4. Verification of 2 metre temperature (2mt) ensemble seasonal forecasts

Introduction: DEMETER (Development of a European Multimodel Ensemble system for seasonal to inTERannual prediction) is an EU initiative to develop a well-validated European multi-model ensemble forecast system for reliable seasonal to inter-annual prediction. It ran from 2000-2003. (SEE Palmer et al, BAMS, 85, 853-872 for further details of the project, and also on the ECMWF.int website). This project offers the opportunity to compare the quality of three 9-member ensembles for three-month average temperature forecasts.

Data files: t2m-ecmwf-JJA-1959-2001.txt, t2m-mf-JJA-1959-2001.txt and t2m-ukmo-JJA-1959-2001.txt. These files contain 1-month lead 2 metre temperature (2mt) forecasts produced in May and valid for JJA for the period 1959-2001 (i.e. 43-years of forecasts) for a grid point in central Pacific (latitude 0°, longitude 140°W) for three coupled seasonal forecast models which were included in the DEMETER project (ECMWF, Meteo-France, and UK Met Office).

The structure of the data files is as follows: (There is no header row)
1st column: year
2nd column: "observed" 2mt in JJA (from ERA-40 reanalysis)
3rd to 11th column: forecast JJA 2mt (ensemble member 1 to 9)

Goal: To compare and contrast the forecasts from the three models.

Analysis suggestions: These are ensemble forecasts (9 ensemble members for each forecast), so can be used to estimate the probability of departure from the normal (climatological) mean temperature or the probabilities for tercile categories (below normal, normal, above normal) for the three month period. The climatological mean can be defined by the average of the observations for the 43 years. The distribution of the observations can then be used to define quantiles e,g below normal, normal and above normal with thresholds at the 33rd, and 67th percentiles of the observed distribution.

It is customary to account for the bias in the models when carrying out the verification. This is done by using the whole ensemble of forecasts (9 members for all 43 years, total 387 temperature values) to compute tercile thresholds of temperature. This is done separately for each of the 3 models. Then probabilities of below normal, normal and above normal can be determined for each ensemble for each of the 43 years by counting the number of ensemble members falling in each tercile category and dividing by the total number of ensemble members (9), using the respective model tercile thresholds, and verified according to the methods of probability verification for discrete probability distributions. (RPS, RPSS, etc). The verification can also be done by considering two- two-category predictions, for P(upper tercile) vs (not upper tercile) and P(lower tercile) vs (not lower tercile). In this case the BS and BSS are the appropriate scores. (Note that the RPS and the BS are equal for two categories)

The forecasts can also be verified using contingency tables, either a 3 by 3 or two- two-category tables with thresholds at 33 and 67%, highest probability category is chosen as the predicted category.

Suggested steps: 1. Plot the ensemble mean vs the observation for each model. This offers a quick assessment of bias.

2. Compute the temperature tercile threshold values for the observations (from the 43-member distribution of observations) and for each of the models. (from the 387 member grand ensemble of forecasts)

3. Verify using the methods applicable to probability forecasts, including the ROC (These are 9-member ensembles so the ROC analysis should have 10 bins representing the 10 possible probability values – bins at 0 and 1 are half-width and the other 8 bins are centered on 1/9, 2/9….8/9). Calculate the RPS and RPSS with respect to climatology and/or evaluate the forecast probabilities for above normal and below normal categories using 2 by 2 contingency tables and associated scores. In either case, compare results over the three models. (With only 43 cases, reliability tables aren't feasible, but an estimate of reliability can be obtained from the factorization of the RPS)

4. Prepare a presentation of the results, identifying the "best" model, indicating why.

Questions to answer: a. If data for a mid-latitude point were analysed, how do you think the results might differ and why?

b. For this dataset, the "observations" come from a global analysis. What effect might that have on the comparative results, and why?

Extra analysis: Verify the ensemble means with respect to the observation using verification metrics for continuous variables. Are these results consistent with the ensemble results?