Increasing Foresight and Forecast Quality with Skillful Low-Cost Empirical Models

Hailiang Du[1], Leonard A. Smith[1,2], Emma Suckling[3] and Erica Thompson[2]

[1] RDCEP, University of Chicago, [2] London School of Economics and Political Science, [3] University of Reading

Email: *hdu@uchicago.edu*, *lenny@maths.ox.ac.uk*

Abstract

Simulation models are widely employed to make probability forecasts on seasonal to annual time-scales and increasingly on decadal scales. While simulation models based on physical principles are often expected, in principle, to outperform purely empirical models, that claim must be established empirically for any given generation of models; direct comparison of the forecast skill of simulation models and empirical models provides information on progress toward that goal which is not available in model-model intercomparisons. More importantly, the blending of forecasts from both sources can lead to better operational forecasts. Direct comparison can also reveal the space and time scales on which simulation models exploit their physical basis effectively, perhaps indicating the origins of their weaknesses. The skill of seasonal and decadal probabilistic hindcasts for global and regional mean temperatures from the ENSEMBLES projects are interpreted in this context. Physically inspired empirical models are shown to display probabilistic skill comparable to that of today's state-of-the-art simulation models as well as to that of the multi-model ensemble. The inclusion of empirical models (blending) with simulation models is shown to significantly improve forecasts. Inasmuch as the cost of building or running empirical models is negligible comparing to large simulation models, it is suggested that the direct comparison of simulation models with empirical models become a regular component of large model forecast evaluations, that rank order evaluations include empirical models whenever the timescales allow, and that blending simulation models with empirical models becomes a regular component of seasonal and decadal forecasting.

Seasonal Forecasts

The ENSEMBLES multi-model ensemble experiment for seasonal forecasting comprises five global coupled atmosphere-ocean climate models (Further details of the ENSEMBLES experiments can be found in [1]). Hindcast simulations considered here were launched on the first day of February, May, August and November each year over the 46 year period from 1960 to 2005 for the Nino3.4 region (each model consists of a nine-member initial condition ensemble). Probability forecasts are generated from the ENSEMBLES simulations via kernel dressing and are blended with climatology to produce seasonal probability forecasts (for a full description see Appendix I). Figure 1. shows an example of the kernel dressed and blended probabilistic forecast distributions for a subset (over the period 1995-2000) of the IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 region, launched in November.

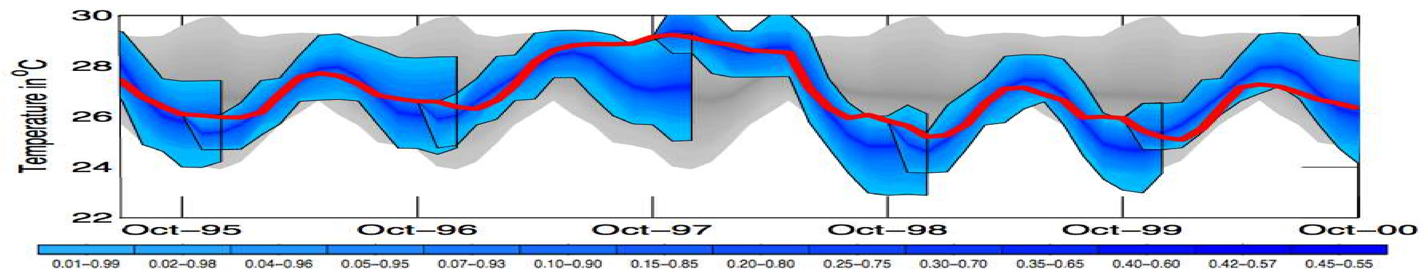

Fig 1: Probabilistic forecast distributions for the IFS(ECMWF) hindcast simulations from ENSEMBLES for the Nino3.4 index, launched in November over the period 1995-2000. The blue shaded regions indicate the forecast percentiles between 1-99% and the red line shows the observed outcome from the ERA40 reanalysis. The grey shaded intervals show the percentiles for the climatological distribution.

Figure 2. shows the empirical Ignorance skill (Defined in the Appendix II) for forecasts of the Nino3.4 index as a function of lead time, in months, relative to climatology. In general at short lead times all the models are substantially more skillful than climatology (that is a negative relative Ignorance) for all four initialization dates. At longer lead times ENSEMBLES models show systematically less skill than at early lead times, as expected. The sampling uncertainty across forecast launches is represented by a bootstrap resampling procedure, which resamples the set of forecast Ignorance scores for each model, with replacement.

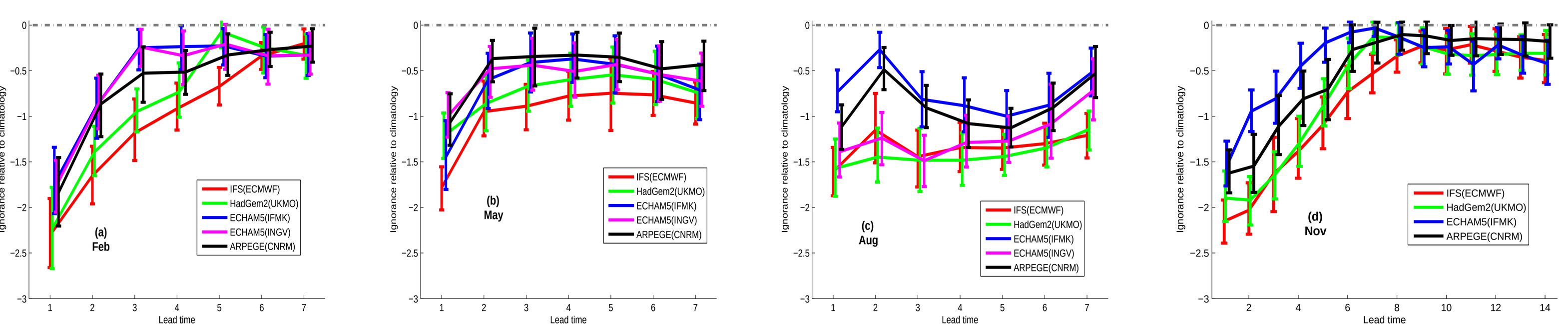

Fig 2: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to climatology as a function of lead time in months. Zero Ignorance indicates a model has no skill relative to climatology. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

Whilst comparing skill between simulations from dynamical models and climatology provides insight into the information gained from forecasting with those dynamical models, other simple empirical models can also serve as appropriate benchmarks to model performance [5,6]. A probabilistic persistence forecast provides an interesting benchmark accounting for the effects both of physical persistence and of any long term drift in the temperature of the target region. The persistence forecasts generated here use the observed SST value over the chosen region in the month prior to the forecast launch, persisted forward in time, and transformed into a probabilistic distribution using kernel dressing parameters that vary with lead time (as described in [5]).

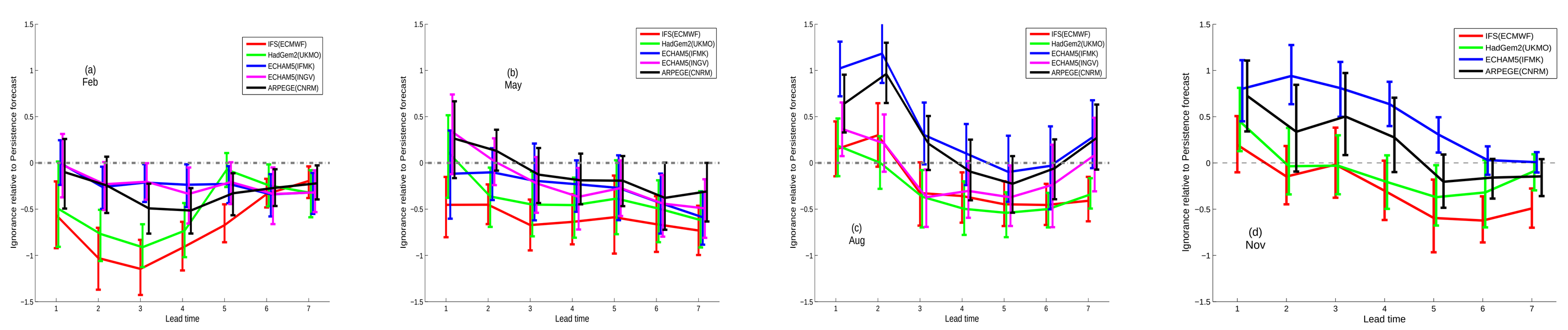

Fig 3: Ignorance score of each model from ENSEMBLES for the Nino3.4 index relative to persistence forecasts as a function of lead time in months. Bootstrap resampling intervals (the vertical bars) reflect the 5% to 95% range as estimated from 512 resamples.

Figure 3. shows the Ignorance score of each of the ENSEMBLES models for the Nino3.4 index relative to persistence. For forecasts launched in February most of the ENSEMBLES models are significantly more skillful than persistence at all lead times. For launch dates in August and November little if any information is added compared to the persistence forecasts for most models at any lead time. In fact at early lead times (up to three months ahead) persistence outperforms the ECHAM5(IFMK) and ARPEGR(CNRM) models. At moderate lead times for the August launch and most lead times in the May launch, on the other hand, the IFS(ECMWF) and HadGEM2(UKMO) models outperform persistence.

Decadal Forecasts

A set of decadal simulations from the Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES) experiment (Hewitt and Griggs 2004; Doblas-Reyes et al. 2010), a precursor to phase 5 of the Coupled Model Intercomparison Project (CMIP5) decadal simulations (Taylor et al. 2009), is considered. For the ENSEMBLES models, each simulation ensemble consisted of only three members launched at 5 years intervals from 1960 to 2005. Figure 4. illustrates the 2 years running mean of simulated global-mean temperature from the four simulation models in the multimodel ensemble experiment of the ENSEMBLES project over the full set of decadal hindcasts. Observations from the Hadley Centre/Climatic Research Unit, version 3 (HadCRUT3) dataset and the 40 years European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) are shown for comparison; HadCRUT3 is used as the verification dataset outcome archive for both the model evaluation and construction of the empirical model. Even at the global scale, the raw simulation output is seen to differ from the observations both in terms of absolute values, as well as in dynamics.

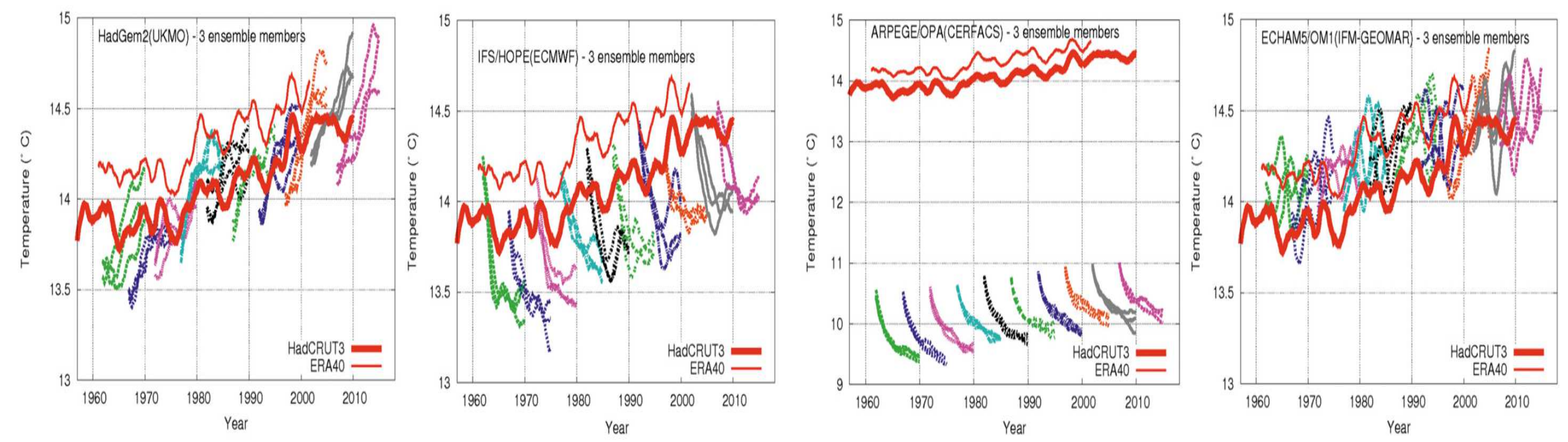

Fig 4: Global-mean temperature (2 years running mean) for the four forecast systems. HadCRUT3 observations and ERA-40 are also shown for comparison. Note that the scale on the vertical axis for the ARPEGE4OPA model is different than for the other three models, reflecting the larger bias in this model.

Empirical models based on historical observations cannot be expected to capture previously unobserved dynamics. Two empirical models typically used in forecast evaluation are the climatological distribution and the persistence model. Alternative empirical models for probability forecasts, more appropriate for a changing climate, define a dynamic climatology based on ensemble random analog prediction (eRAP) (Smith 1997; Paparella et al. 1997). The methodology of the dynamic climatology (DC) is described in Appendix III. Figure 5a. shows the performance of each of the ENSEMBLES models and the DC model relative to forecasts of persistence. The DC model consistently shows relative ignorance scores below zero across most lead times, while the ARPEGE4/OPAmodel scores below zero for early lead times (up to a lead time of 5 years), suggesting that forecasts from these models are more skillful than a persistence forecast over this range. The skill of the ENSEMBLES simulation model forecasts is illustrated relative to the DC model in Figure 5b. None of the models in the ENSEMBLES multimodel ensemble demonstrates significant skill above the DC model at any lead time for global-mean temperature. In fact, all four simulation models show systematically less skill than the DC model.

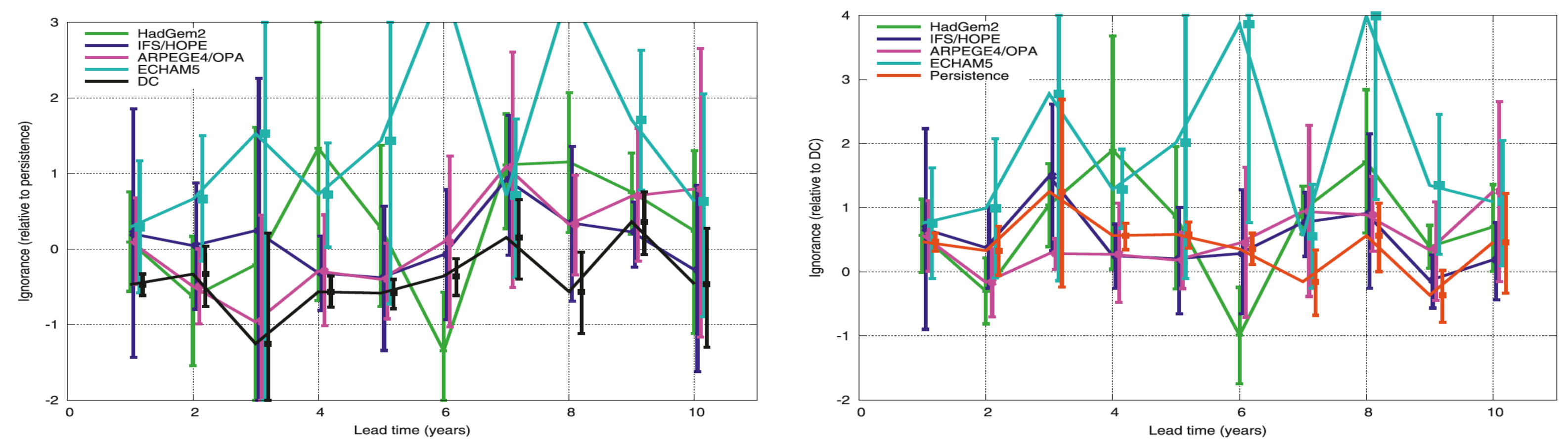

Fig 5: a) Ignorance of the ENSEMBLES models and DC relative to persistence forecasts as a function of lead time; b) Ignorance of the ENSEMBLES models relative to DC as a function of lead time. The bootstrap resampling intervals are illustrated at the 10th-90th percentile level.

Appendix I: 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:

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

where K is a kernel. Here take

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

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:

$$p(\cdot) = \alpha p_m(\cdot) + (1 - \alpha)p_c(\cdot) \tag{3}$$

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 (see [1] for a full description).

Appendix II: Evaluate probabilistic forecasts

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

$$S(p(y), Y) = -log(p(Y)), \tag{4}$$

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

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

Appendix III: Dynamic Climatology

The dynamic climatology approach used here considers the $l - step$ differences in the observational record (e.g., a 1-step difference might be the temperature difference between the current state and its immediately preceding state). A distribution is formed for each value of l from the corresponding differences using all the observations after some start date; thus, the size of the ensemble decreases linearly with lead time because of the finite size of the archive. For a forecast of a scalar quantity, such as the global-mean temperature below, the DC ensemble at lead time l launched at time t consists of the set of N_l values,

$$e_i = S_t +^l \Delta_i, i = 1, ..., N_l, \tag{6}$$

where S_t is the initial condition at time t and $^l\Delta_i$ $i = 1, ..., N_l$ is the set of lth differences in the observational record.

Conclusions

Probabilistic seasonal forecasts based on the ENSEMBLES stream II experiment clearly outperform climatological probability forecasts in many cases. The fact that empirical persistence-based probability forecasts provide a significantly stronger challenge suggests that, in practice, the skill of operational forecast systems can be enhanced with information from the richer empirical models. The quality of decadal probability forecasts from the ENSEMBLES simulation models has been compared with that of reference forecasts from several empirical models. In general, the stream 2 ENSEMBLES simulation models demonstrate less skill than the empirical DC model across the range of lead times from 1 to 10 years. The DC probability forecasts often place up to 4 bits more information (or 24 times more probability mass) on the observed outcome than the ENSEMBLES simulation models. In the context of climate services, the comparable skill of simulation models and empirical models suggests that the empirical models will be of value for blending with simulation model ensembles. It also calls into question the extent to which current simulation models successfully capture the physics required for realistic simulation of the Earth system and can thereby be expected to provide robust, reliable predictions (and, of course, to outperform empirical models) on longer time scales. The comparison of near-term climate probability forecasts from Earth simulation models with those from dynamic climatology empirical models provides a useful benchmark as the simulation models improve in the future. The blending (Brocker and Smith 2008) of simulation models and empirical models is likely to provide more skillful probability forecasts in climate services, for both policy and adaptation decisions.

Acknowledgment

This research was supported by the LSE's Grantham Research Institute on Climate Change and the Environment and the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council and Munich Re. Additional support for H.D. was also provided by the National Science Foundation Award No. 0951576 "DMUU: Center for Robust Decision Making on Climate and Energy Policy (RDCEP)". L.A. S. gratefully acknowledges the continuing support of Pembroke College, Oxford.

References

[1] Smith, L.A., Du, H., Suckling, E.B. and Niehoerster, F. ?Probabilistic skill in ensemble seasonal forecasts?, Quarterly Journal of the Royal Meteorological Society (in press).
[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 (2008).
[5] Suckling, E.B. and Smith, L.A. (2013) 'An evaluation of decadal probability forecasts from state-of-the-art climate models', Journal of Climate, 26 (23): 9334-9347.
[6] Smith, L. A., 1992: Identification and prediction of low-dimensional dynamics. Physica D, 58 (1?4), 50?76.
[7] Smith, L. A., 1997: The maintenance of uncertainty. Proc. 133rd Int. School of Physics "Enrico Fermi" Course, Varenna, Italy, Societa Italiana di Fisica, 177?246.
[8] Paparella, F., A. Provenzale, L. A. Smith, C. Taricco, and R. Vio, 1997: Local random analogue prediction of nonlinear processes. Phys. Lett., 235A, 233?240.
[9] Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2009: A summary of the CMIP5 experimental design. CMIP5 Rep., 33 pp.
[10] Hewitt, C. D., and D. J. Griggs, 2004: Ensembles-based predictions of climate and their impacts. Eos, Trans. Amer. Geophys. Union, 85, 566.
[11] Doblas-Reyes, F. J., A. Weisheimer, T. N. Palmer, J. M. Murphy, and D. Smith, 2010: Forecast quality assessment of the ensembles seasonal-to-decadal stream 2 hindcasts. ECMWF Tech. Memo. 621, 45 pp.