All models are wrong: Which are worth paying to look at?

A case study for Global Mean Temperature

Emma Suckling and Leonard A. Smith

Centre for the Analysis of Time Series, Department of Statistics, London School of Economics, UK

http://www2.lse.ac.uk/CATS

Introduction

Dynamical simulation models (GCMs), often used to provide decision support in the context of climate variability and change, typically have complex structures, rendering them computationally intensive to run and expensive to develop. In extrapolation the models which 'capture the physics' must justify their cost to users by demonstrating that they outperform simpler statistical models by placing significantly more probability mass on the verification. But do today's 'best available' models do so?

An approach is presented towards a robust measure of the in-sample skill of ensemble forecasts and the performance of a set of decadal simulations from ENSEMBLES for global mean temperature is assessed against a benchmark statistical model based on the random analogue prediction method. The ensemble forecasts are expressed as probability distributions through the kernel dressing procedure and their quality quantified according to the Ignorance skill score.

Can we determine where simulations win?

The performance of decadal predictions over global mean temperature is initially considered since in the absence of second order effects, simulation models are expected to perform better over larger spatial and temporal scales. It is then essential to understand how the performance (as well as the value in terms of providing decision support) of complex models will change against simple data-based models moving from global annual averages to local daily extremes.

Figure 1 illustrates the simulated global mean temperature (as a 2 year running mean) for the HadGEM2 model, containing 3 initial condition ensemble members, over the full set of decadal predictions from ENSEMBLES, which are initialised to observations for a series of November launch dates [1]. HadCRUT3 observations and ERA40 reanalysis data have been treated in an identical manner and are also shown.

Even at global scales, the raw model forecasts are seen to differ somewhat from the target observations.

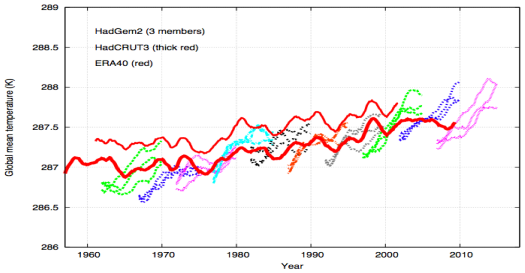

Figure 1: Global mean temperature (2 year running mean applied) for the HadGEM2 model of the ENSEMBLES decadal forecasts. HadCRUT3 observations and ERA40 reanalysis are included.

Figure 2 shows the forecast distributions for a subset of the simulations. While the pattern of temperature change is captured well over some individual forecasts, in several instances the verification falls within the tail of the distribution even after a complicated bias correction procedure is applied, which is based on the mean forecasts error as a function of lead time (figure 3).

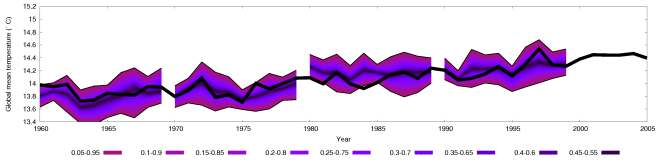

Figure 2: Predictive distributions (percentile ranges as indicated) of global mean temperature for the HadGEM2 model from the ENSEMBLES decadal simulations for launch dates at 10 year intervals.

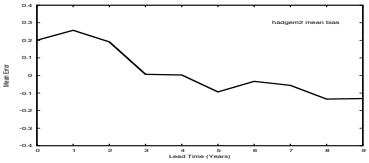

Figure 3: Mean forecast error for global mean temperature as a function of lead time (T(obs)-T(model)).

References

[1] A. Troccoli and T. N. Palmer, Phil. Trans. R. Soc. A **365**, 2179-2191 (2007).

[2] L. A. Smith, *The Maintenance of Uncertainty*, Proc. International School of Physics "Enrico Fermi", Course CXXXIII, pg 177-246 (1997).

[3] J. Bröcker and L. A. Smith, Weather and Forecasting, **22** (2), 382-388 (2007).

[4] H. Du, F. Niehoerster and L. A. Smith, *Improvement in Full Probability Forecasting at Seasonal Lead-times*, submitted.

Random Analogue Prediction model

The random analogue prediction model (RAP) [2] provides a simple reference for comparison against the performance of complex simulation models since it contains few model structure assumptions but is expected to be more skilful than climatology. A set of forecasts are produced for the RAP model, initialised to the observations.

Figure 4 illustrates this approach, in which an ensemble is built from available analogue states over the full time series using the direct method (so that a forecast for lead time, n, is produced by considering the full set of nth differences, leaving out the forecast year itself).

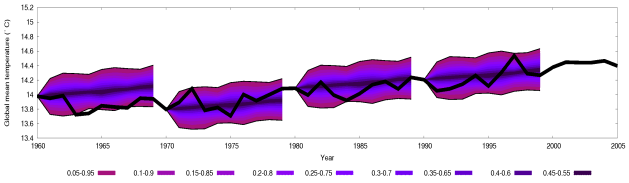

Figure 4: Predictive distributions for global mean temperature using the direct RAP model.

Ensemble forecast evaluation

Figure 5 shows a comparison between the performance as a function of lead time for the HadGEM2 and RAP models. A measure of the information contained in each ensemble of model forecasts is quantified by transforming the forecasts into a continuous probability distribution function through the kernel dressing procedure [3]. Gaussian functions are applied to each ensemble member with optimised kernel mean and spread as a function of lead time that are obtained by minimising a cost function based on the Ignorance skill score, defined as $I=-\log_2(p(x))$, where $p(x)$ is the probability assigned to the verification, x. The mean is taken over a set of forecast-verification pairs using a leave-one-out cross validation methodology.

RAP performs to a similar quality as HadGEM2 over some lead times, although a small sample size of forecast-verification pairs in the ENSEMBLES simulations leads to large uncertainties. The blue line illustrates a true leave-one-out methodology, pointing to the importance of careful cross validation in decadal forecasting [4].

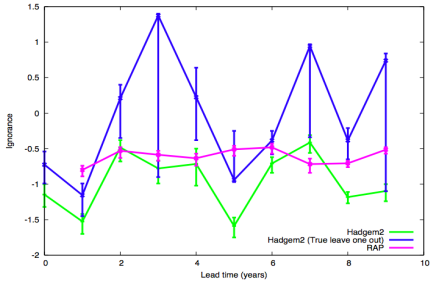

Figure 5: Ignorance as a function of lead time for the HadGEM2 and RAP models. The uncertainty bars are the 70th percentile from re-sampling from the forecast set. Lower Ignorance values indicate better skill.

While statistical models such as RAP are not expected to capture changes due to previously unobserved physical feedbacks, current simulation models may or may not capture such feedbacks. An important question therefore arises as to how a model may be deemed fit for purpose in decision support tasks in the longer range. At decadal scales, direct comparisons can be made as illustrated here.

Outlook

The performance of decadal forecasts from ENSEMBLES has been contrasted with a simple benchmark model. It is found that at some lead times simulations outperform RAP (~1 bit), placing twice the probability mass on the verification. Small sample sizes, however, lead to large uncertainties in terms of computing skill.

Given the practical limitations in terms of producing ensemble forecasts, establishing statistical significance for any model requires clear experimental design, including prior specification of which predictions will be evaluated.

Under EQUIP, this methodology will be used to determine the user-value derived by employing GCMs in addition to statistical forecasts.

Acknowledgements

The support of the Economic and Social Research Council (ESRC), as well as NERC under the EQUIP project is gratefully acknowledged.