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:

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

where Y is the verification and 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 $(p_t, Y_t, t = 1, ..., K)$, the empirical average Ignorance skill score is:

$$S_{Emp}(p(y), Y) = \frac{1}{K}\sum_{i=1}^{K} -log(p_i(Y_i)) \qquad (2)$$

We evaluate the ENSEMBLES models [5] by their empirical Ignorance score. A bootstrap resampling procedure which samples the forecast Ignorance score with replacement is used to reflect the uncertainty in the empirical Ignorance. In general, all models (Fig. 1) are substantially more skillful than climatology in predicting short lead times for all initialization dates. Although less information is provided in the longer lead time, they still add significant information to the climatology up to a lead time of 14 months. On average, the IFS(ECMWF) and HadGem2(UKMO) models score best.

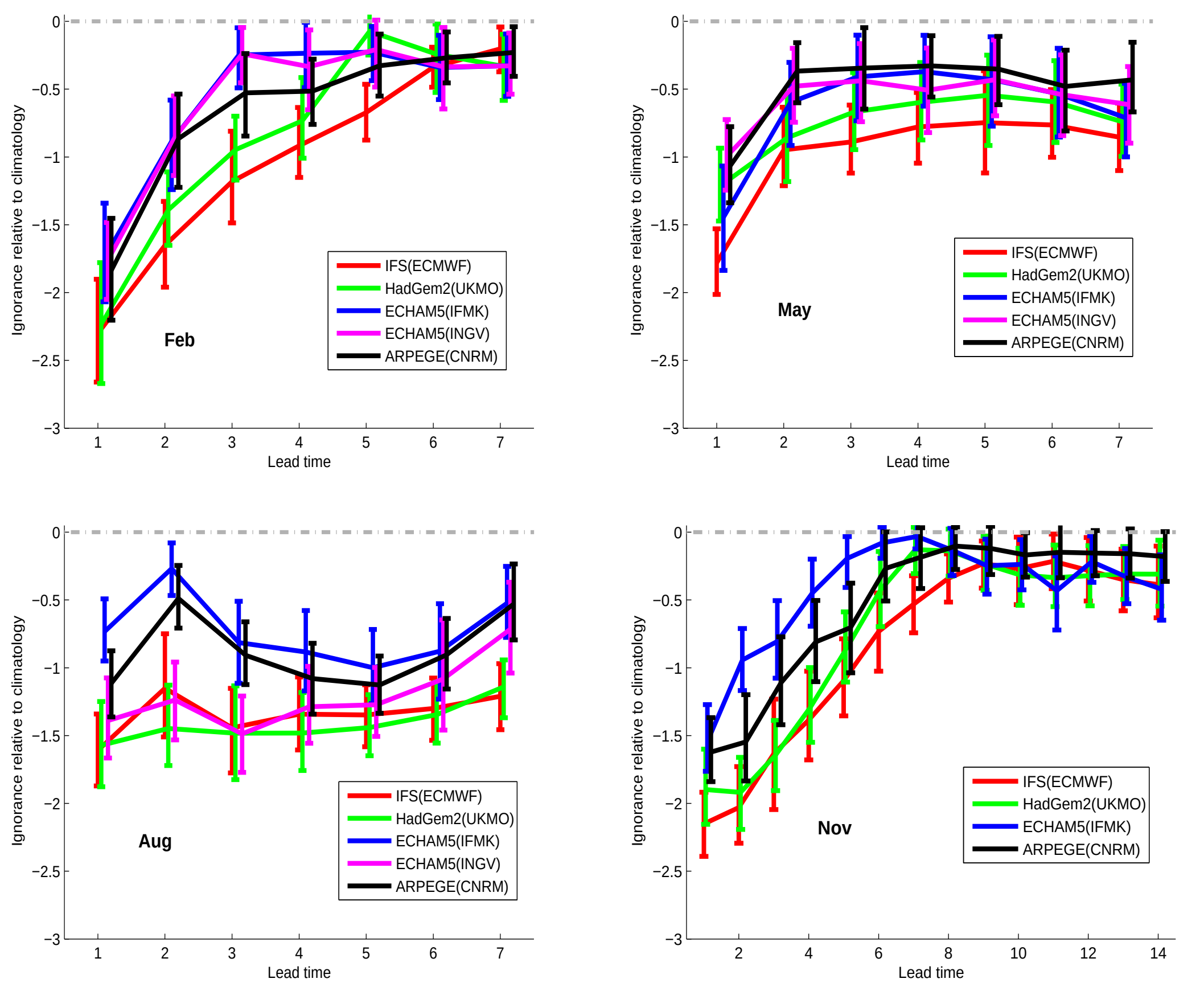

Fig 1: Ignorance score of each model forecast of SST in the Nino3.4 region as a function of lead time. The uncertain bars are the 90 percent bootstrap re-sampling bounds, calculated from 512 bootstrap re-samples. Ignorance of each model from ENSEMBLES project relative to (monthly) climatology is represented, each picture corresponds to a different launch date.

The meaning of the (ensemble) mean and value of multi-model ensembles

It is often suggested that the multi-model mean is more skillful than the best ensemble forecast [6]. This statement requires a careful examination of the definition of skill. The RMS error of the ensemble mean, for example, can be a very misleading measure of forecast skill. This is demonstrated in Fig. 2 where we compare the IFS(ECMWF) ensemble mean November forecast with forecasts in which we simply reduce the variance of the ensemble mean forecasts over all forecast years for each lead time; this can be seen as adding "an ensemble" of random numbers with zero mean to each forecast. Fig. 2a shows including such zero skill forecasts would improve the RMS score! At short lead times (where the ensemble has significant skill) decreasing the forecast variance increases the RMS error. The Ignorance score reveals that the actual forecast skill (the probability on the outcome) is not improved by including a zero skill forecast (see Fig. 2b). The problem here is with RMS error measures.

Fig. 2c shows the standard deviation of the ECWMF and multi-model ensemble mean as a function of lead time. The variance of the multi-model ensemble actually becomes smaller than that of ECMWF ensemble after a lead time of 8 months. Simply this fact, in the same fashion as above, could show increased "RMS skill" of the multi-model mean, even if the other models provide no additional skill (information regarding the outcome).

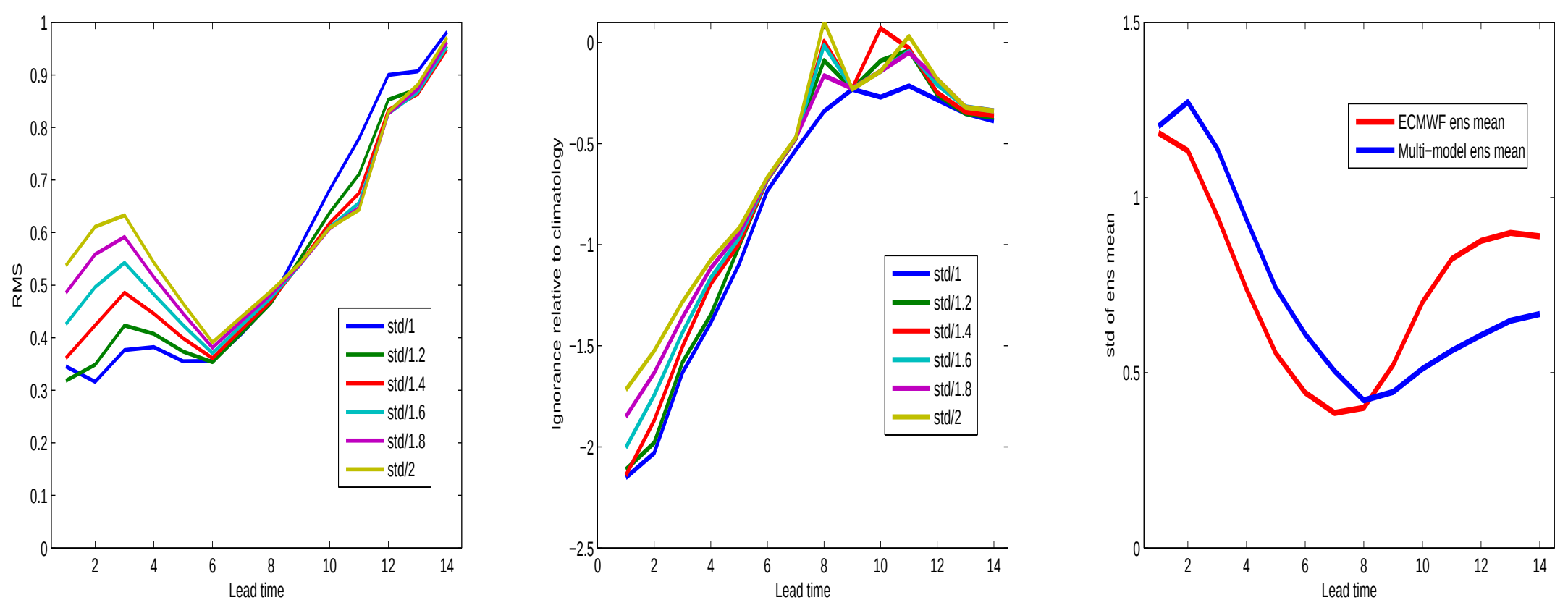

Fig 2: a) RMS error for the forecast using IFS(ECMWF) ensemble mean with reduced variance. b) Ignorance score for forecast using IFS(ECMWF) ensemble mean with reduced variance. c) standard deviation of Multi-model and IFS(ECMWF) ensemble mean as a function of lead time

Does one gain more information from increasing the number of ensemble members from a good model, or, combining the forecasts of different models to a multi-model ensemble? Fig. 3 shows the Ignorance score of multi-model forecasts with various ensemble sizes relative to the full 9-member ECWMF forecast (zero line). The full multi-model ensemble (with all 36 members from simulations of the IFS(ECMWF), HadGem2(UKMO), ECHAM5(IFMK) and ARPEGE(CNRM) models, blue line in Fig. 3 outperforms the 9-member ECMWF ensemble in longer lead times while under-performs in short lead times. To compare ensemble of the same size, multi-model ensembles with smaller number of ensemble members are generated by random draws (without replacement) from all 36 simulations. The 9-member ECMWF forecast outperforms 4-member multi-model ensembles significantly in short lead time and often at longer lead times (Fig. 3a). Comparing like with like 9-member ensembles, the single-model ECMWF forecast outperforms the multi-model forecast most of the time (Fig. 3b). We see no evidence that multi-model ensembles significantly outperform a single model ensemble of the same size.

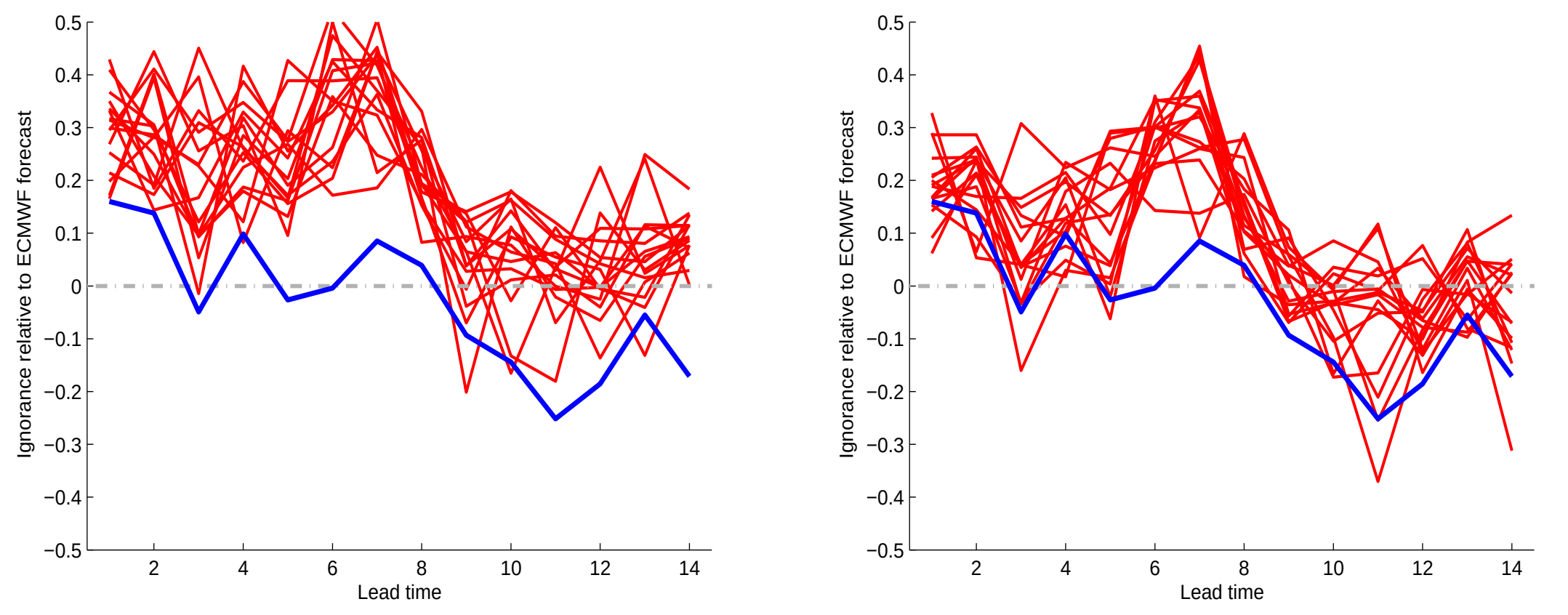

Fig 3: Ignorance of multi-model forecasts of various ensemble sizes relative to the 9-member ECMWF November forecast for the Nino3.4 index. The blue line represents the multi-model forecast using all 36 ensemble members (including ECMWF). The red lines are multi-model forecasts using random combinations of 4-members (left) or 9-members (right) from the full ensemble. The dashed line of zero represents the 9-member ECMWF forecast.

True cross-validation when data is precious

When the forecast-outcome library is small, misleading expectations of out-of-sample performance can arise due to "informative contamination". Maintaining true cross-validation is critically important. The results presented above adopt the leave one out cross-validation (described below) to fit the offset u, the kernel width σ and the weight assigned to the model α. Forecast details in the appendix are given.

In short, define the forecast probability distribution to be $p(\cdot, X_t, \Theta)t = 1, ..., N$, where X represents the ensemble forecast at time t, vector Θ contains u,σ and α and N is the number of forecasts. Given the corresponding outcome s_t, for each forecast at time j, we leave out (X_j, s_j) and using the rest of the data to fit the parameter Θ by minimizing the empirical score. Let the fitted value to be $\hat{\Theta}_j$. We use the median of those fitted values (noted $\tilde{\Theta}$) to compute the forecast empirical Ignorance, i.e. $\sum_{j=1}^{N} -log_2 p(s_j, X_j, \tilde{\Theta})$.

The leave one out procedure described in the previous paragraph does not provide true cross-validation, as $\tilde{\Theta}$ is not completely independent of (X_j, s_j). To achieve true cross-validation, one can adopt the procedure described as follows. After firstly leaving out (X_j, s_j), for the remaining set one apply the leave one out procedure again to obtain the fitted parameter values $\tilde{\Theta}_j$ for the archive that does not contain (X_j, s_j). Now $\tilde{\Theta}_j$ is independent with (X_j, s_j).

Fig. 4 shows the relative Ignorance using true cross-validation; the bootstrap resampling bars tend to be wider. Arguably there is no statistical significant skill after lead time 8. Fig. 4b quantifies the apparent "loss of skill", which is in fact false skill, and may lead to over confidence. Finding an ideal cross-validation procedure requires further investigation. Using true cross-validation indicates using less informative model as one put less data to build the model. In practice one will use all the previous years forecasts to forecast next year, while using all the previous forecasts in-sample yields over confidence.

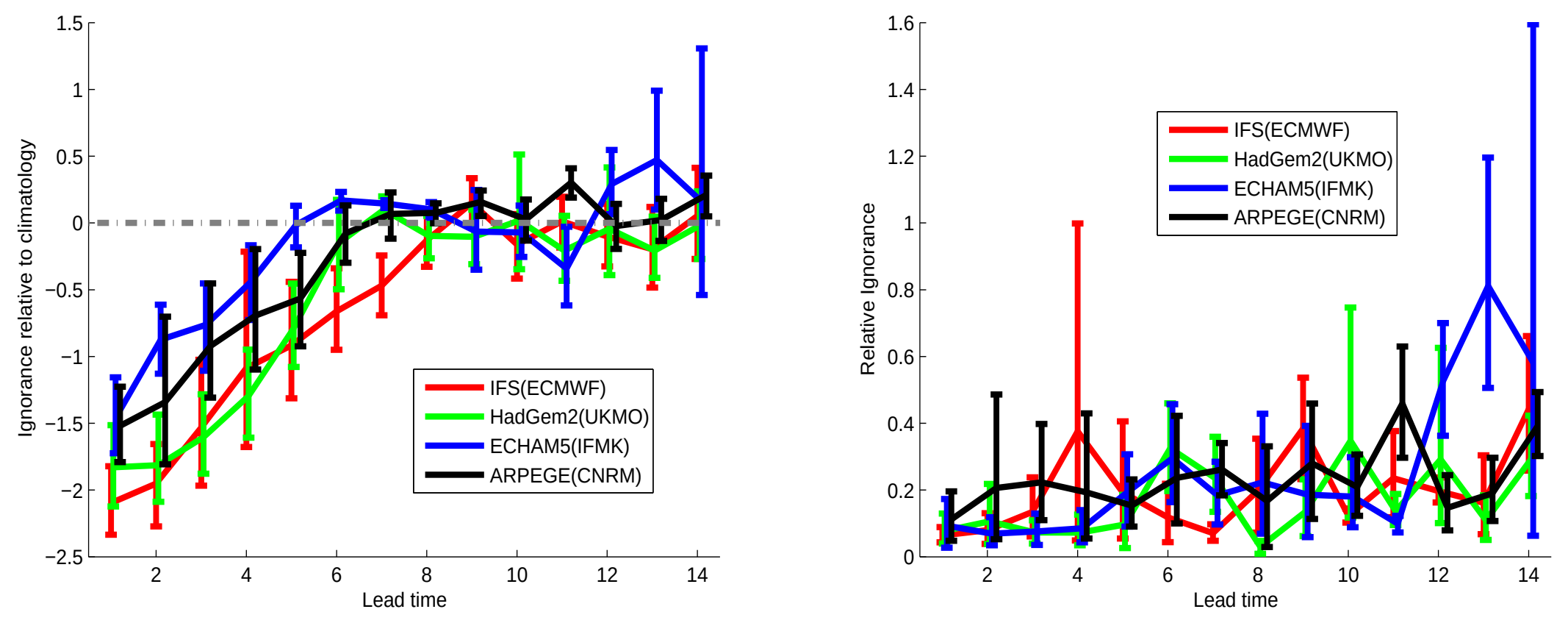

Fig 4: a)Ignorance score of ECMWF ensemble forecasts relative to climatology using truly leave one out cross-validation, b) Ignorance score of ECMWF ensemble forecasts, true leave-one-out cross-validation relative to leave one out cross-validation.

Summary

The current generation of seasonal forecasts will retire before the forecast-outcome archive grows significantly: seasonal forecast-outcome data is precious. ENSEMBLES-based PDFs have probabilistic skill at long lead times. Using RMS as a measure of skill can obscure true skill with mere statistical effects. 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 may prove of value.

Appendix: 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:

$$p(y : X, \sigma) = \frac{1}{N\sigma}\sum_{i}^{N} K\left(\frac{y - x^i - u}{\sigma}\right), \qquad (3)$$

where K is a kernel. Here we take

$$K(\zeta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\zeta^2), \qquad (4)$$

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:

$$p(\cdot) = \alpha p_m(\cdot) + (1 - \alpha)p_c(\cdot) \qquad (5)$$

where p_m is the density function generated by dressing the ensemble and p_c is the estimate of climatological density. The parameter α reflects the contribution of the model to the forecast.

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 (discussed in the poster).

Acknowledgements

The support of the Economic and Social Research Council (ESRC) and of Munich Re is gratefully acknowledged.

References

[1] J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(7):686-690, (1979).
[2] D.J. Good, Rational decisions. *Journal of the Royal Statistical Society*, XIV(1) 107 (1952).
[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, et al. 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).
[6] R. Hagedorn, F.J. Doblas-Reyes and T.N. Palmer, The rationale behind the success of multi-model ensembles in seasonal forecasting. *Tellus A*, 57 (2005).
[7] L.A. Smith, H. Du and F. Niehorster, Skill of Ensemble Seasonal Probabilistic Forecast, working paper (2011).