Blending Ensembles from Multiple Models

Sarah Higgins[1] and Leonard A. Smith[1,2]

[1] Centre for the Analysis of Time Series, Dept. of Statistics, London School of Economics, U.K.
[2] Oxford Centre for Industrial and Applied Mathematics, University of Oxford

Introduction

The most accurate seasonal weather forecasts combines multiple models developed by different countries using equal weights. A methodology to blend each model's forecasts using weights determined by the skill of each model is examined. As there is only a small forecast-outcome archive available for seasonal forecasts we look at combining multiple imperfect models from a non-linear system using a proper skill score to determine the weights.

DEMETER

The DEMETER data set is a multi-model ensemble of seasonal forecasts for atmospheric variables which is modeled in 10^6 dimensions. Each hindcast has been integrated for six months and has nine members in its forecast ensemble. The dataset contains 22 seasonal forecasts.

Moran Ricker

The low dimensional chaotic system used in this experiment is the Moran Ricker Map:

$$x_{i+1} = x_i e^{\alpha(1-x_i)} \tag{1}$$

Where α is set to 2.9. The experiment uses three imperfect models of the Moran Ricker Map and a climatology model. The climatology model was generated from points iterated thousands of times through the Moran Ricker Map to ensure they lay on the model's attractor.

Large Data Sets

The initial conditions for the large data set were drawn from the Moran Ricker Map's attractor. To mirror the DEMETER forecasts the three imperfect models (1000, 0100 and 0010) each had nine initial conditions. The outcomes were from the perfect model, the Moran Ricker Map. For each model the nine point forecast was converted into a probability distribution function by fitting a Gaussian kernel on top of each member of the ensemble. The first half of the large dataset (1,000 points) was used to calculate the model parameters. The forecasts from all three models were blended with equal weights using every possible model combination. For time step five the skill of each model combination was calculated using a proper skill score, ignorance, using the second half of the dataset. The relative ignorance was then calculated by subtracting the climatological skill score from the blended model skill score. If the relative ignorance was less than zero then the forecast had more skill than climatology. With ignorance the lower the value, the more skill the model has. Fig 1 shows box plots of the relative ignorance of all model combinations.

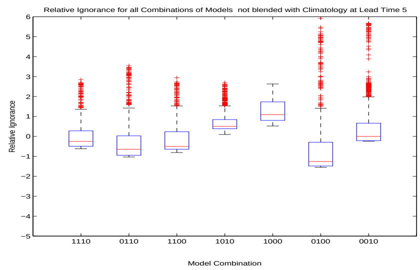

Fig 1: Relative ignorance at time step 5 for all combinations of equally weighted imperfect models using a large dataset

The model combination with the most skill is not all three models combined (1110) but a single model 0100. The worst model forecast is again a single model 1000 with the highest relative ignorance score. The high values of ignorance on the boxplots mark forecast busts where the verification fell some distance outside the forecast distribution.

Limited Data Sets

As the DEMETER data set had only 22 points the experiment was repeated limiting the forecast-outcome archive to this number but using the same climatology. The model parameters were fitted by minimizing ignorance using a leave one out methodology where the parameters were fitted across 21 points and validated against the 22^{nd} point. The model's forecasts were again blended using equal weights for every possible combination of models.

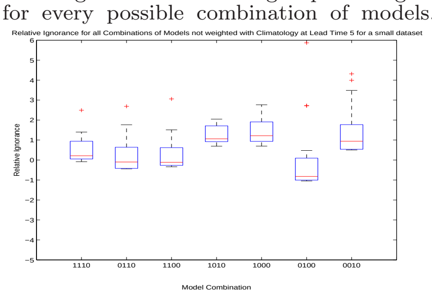

Fig 2: Relative Ignorance at time step 5 for all combinations of equally weighted imperfect models using a 22 point dataset

Fig 2 shows that even with a significantly smaller forecast-outcome archive the skill of each model combination can still be roughly estimated as shown by the resemblance between Fig 1 and Fig 2.

Combining with Climatology

Boxplots of the 22 relative ignorance scores for each combination of equally weighted models and climatology are shown in Fig 3. Most of the relative ignorance scores are now lower than Fig 2 showing that blending with climatology has increased the skill.

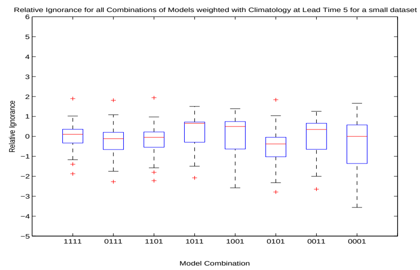

Fig 3: Relative Ignorance at time step 5 for all combinations of equally weighted imperfect models and climatology with a 22 point dataset

DEMETER Dataset

The DEMETER models were blended with climatology with the weight set by minimizing ignorance. August forecasts for the sea surface temperature at Nino 3.4 were used.

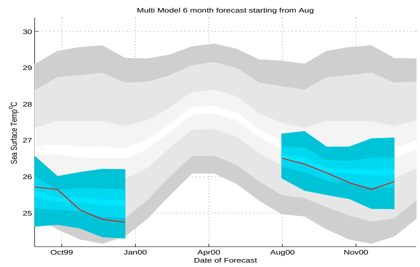

Fig 4: Six monthly August forecasts using DEMETER multi models blended with climatology for SST for 1999 and 2000

In Fig 4 grey shows the pdf of climatology, blue is the seasonal forecast pdf of all the DEMETER models blended with climatology and the red line is the actual observations. Blending the models with climatology produces a forecast that is more accurate than climatology alone as the blue pdf has a smaller spread than climatology yet still captures the red observations.

Conclusions

The Moran Ricker experiment shows using equally weighted multi models does not always provide the best forecast and that the skill of individual models should be taken in to account. Blending the Moran Ricker models with climatology normally increases the skill of the forecast. Further work needs to be done to estimate at what size limited datasets can provide an accurate assessment of large datasets.

References

Hagedorn, R., Doblas-Reyes, F.J. and Palmer T.N.(2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept Tellus A 57(3):219

Brocker, Jochen and Smith, Leonard A. (2008) From ensemble forecasts to predictive distribution functions Tellus A 60(4) p 663-678