nine member ensemble.

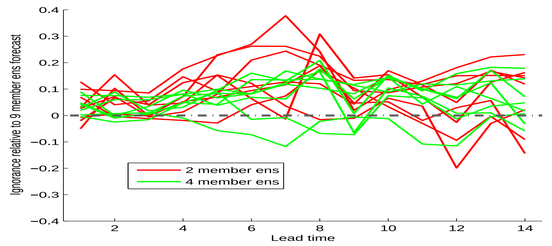

Fig 3: Ignorance score of IFS(ECMWF) model forecast of SST in the nino3.4 region as a function of lead time. The green lines represents the Ignorance of 4 member (random drawn from the original 9 member ensemble) ensemble forecasts relative to 9 member ensemble forecast; the red lines 2 member ensemble forecasts

4 Constructing PDFs from multiple models

Each model provides a distribution of simulations: how do we best combine them without over-fitting given that we have only 50 independent launches? Fig. 4 illustrates that that such combinations will be lead-time dependent. At shorter lead times, where the better models have significantly more skill, combining only one or two of the best models does well, while including all models does poorly in months one to eight, and then arguably outperforms the other combination in months ten through fourteen.

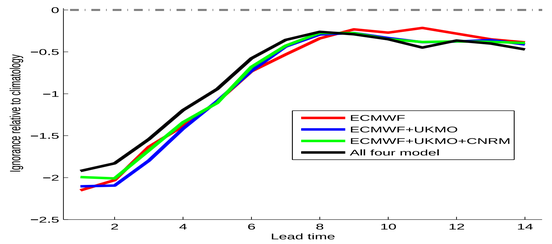

Fig 4: Ignorance score of multi-model forecasts of Nino3.4 SST as a function of lead time. Multi-model forecasts are constructed by assigning equal weights on each model forecast PDF.

Summary

The current generation of seasonal forecasts will retire before the forecast-verification archive gets significantly larger: seasonal verification data is precious. ENSEMBLES-based PDFs have skill at 14 months lead-time, a skill significant improvement on the DEMETER models. Different skill scores can obscure real skill from proper scoring rule from mere statistical effects reflected in RMS scores. The evidence of skill at long lead-times is of nontrivial value in various applications, and distinguishing the limitations of this skill for decision making from the limitations of our current skill scores will prove of great value.

References

[1] J. M. Bernardo. Expected information as expected utility. Annals of Statistics, 7(7):686C690, (1979).

[2] G.W. Brier, Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1C3, (1950).

[3] J. Brocker, L.A Smith, Scoring Probabilistic Forecasts: On the Importance of Being Proper, *Weather and Forecasting*, 22 (2), 382-388, (2006).

[4] J. Brocker and L.A. Smith, From ensemble forecasts to predictive distribution functions, *Tellus A*, 60, 663-678 (2007).

[5] A. Weisheimer, F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Deque, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel, ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictionsSkill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, 36, L21711 (2009).

[1]As only 42 years data are provided, the estimation of these two parameters is lack of robustness. If one has 4000 years data, one can draw multiple 42 years data set from them and estimate the parameters for each sample set. The variation of the estimates is large.