Calibration and validation of seasonal forecasts

Laura Ferranti

ECMWF, Shinfield Park, Reading, UK
laura.ferranti@ecmwf.int

1 Introduction

Producing a seasonal forecast from a coupled model does not solely involve running the model and viewing the results. After running a coupled model, regardless of its complexity, corrections need to be made for systematic errors; due to the model's climate and the observed climate being different. Typically; simple procedures, applied a posteriori, are used to correct these systematic errors. The following section discusses different types of systematic errors and methods to correct them. The seasonal forecast information is not complete if we have no idea of the forecast skill. Computing and providing skill measurements is another crucial step for the use and further development of seasonal forecast. The last section describes the verification procedures and discusses some outstanding issues that affect the skill assessment. Typically, the systematic error estimates and skill measurements are based on a re-forecast data set. A re-forecast data set consists of a collection of forecasts with start and valid dates from the past, usually going back for a considerable number of years. In order to ensure consistency between re-forecasts and real-time forecasts, re-forecasts are produced specifically with the same model system that is used to produce the real-time forecasts. The re-forecast data set is an integral part of the real-time seasonal predictions. In fact, without this data set, it would be difficult to construct a calibrated forecast and any measurement of its skill. The re-forecasts for the current ECMWF operational forecasting system (S4) start on the 1st of every month for the years 1981-2010. The ensemble size is 15 members. The data from these forecasts is available to users of the real-time forecast data, to allow them to calibrate their own real-time forecast products. Currently ECMWF is running additional re-forecast ensemble members for a sub-set of dates, to allow a better sampled characterization of skill. This is particularly important for regions and times when the forcing signal is low – a large ensemble size is needed to avoid spurious "signals" due to inadequate sampling.

2 Calibration and systematic model errors

With systematic model errors we intend any difference between the observed and model climatology. The simplest form of systematic error is the mean bias. Figure.1 shows the mean bias in Sea Surface Temperature (SST) for the summer months (June to August). All runs used here are started the first of May and they covered the 30 years period between1981-2010. Among other signals, the plot shows that the ECMWF model climatology is too cold over the Equatorial Pacific. The panel on the

bottom shows the SST bias for the autumn months estimated by the same simulations. The mean state of the model at a further forecast range, shows that the cooling over the Equatorial Pacific tends to amplify. Figure2 shows the mean biases for SST from 4 different coupled systems that contribute to the Eurosip multi-model system. It is interesting to note that:

- mean biases are a feature of any model
- in some regions, the biases are common to different models. For example the cold tongue over the equatorial Pacific is a common feature to system a) and b).

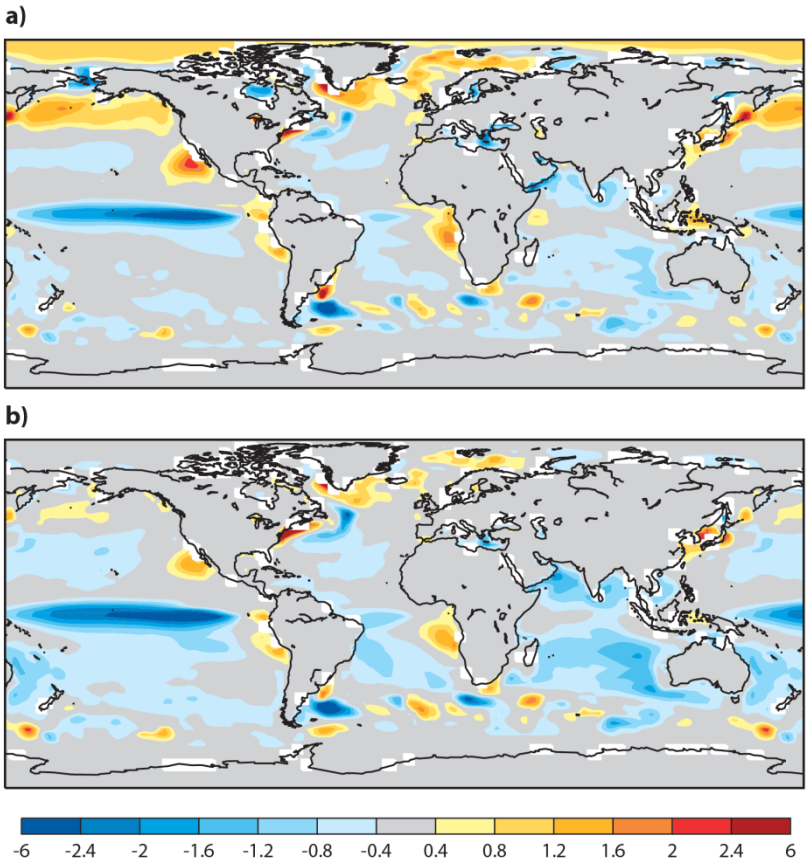

Figure 1: SST biases for forecast initiated in May valid for June to August. The biases are computed using the re-forecast over 30 years period (1981-2010).

The model can be systematically different from the observed climate in terms of its variance. Figure3 shows SST anomalies averaged over the Central Eq. Pacific (Niño3.4 area) for the period 1981 to 2010 valid for June to August season. The blue dots represent the ensemble means and the red dots are the observed values. The green boxes indicate the ensemble forecast distribution for each year. It is easy to note that while the inter-annual variations are well captured (anomaly correlation is 0.88) the variance of the simulated SST anomalies is larger than the observed variance. For this particular season, the ratio between the model inter-annual variance and the observed one is in fact 1.30. The over estimation of the SST amplitudes in the Equatorial Pacific is seasonally varying and it reaches its maximum for a lead time of two months, verifying the boreal spring. In order to correct for some of these systematic model

errors, the model output is adjusted a posteriori. This procedure is called calibration. It is important to note that this "a posteriori" correction does not remove the effect of having an incorrect mean state, since the evolution of the anomalies in a climate system is not a linear process.

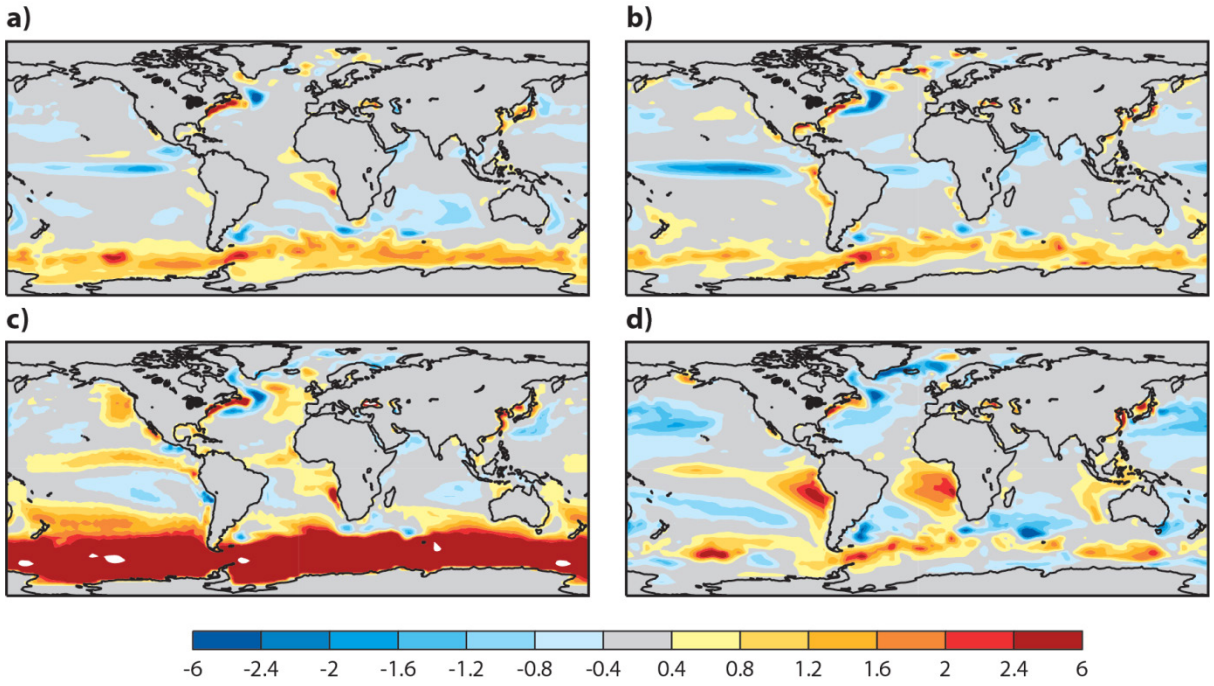

Figure 2: SST biases from the 4 coupled seasonal forecast models contributing to the Eurosip multi-model system. Biases are valid for the season December to February and are computed for the 14 years period 1996-2010.

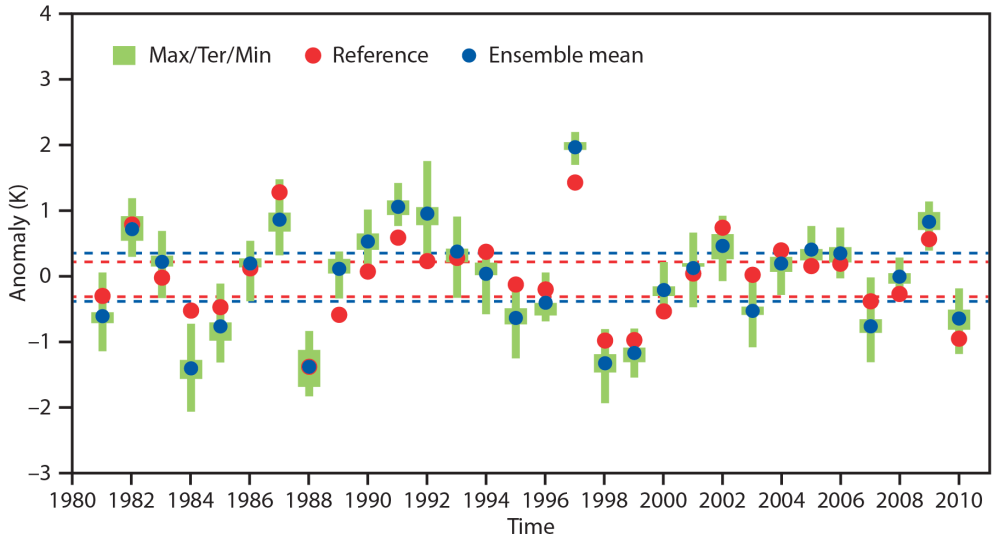

Figure 3: Time series of SST anomalies averaged over the NINO3 area covering the period 1981-2010 and valid for JJA. Blue and red dots represent respectively the ensemble mean and the observed values. The green boxes represent the distribution of the ensemble forecasts.

The calibration used at ECMWF is very simple:

- Seasonal predictions for any atmospheric variable are issued in term of anomalies, in this way the systematic errors associated with the climate drift is removed.
- The SST anomalies over the tropical Pacific (NINO indices) are corrected by the process of bias removal and have re-scaled amplitudes.

For the SST Nino Indices, the variance of the model output is scaled to match the observed variance. (As with other skill assessment, this is done in cross-validated mode for the re-forecast period, which decreases the calculated scores slightly). Note that this technique is not the same as optimizing the amplitude of the forecasts to give the best root mean square error, which might give the best re-forecast statistics but risks damping forecast anomalies unrealistically towards zero, if the forecast performance is poor. Figure 3 demonstrates the impact of such a variance correction. The uncorrected S4 forecasts are appreciably worse than the previous operational system 3 (S3) in terms of mean-square skill score (MSSS), and have a large overestimation of amplitude. The corrected S4 forecasts have much higher skill than S3. Note that because S3 is underactive, re-scaling the amplitude does not typically help the MSSS – in most cases, it makes it worse. The reason why this re-scaling is so successful is that S4 has a very high anomaly correlation skill in the east Pacific.

The re-scaling does not prevent the fact that during the simulation the atmosphere interacts with too large SST anomalies. However, diagnostics show that the atmospheric response to a 1 deg SST anomaly in S4 typically has a realistic spatial structure, but the amplitude is too weak (Molteni et al. 2011) – that is, for S4 the amplitude of SST teleconnections are more realistic if the SST anomalies are larger than observed. In other words, the effect of too large ENSO SST anomalies is compensated to a large extent by a reduced sensitivity to those SST anomalies.

Sometimes people tend to make a distinction between calibrated model output which has been corrected for systematic errors and re-calibrated model output which has been corrected for model skill in addition to systematic errors. Calibration offers significant prospects for forecast improvement. However, one has to acknowledge that there is probably no universally optimal calibration of probabilistic forecasts. Different users may want to adapt the calibration procedure according to their needs. Someone, who would like to predict values at a station location, will calibrate using data from that specific station; while someone else who is interested in values on scales of, say order of 100 km, would calibrate against analyses or up-scaled observations. Some users may need predictions of the joint probability distribution of several variables. An alternative way to calibrate probabilistic forecasts is by combining output from several models. This is called “the multi-model approach”. The combination of several independent models widens the ensemble spread by sampling model errors. Typically, the multi-model forecast, a better representation of the full range of uncertainties, is more reliable than a single model forecast.

3 Skill assessment

A comprehensive assessment of the seasonal forecast skill is crucial for the correct use of the forecast information. In addition, verification statistics provide important feed-back to the model developers. Typically, for the verification of seasonal forecast ensembles, a set of deterministic scores is applied to assess the skill of the ensemble mean. Since no single metric can fully represent the quality of the probabilistic forecasts, probabilistic forecasts are verified using a wider set of probabilistic scores. The Commission for Basic Systems (CBS) of the World Meteorological Organisation (WMO) has produced a comprehensive documentation of skill levels measured according to a common standard. This initiative was taken to promote: the assessments of the scientific quality of long-range forecasts using a standard method and its provision to users. Providing verification for a few seasons or even over a few years may be misleading and may not give a fair assessment of the skill of any long range prediction. Seasonal predictions should be verified over as long a period as possible using consistent re-forecast data. Although there are limitations on the availability of verification data sets and, in spite of the fact that validating numerical forecast systems in re-forecast mode requires large computer resources, the re-forecast period should be as long as possible. WMO Standard Verification System suggests a minimum period of 20 years. Typically verification is performed in cross-validation mode. Since the seasonal forecast skill depends strongly on the season, forecast averages for 3 months are evaluated separately for different starting months.

Figure 5 is an example of deterministic scores for SST forecasts averaged over two key areas over the equatorial Pacific (NINO3.4: 5N-5S 170-120w and NINO3 5N-5S 90W-150W). The top panels show the Root Mean Square Error (solid lines) and the spread of the ensemble (dashed lines), with red indicating values for S4 and blue values for S3. The bottom panels show the anomaly correlations for the same areas. If an ensemble is efficient in predicting forecast uncertainties, the observations should be statistically indistinguishable from the ensemble members of the forecast. In other words, most of the time the verification is included in the range of predicted values. The performance of the ensemble is therefore assessed by the relation between the spread and the RMSE for a long term sample. The closer the RMSE is to the spread, the better is the forecast in representing the uncertainties. It is clear from both top panels that the relationship between the RMSE and spread is better reproduced in S4 than in S3.

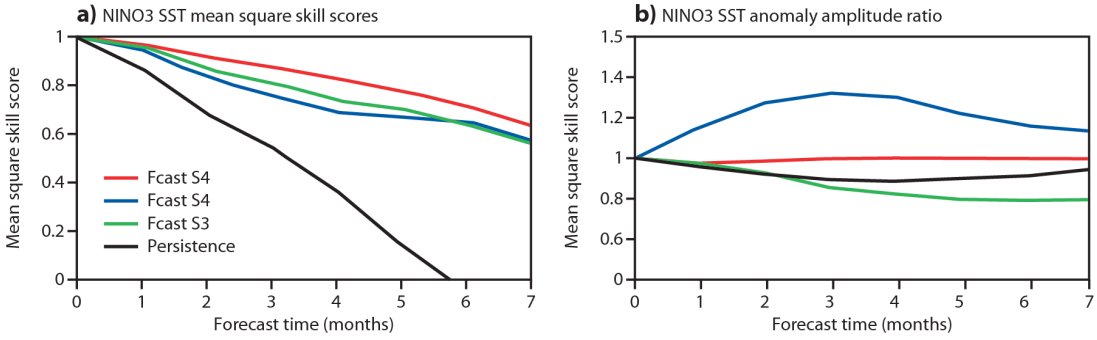

Figure 4: Nino 3 statistics for the ECMWF current operational system (S4) with (red) and without (blue) variance correction and previous operational system S3 (green). Left: mean-square skill scores; right: anomaly amplitude with respect to observations.

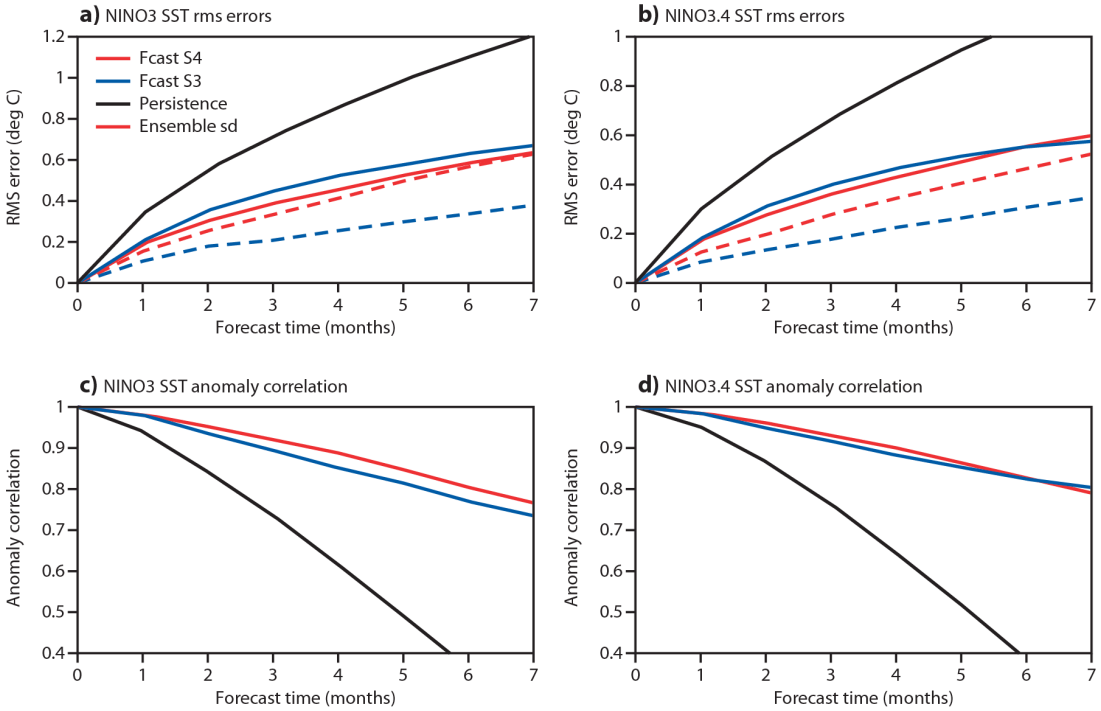

Figure 5: NINO 3 and NINO3.4 deterministic scores based on 30 years period. Dashed lines represent the ensemble spread.

The performance of the seasonal predictions is not only measured in terms of ENSO scores. Figure 6, for example, shows the 2m temperature local correlation between the ensemble mean and era-interim. The regions with orange to red shadings are the regions with correlations larger than 0.6. Fig.6 compares the anomaly correlation between the current system, S4, and the previous one, S3. The forecasts have starting dates on first of May and are valid for the June to August season. The most recent system shows an overall higher level of skill over East Pacific and West Africa. Over Scandinavia the skill is also improved, although we would need a much larger number of cases to estimate the significance of such improvement, since this is a region where the signal to noise ratio is rather small. In fact, it is easy to note that there is higher predictive skill in the tropics and a much lower skill over the mid-latitudes.

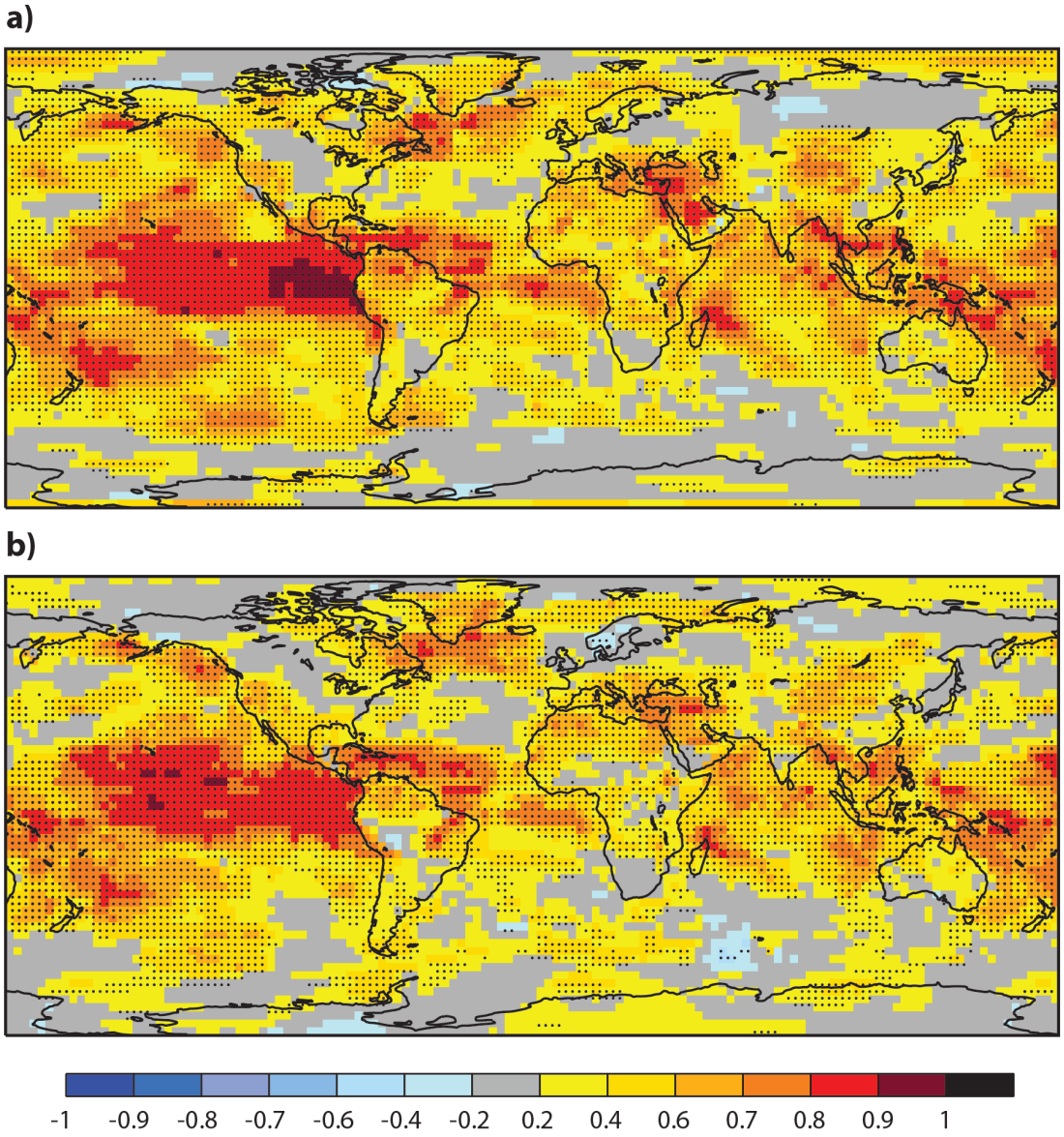

Figure 6: Ensemble mean anomaly correlation for 2m temperature predictions valid for JJA: S4 (top) S3 (bottom).

The most transparent way to illustrate the performance and characteristics of a probabilistic forecast system is the reliability diagram (Figure7), where the x-axis is the predicted probability and the y-axis the frequency with which the forecasts verify. When the forecast probabilities agree with the frequency of events for this particular probability, the distribution should lie along the 45° diagonal. In such a case, the probability forecasts are considered reliable. The frequency of forecast probabilities is represented by red circles of varying sizes. The probabilities should cluster as far away as possible from the climatological frequency, here assumed to be 50%. If climatological probability averages were used as forecasts, they would yield perfect reliability, since the distribution would be exactly on the 45° diagonal, but this would not be very useful. Ideally, we want the forecast system, while mainly reliable, to span as wide a probability interval as possible, with as many forecasts as possible away from the climatological average and close to 0% and 100%. The property of a probabilistic forecast system to spread away from the climatological average is called sharpness. Figure 7 shows the reliability for predictions of 2m temperature anomalies being in the upper third of the model distribution computed for the 30 years period 1981-2010. It shows that forecast is more reliable over the tropics where the ENSO signal is dominant and the predictability is high. Over the extra-tropical regions like Europe (Figure 8a) the effect of ENSO is less strong the signal to noise ratio is much smaller so that the predictability is limited and the uncertainties on the skill estimates are larger.

There are strong variations in the estimate of the skill measure, depending on the sample size and the predictability range. Kumar 2009 showed that for high predictive ranges like for the tropical regions the spread of the estimates using different sample sizes is smaller than for lower predictive ranges. He estimated that to have an accurate measure of the anomaly correlation (let's say an estimate with spread < 0.1) we need to consider a sample of 20 years. On the other hand over Europe where the predictive skill is much lower (typically acc<0.6) in order to have the same level of spread in the estimate of the anomaly correlation we need to consider a sample of about 40 years. For a reduced verifying sample the uncertainty on the scores can be large. It is then particularly relevant to estimate some confidence intervals in order to set some bounds on the expected value of the verification score. This also helps to assess whether differences between competing forecast systems are significant.

The skill of the seasonal predictions is mainly associated with the ability to predict ENSO and its influence over remote regions (teleconnections). It follows that in order to sample the ENSO variability, the skill analysis should be performed for a sufficiently long period. Since the skill assessment of the seasonal predictions is based on the re-forecast performance, the size of the re-forecast (length and ensemble size) can affect the skill estimates. The size of the re-forecast (length and ensemble size) is often a subject of debate. Ideally, verification statistics should be based on the real-time forecasts. However, this would make it difficult to assess the conditionality to low frequency variability such as ENSO. Currently there is no clear answer to this dilemma: long re-forecast (30 years) with a large number of ensembles can be simply too expensive.

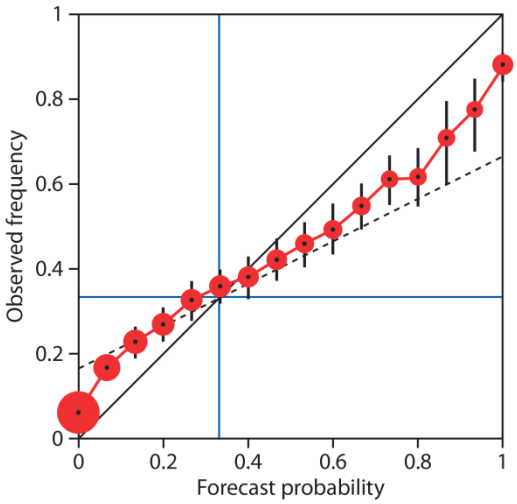

Figure 7: Reliability diagram for JJA 2m temperature in the upper tercile category over the tropics based on 30 years of re-forecasts.

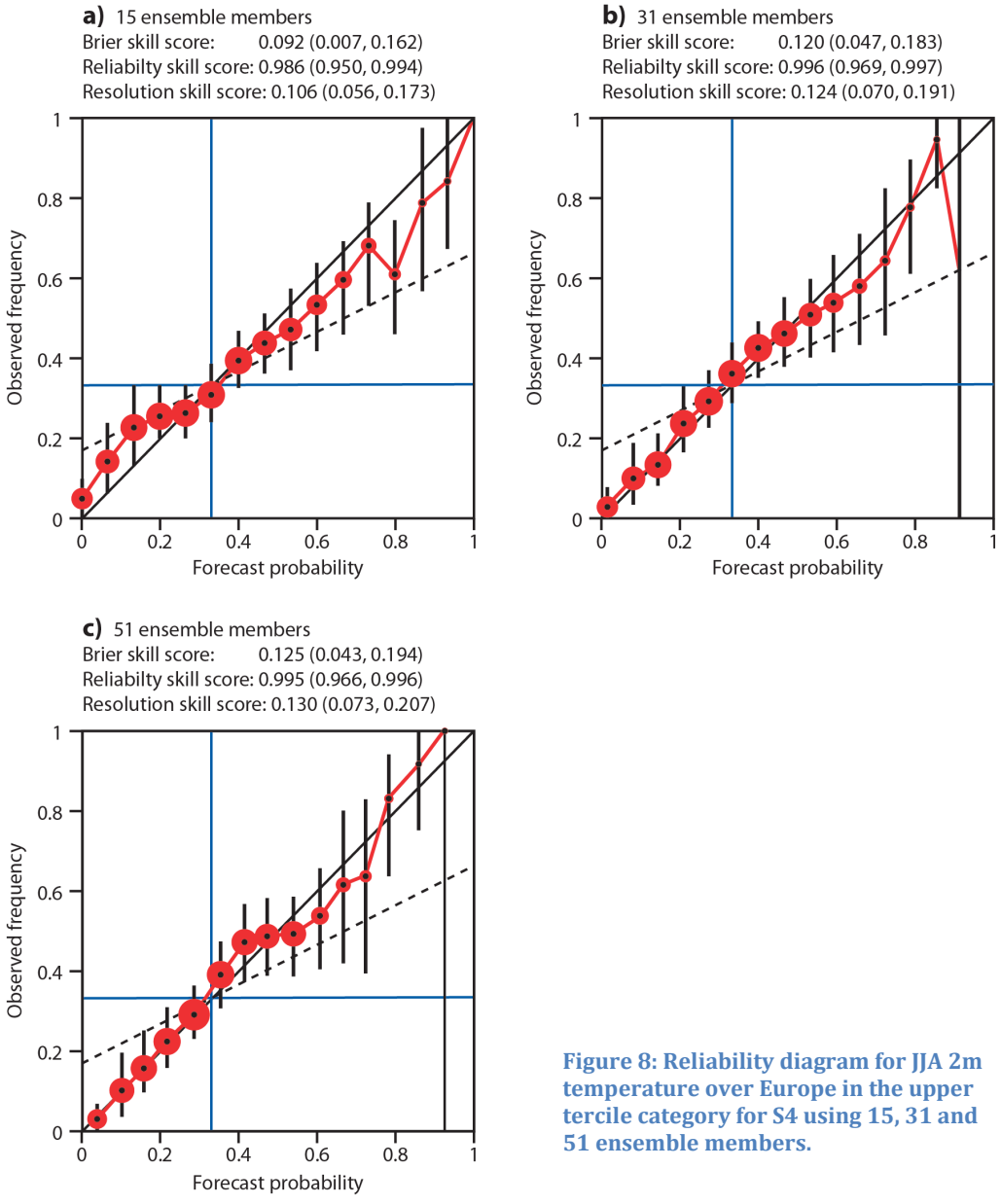

Figure 8: Reliability diagram for JJA 2m temperature over Europe in the upper tercile category for S4 using 15, 31 and 51 ensemble members.

Barnston et al. 2012 analysed the real-time ENSO prediction skills for the period 2002-2011 looking at 20 (12 dynamical 8 statistical) models. Those models contribute to the IRI ENSO prediction plume. The study showed that during the 9 years period (2001-2011) the ENSO events had smaller amplitudes and presented a larger number of alternations among consecutive years. Barnston concluded that the ENSO decadal variability can hide the higher skill of todays' models. Long term trends present in the verification period affect the skill assessement. If we consider the surface air temperature during the last 30 years this exhibits a warming trend. This global warmth in the last decades is a continuation of the upward warming trend observed since the mid-20 century in response to the increase of Green House Gases (GHGs). Several studies discussed the importance of an adequate representation of the GHGs in the coupled climate models used for seasonal predictions (Doblas-Reyes at al. 2006, Cai et al. 2009). In the skill assessment based on 30 years period, the ability of reproducing the effect of climate change will be accounted for, as well as, the actual skill in predicting the year-to-year variations of anomalies.

Several authors have studied the dependence on ensemble size of the probabilistic scores. (e.g. Richardson 2001; Kumar et al. 2001, Mason 2004). Kumar et al. 2001 showed that the ensemble size of 10-20 members is sufficient to estimate the skill only for moderate ENSO cases. Weigel et al. 2007 suggested the use of a de-biased Brier and ranked probability skill score to avoid the dependency on the ensemble size and to assess forecast with small ensemble size. Fig.8 shows the reliability over Europe for the current operational system computed by using the re-forecast with 15 members, 31 and 51 members. As expected the reliability increases substantially going from 15 to 51 members. A systematic analysis of reliability estimates as a function of the ensemble size is in progress. So far we can only say that the reliability estimates over the extra-tropics, being based on a reduced sample size, (15 members of the reforecast versus 51 of the real-time forecast) are likely to underestimate the actual reliability of the current seasonal forecast system.

4 References

Barnston, A.G., M. K. Tippett, M.L. L'Heureux, S. Li, and D.G. DeWitt, 2012: Skill of Real-time. Seasonal ENSO Model Predictions during 2002-2011—Is Our Capability Increasing? Bull. American Met. Society, DOI:10.1175/BAMS-D-11-00111.1

Cai W and T. Cowan 2009: Trends in Southern Hemisphere Circulation in IPCC AR4 Models over 1950–99: Ozone Depletion versus Greenhouse Forcing. J. Climate, 20, 681–693. doi: http://dx.doi.org/10.1175/JCLI4028.1

Doblas-Reyes, F.J., R. Hagedorn, T.N. Palmer and J.-J. Morcrette (2006). Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. Geophysical Research Letters, 33, L07708, doi:10.1029/2005GL025061.

Mason 2004: On using "climatology" as reference strategy in the Brier and ranked probability skill scores. Mon. Wea. Rev. 132 1891-1895.

Molteni and Coauthors 2011: The new ECMWF seasonal forecast system: system 4. ECMWF Tech. Memorandum No.656.

Richardson 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. Quart. J.Roy. Meteor. Soc.127 2473-2489.

Weigel P. A. Liniger and G. Appenzeller 2007: The discrete Brier and ranked probability skill scores. Mon.Wea.Rev. 135 118-124.