On the Reliability of Seasonal Forecasts

T.N. Palmer and A. Weisheimer

University of Oxford and ECMWF

1 Introduction

On a scale of 1-5, where 5 is very good, how skilful are seasonal forecasts today? On a similar scale, how skilful can we expect seasonal forecasts to be 30 years from now? These types of question are sufficiently vague and open-ended that they appear impossibly difficult to answer in any meaningful way. And yet precisely these types of question are being asked by policy makers e.g. by the UK Government as it considers options for future investment in science (*Foresight, 2012*).

Whilst forecast skill for El Nino itself is impressive, nobody would suggest that seasonal forecast skill currently merits a "5" for example over Europe. But even if we are optimistic and suppose that a "5" is achievable in all regions of the world in 30 years time, what would this mean? Would it mean that every time there is to be a drought in Eastern Africa, or a BBQ summer in the UK, it will be predicted unequivocally six months ahead of time? That is clearly a scientifically impossible goal. Climate is chaotic and seasonal forecasts must necessarily be probabilistic, reflecting the amplification of inevitable initial and model uncertainties.

In this paper we discuss what a "5" might mean in a probabilistic sense, how close we are to achieving a "5" today, and what is needed to achieve a "5" in 30 years time. It will be concluded that focussing on a single skill score may not be the best way of assessing the utility of a forecast system, and we propose a rather different measure by which a system can be rated "5", based on its reliability when non-climatological probabilities are forecast.

2 Probabilistic Skill

Forecasts are used to make decisions. For example, an agronomist might be tasked to advise farmers on what type of crop to plant in the coming season. Suppose there is a choice between two types of crop: A and B. The crop yield (tons/hectare) C_A and C_B of A and B depends on a number of meteorological variables such as temperature and precipitation. These are collectively labelled by X. Hence $C_A = C_A(X)$ and $C_B = C_B(X)$. Suppose a forecast system predicts over a given season a probability distribution $\rho(X)$ for X. Then the expected crop yield for A and B is

$$\langle C_A \rangle = \int_X C_A(X)\rho(X)dX$$

$$\langle C_B \rangle = \int_X C_B(X)\rho(X)dX$$

If $\langle C_A \rangle > \langle C_B \rangle$ the agronomist might recommend A over B and vice versa. (In practice of course there may be many factors other than climate which determine the agronomist's advice, but let us suppose here that climate is the only relevant one.)

Now in general, one can expect C_A to be a nonlinear function of *X*. Hence $\langle C_A \rangle$ will depend on more than just the mode of ρ. The uncertainty, given by the spread of the forecast distribution, might also have a substantial impact on the estimate $\langle C_A \rangle$.

It can be assumed that the agronomist knows the climatological distribution $\rho_C(X)$ of X. Let us assume that

$$\langle C_A \rangle_C \equiv \int_X C_A(X)\rho_C(X)dX > \langle C_B \rangle_C \equiv \int_X C_B(X)\rho_C(X)dX$$

Now let's suppose that the forecast system is reliable, but in the majority of forecast occasions, ρ(X) is not significantly different from ρ_C.(X). Then whilst the agronomist is not going to gain any specially useful information from the forecast system, (s)he is not going to mislead the farmer with unreliable information. However, on occasions where $\rho(X) \neq \rho_C(X)$ such that $\langle C_A \rangle < \langle C_B \rangle$, knowing that the forecast system is reliable is essential if the agronomist is to recommend B over A to farmers.

One way to assess whether such ρs are reliable when $\rho \neq \rho_C$ is to study so-called Attributes (or "Reliability") Diagrams. Such diagrams are shown in the next Section.

In this paper we are going to study the reliability of ECMWF's System 4 seasonal forecasts, focussing on the situations where $\rho \neq \rho_C$.

3 Reliability of System 4

Here we assess Attributes (or Reliability) Diagrams for the ECMWF System 4 seasonal forecast system as one of the best state-of-the-art dynamical seasonal forecasting systems in the world. It is based on the coupled atmosphere-ocean model IFS/NEMO with a horizontal atmospheric spectral resolution of TL255 (~80km) and a 1° x 1° resolution for the ocean component in mid-latitudes and enhanced meridional resolution near the equator. System 4 became operational in November 2011 and produces probabilistic forecasts of global seasonal-mean climate conditions every month. We focus our analysis on the December to February (DJF) and June to August (JJA) seasons initialised on 1st November and 1st May, respectively. Attributes Diagrams are computed from the retrospective forecasts (re-forecasts) over the 30-year period 1981-2010 using 15 ensemble members. The verification data we use are re-analyses of 2m temperature and sea surface temperatures (SST) (*Dee et al., 2011*) and GPCP for precipitation (*Adler et al., 2003*). All data considered are either observed or modelled anomalies with respect to the 30-year re-forecast period climatological mean estimated in leaving-one-year-out cross-validation mode.

In the following we are going to consider dichotomous events *E* based on terciles of the climatological forecast distribution of seasonal forecast anomalies. If *E* is defined as falling into the lower tercile of the distribution, the event is called "cold" for 2m temperature and SSTs, and "dry" for precipitation. Similarly, if *E* relates to the upper tercile, the event is called "warm" or "wet".

Attributes Diagrams summarise for a given event the correspondence of the forecast probabilities with the observed frequency of occurrence of the events given the forecast. For example, consider the wet event *E* that the precipitation anomaly lies in the upper tercile of the climatological forecast probability distribution. Suppose the seasonal forecast probability for the event is equal to 0.8. Then, in a reliable seasonal forecast system, *E* would actually occur, taking into account sampling uncertainty, on 80% of occasions where *E* was predicted with a probability of 0.8.

Figure 1 shows an example of an attributes diagram for the warm SST event *E* during DJF in the Nino3.4 region of the central tropical Pacific. The vertical and horizontal line intersection at 1/3 indicates the climatological frequencies of 1/3 of *E* for the forecasts and observation, respectively. The diagonal in the diagram is the line of perfect reliability. The black dots show how well the binned forecast probabilities for this event verify in terms of frequency of occurrence. Note the size of the dots is proportional to the number of cases falling into that frequency bin. Each of the dots has 95% confidence intervals for the estimate of the observed frequency attached to them (for big dots the intervals might become invisible). Here, the confidence intervals were estimated from a bootstrap procedure with 1000 resamples. A weighted linear regression has been fitted to the data points as the "reliability curve" and is shown by the black line. If the reliability curve has a slope larger than the slope indicated by the dashed grey line, the forecast will have a positive Brier Skill Score compared with a climatological reference forecast. If the reliability curve is flatter than the dashed no-skill line, then the forecast system is overconfident. For flat curves there is no relationship between the forecast probabilities and the observed frequencies of occurrence and the system is not reliable beyond climatology.

How would a reliability curve link to the 1-5 scale mentioned above? As discussed above, it is unrealistic to expect an "ideal" seasonal forecast system to produce unequivocal forecasts on all occasions. Indeed, it may be unrealistic to expect an "ideal" seasonal forecast system to produce forecast probabilities that differ from climatology on all occasions – seasonal predictability may be a more intermittent property of nature. Hence we shouldn't penalise a seasonal forecast system because it predicts climatological probabilities on occasion. That is to say, a "5" shouldn't necessarily be defined in terms of high Brier Skill Score. Rather, we take the view here that an "ideal" seasonal forecast system should be rated "5" (for given region and event) if the corresponding reliability curve is, within the error bar uncertainty, on or very close to, the diagonal. A "4" will be given to cases where the slope of the regression line is positive and larger than the slope of the dashed no-skill line. If the slope of the reliability curve is positive but less than the dashed line, we will give it a "3". Near-horizontal reliability curves will score as "2" and those curves that have a negative slope will be getting a "1".

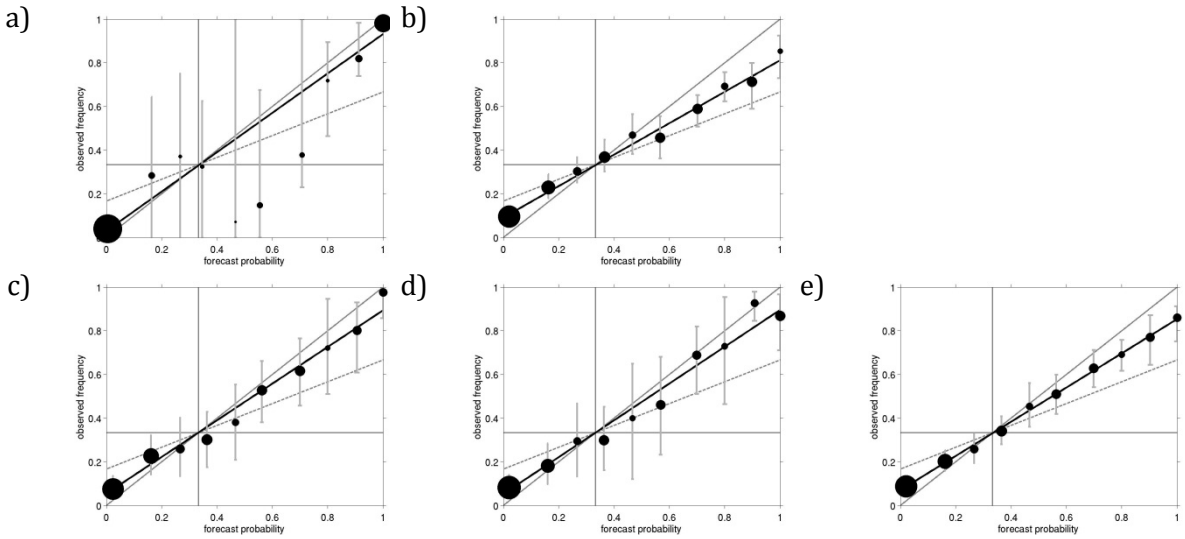

Figure 1: Attributes (Reliability) Diagrams for warm SSTs in DJF in the a) tropical central Pacific Nino3.4, b) tropical Atlantic, c) western tropical Indian Ocean, d) eastern tropical Indian Ocean and e) tropical Indian Ocean.

El Nino Southern Oscillation (ENSO) as the dominant coupled atmosphere-ocean mode of variability on seasonal time scales is the central phenomenon for seasonal forecasting. Figure 1 shows the Attributes Diagrams for warm SST events in DJF for different tropical ocean areas. The seasonal forecasts of SSTs in the central tropical Pacific Nino3.4 region (Fig 1a) are extremely reliable with the two principal forecast probabilities of ~0 and ~1 lying almost exactly on the diagonal. The remaining forecast probabilities between 0 and 1 were populated by very few data and thus have a small relative weight but large error bars. Fig 1a indicates that the SST forecasts for this event are very close to perfect deterministic forecasts where the event is either correctly predicted to occur with certainty, or correctly predicted to not occur with certainty. We would thus rate the Fig 1a ENSO reliability with a "5".

SST forecasts for the tropical Atlantic, as displayed in Fig 1b, show a positive slope of the reliability curve which is larger than the no-skill line slope. We score these forecasts with a "4" according to our scoring rules mentioned above.

Forecast evaluations for the tropical Indian Ocean SSTs are shown in Fig 1c-e. Here, the western Indian Ocean (Fig 1c) and the eastern Indian Ocean (Fig 1d) perform with a near-perfect reliability ("5") whereas the entire tropical Indian Ocean basin in Fig 1e would rate at a "4". Note that in contrast to the tropical central Pacific, for both the tropical Atlantic and the tropical Ocean the population of the forecast probability bins is much more homogeneous across the range of possible forecast probabilities, indicating how important it is to have a probabilistic forecasting system.

How reliable are the S4 seasonal forecasts for near-surface temperature and precipitation on the global scale? Forecasts of cold boreal winters (Fig 2a) and warm boreal summers (Fig 2b) computed from all model grid points on a 2.5° grid show a reliability curve that lies in the skilful area of the diagram and thus would be ranked a "4". However, global precipitation forecasts are less reliable and can only be scored as "3" (Fig 2d for dry JJA) to "4" at most (Fig 2c for wet DJF).

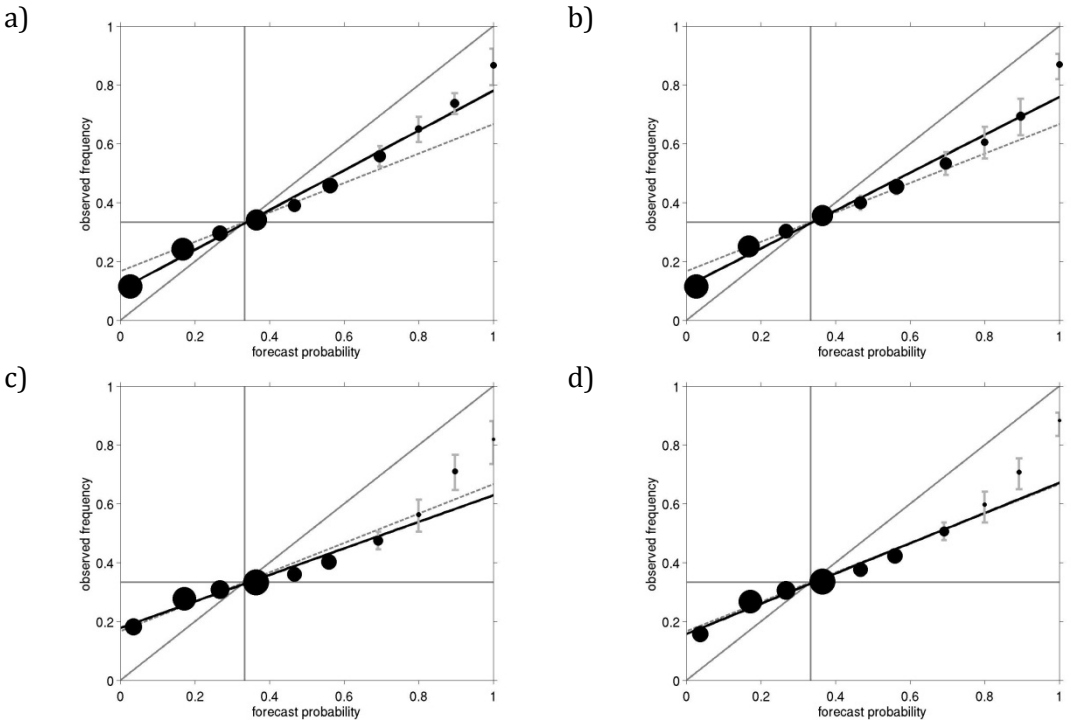

Figure 2: Attributes (Reliability) diagrams for global a) cold 2m temperature in DJF, b) warm 2m temperature in JJA, c) wet precipitation in DJF and d) dry precipitation in JJA.

In the following we analyse the temperature and precipitation reliability performance of S4 for selected regions over land, as these are areas that are central for the use of seasonal forecast information. For the definition of the areas, see *Giorgi and Francisco (2000)*. As demonstrated in Fig 1a, forecasts of the tropical Pacific ENSO SST events over the next season are highly skilful. The continental areas that are most directly affected by ENSO teleconnection patterns in the atmosphere are South America and the Maritime Continent of South-East Asia and one would thus expect that these areas also show good skill in forecasting seasonal climate anomalies. Indeed, warm and wet DJF forecasts in the Amazon region (Fig 3a and 3b) show a positive slope of the reliability curve rating with a "5" and a "4", respectively. Similarly, cold and wet JJA over South-East Asia in Fig 4 are also reasonably reliable with scores of "4".

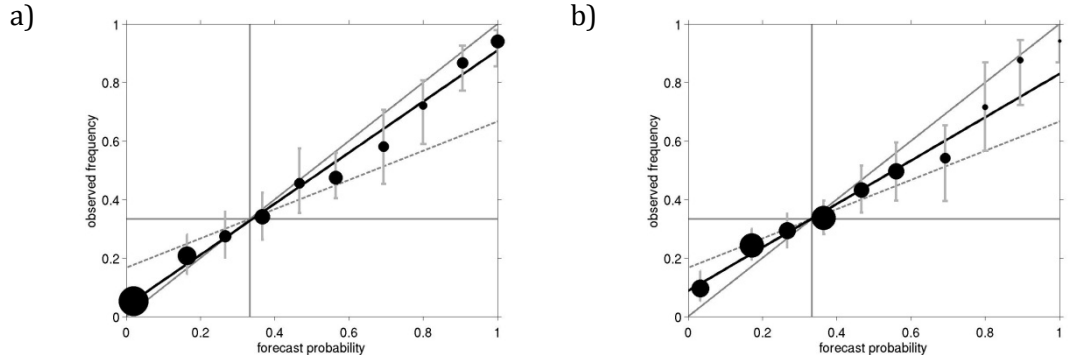

Figure 3: Attributes (Reliability) Diagrams over the Amazon region for a) warm 2m temperature in DJF and b) wet precipitation in DJF.

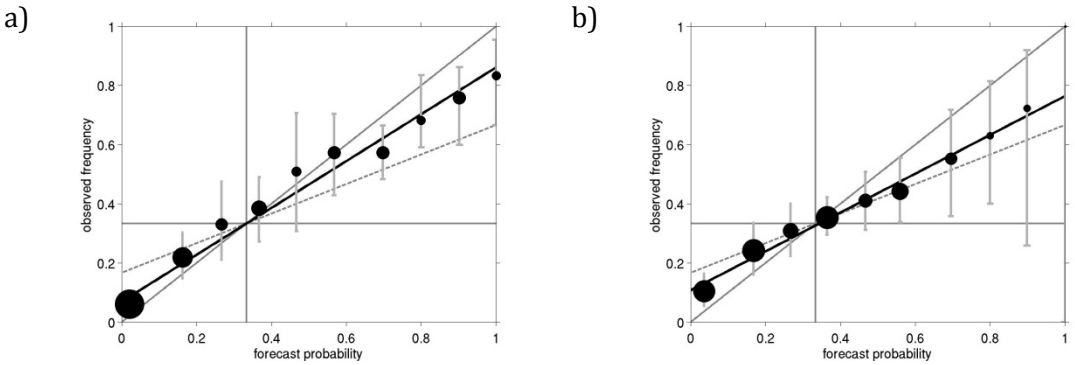

Figure 4: Attributes (Reliability) Diagrams over the South-East Asia region for a) cold 2m temperature in JJA and b) wet precipitation in JJA.

How reliable are forecasts of the Indian monsoon? In Fig 5a we show the attributes diagram for wet JJA over South Asia. The reliability curve has a positive slope which means that in general there is a positive relationship between forecast probabilities and the frequency of occurrence. However, the curve does not fall into the skilful area of the diagram as bounded by the dashed no-skill line and thus rates as a "3". Note that there are very few cases of forecasts with probabilities > 0.5 and that the forecasts mainly cluster around the climatological probability of 1/3. The forecasts for wet months of May which corresponds to the first month of the seasonal forecasts initialised on 1st May indicate a better reliability (Fig 5b) with a positive and skilful reliability curve rated at "4".

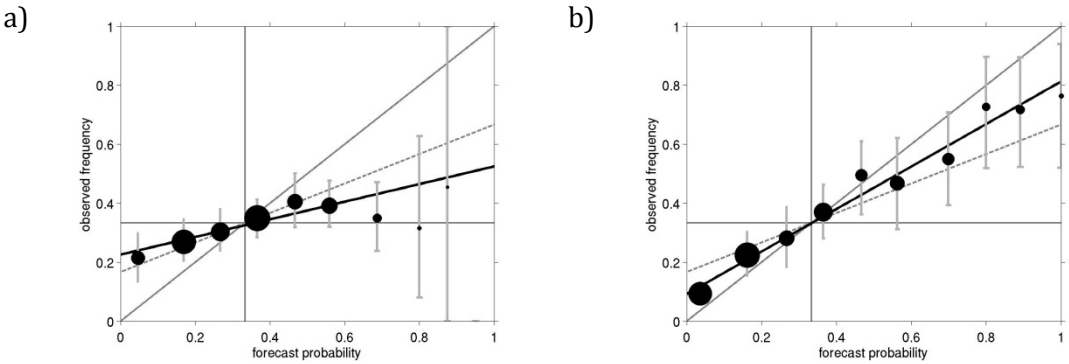

Figure 5: Attributes (Reliability) Diagrams over the South Asia (Indian sub-continent) region for a) wet precipitation in JJA and b) wet precipitation in May (1st forecast month).

The above diagrams were examples of moderately to very good seasonal forecast skill in terms of forecast reliability. However, seasonal predictability in the extra-tropics is, in general, lower due to the internal variability of the coupled atmosphere-ocean system, the lack of relevant teleconnection mechanisms and difficulties of general circulation models to simulate these. For example, Fig 6a shows the attributes diagram of wet summers (JJA) over central North America. For this case, the fitted reliability curve has a negative slope indicating a weak inverse relationship between the forecast probability of the event and the frequency of the event eventually occurring. This means, for example, that if the event *E* is forecast with a high probability, it actually is not likely to occur. Such forecasts clearly rate as a "1" (very poor!) in our scale of forecast skill.

It is interesting to note though, that on the monthly time scale the reliability of the forecasts for the same event is extremely good (marked at a "5"), as shown in Fig 5b. This is an example of a forecasting situation where the impact of the initial conditions diminishes after the first month and the lack of useful sources of seasonal forecasting skill and imperfect representations of model uncertainty lead to a very poor performance for longer lead times.

A similar though not as drastic situation arises for Northern European wet winters (Fig 7a). Here, most of the forecast probabilities cluster around the climatological background probability. The slope of the regression reliability curve is positive but less than for a skilful system. We give such a forecasting performance a "3". Again, the reliability for the first month of the forecast is clearly better than for the first season of the forecast (Fig 7b) and would rate as a "4" out of 5.

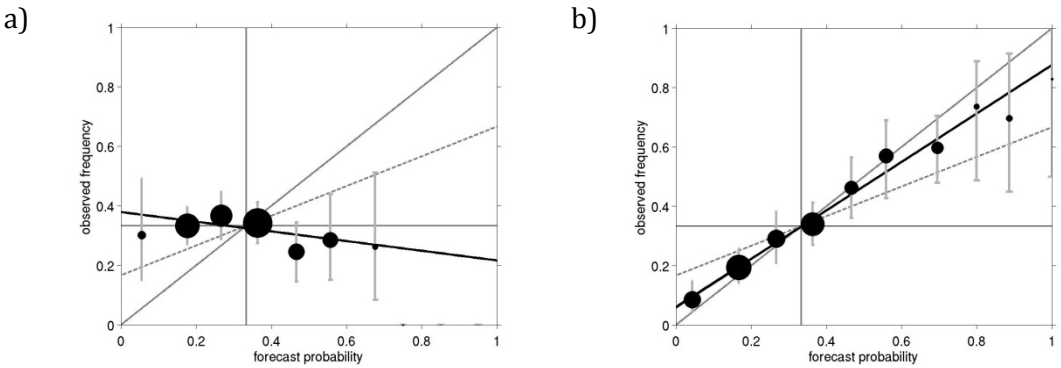

Figure 6: Attributes (Reliability) Diagrams over the Central North America region for a) wet precipitation in JJA and b) wet precipitation in May (1st forecast month).

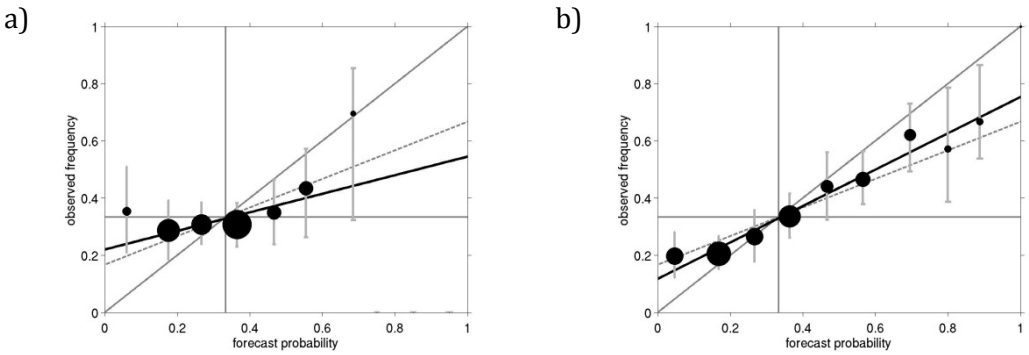

Figure 7: Attributes (Reliability) Diagrams over the Northern Europe region for a) wet precipitation in DJF and b) wet precipitation in November (1st forecast month).

Our last example is the forecast skill for Southern European dry summers in Fig 8a. The reliability curve is nearly flat which indicates virtually no relationship between the forecast and the observations: the forecast probability is mostly irrelevant for predicting the occurrence of dry Mediterranean summers. Such a forecast scores as a "2". In contrast, the forecasting system performs better for predicting dry conditions in May (Fig 8b) with a positive and just skilful slope of the reliability curve ("4").

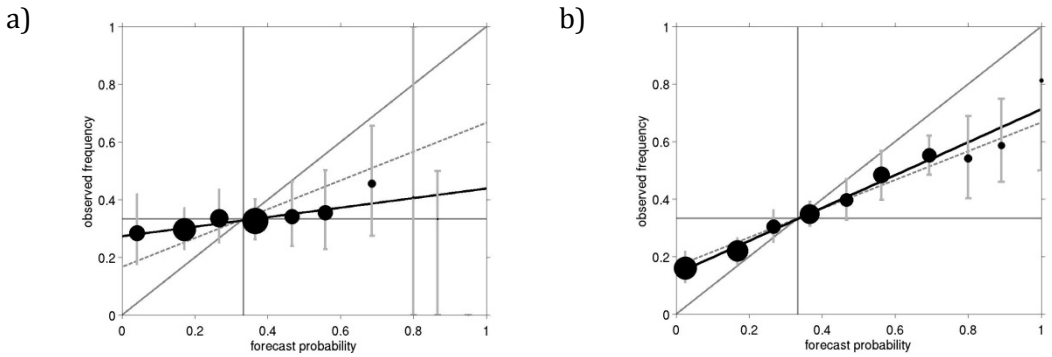

Figure 8: Attributes (Reliability) Diagrams over the South Europe/Mediterranean region for a) dry precipitation in JJA and b) dry precipitation in May (1st forecast month).

4 How can seasonal forecast reliability be improved?

The results above suggest that we still have some way to go before we can say we have achieved the goal of providing users with reliable forecasts, particularly for precipitation and away from the El Nino region. Of course, it is always possible to calibrate the forecasts so that they become reliable *a posteriori.* However, one should not rely on calibration to provide reliability. Firstly such calibration would effectively remove any sharpness from the distributions and thus the resulting system may have little value over that of a climatological forecast. Secondly, one cannot be certain that the system will remain reliable out of sample.

There can be little doubt that the ability to represent physical processes accurately is key to improved reliability. In a recent study based on Athena integrations (*Jung et al., 20112*), *Dawson et al. (2012*) was able to show in AMIP integrations that the ECMWF model could simulate the non-Gaussian structure of observed Euro-Atlantic weather regimes more accurately in a T1279 model than a T159 model. It is plausible that the improved simulation of such weather regimes in a T1279 model is associated with better representation of topography on the one hand, and with a more realistic representation of Rossby wave breaking on the other.

A better representation of other Earth System components is also likely to improve reliability. For example, *Weisheimer et al. (2011a)* showed that a better representation of land surface processes led to remarkably good probabilistic forecast of the summer 2003 heat wave.

On the other hand, since the climate system is chaotic, it is necessary to represent inevitable uncertainties in the representation of processes which have to be parameterised. There has been a programme to represent parameterisation uncertainty using stochastic methodologies for some time at ECMWF (*Buizza et al., 1999; Palmer, 2001; Palmer, 2012*). On the monthly and seasonal timescales there is evidence that it is competitive with, and for temperature predictions can outperform, the more standard multi-model ensemble approaches to the representation of model uncertainty (*Weisheimer et al., 2011b*).

There can be little doubt about the value to society of reliable non-climatological predictions of seasonal climate. However, to develop a high resolution system with accurate stochastic representations of model uncertainty in all relevant components of the Earth System, is not only a formidable technical challenge, it is one that will require computing resources which are unavailable to individual institutes in the foreseeable future. A possible route to achieve the goal of a reliable seasonal climate prediction system, based on much stronger international collaboration, has been presented elsewhere (*Shukla et al., 2010; Shapiro et al., 2010; Palmer, 2011; Palmer, 2012*).

5 Conclusions

Let us return to the question posed in the Introduction. What constitutes a "5", to which a seasonal forecast system should aspire? Here we propose the following broad criterion for rating a seasonal forecast system a "5": when the system predicts probabilities $\rho(X)$ that are substantially different from the climatological distribution $\rho_c(X)$ then these probabilities can be relied on, and acted on by decision makers. Note that we make no firm statement about how often such situations arise. It may be that in many cases the forecast system does not predict probabilities that differ substantially from $\rho_c(X)$. If this is the case, then the probabilistic skill score may not be particularly high. However, for such a forecast system, a user will not make a bad decision based on unreliable forecast information.

The ECMWF System 4 cannot be rated a "5" for all regions of the world, and for all variables. We have shown that for temperature, and even more for precipitation, forecast probabilities are not reliable when different from climatology and away from the El Nino region. Based on current performance and expected increases in resolution and better stochastic representations of model uncertainty our current capability to forecast seasonal climate could perhaps be rated 2/5 overall, with the potential to rise to 4/5 by 2040.

6 References

Adler, R.F. and coauthors, 2003: The Version – 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979 – present). *J. Hydrometeor.*, 4**,** 1147 – 1167.

Buizza, R., M.J. Miller and T.N. Palmer, 1999: Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **125**, 2887-2908.

Dawson, A., T.N. Palmer and S. Corti, 2012: Simulating regime structures in weather and climate prediction models. *Geophys. Res. Lett.*, **39**, L21805. doi:10.1029/2012GL053284.

Dee, D.P., S.M. Uppala, A.J. Simmons and coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data ssimilation system. *Q. J. R. Meteorol. Soc.*, **137**, 553-597. doi: 10.1002/qj.828.

Foresight, 2012: Reducing the Risks of Future Disasters: Priorities for Decision Makers. *The Government Office of Science, London*.

Giorgi, F. and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.,* **16**, 169-182.

Jung, T. and coauthors, 2012: High-resolution global climate simulations with the ECMWF model in Project Athena: Experimental design, model climate and seasonal forecast skill. *J. Climate*, **25**, 3155-3172. doi:10.1175/JCLI-D-11-00265.1.

Palmer, T.N., 2001: A nonlinear dynamical perspective on model error: A proposal for nonlocal stochastic-dynamic parametrisation in weather and climate prediction models. *Q. J. R. Meteorol. Soc*., **127**, 279-304.

Palmer, T.N., 2011: A CERN for climate change, *Physics World*, **24 (3),** 14-15.

Palmer, T.N., 2012: Towards the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Q. J. R. Meteorol. Soc.,* **138**, 841-861.

Shapiro, M. and co-authors, 2010: An Earth-System Prediction Initiative for the Twenty-First Century. *Bull. Amer. Meteor. Soc*., **91**, 1377–1388, doi: 10.1175/2010BAMS2944.1 .

Shukla, J. and coauthors, 2010: Toward a New Generation of World Climate Research and Computing Facilities. *Bull. Amer. Meteor. Soc.,* **91**, 1407–1412, doi: 10.1175/2010BAMS2900.1.

Weisheimer, A., F.J. Doblas-Reyes, T. Jung, and T.N. Palmer, 2011a: On the predictability of the extreme summer 2003 over Europe, *Geophys. Res. Lett.,* **38**, L05704, doi:10.1029/2010GL046455.

Weisheimer, A., T.N. Palmer, F.J. Doblas-Reyes, 2011b: Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles, *Geophys. Res. Lett.,* **38**, L16703, doi: 10.1029/2011GL048123.