Practices for Seasonal-to-Interannual Climate Prediction

by Lisa Goddard & Martin P. Hoerling

Accuracy in seasonal-to-interannual climate forecasts for the United States (US) remains a challenge. This despite advances in understanding sources of climate variability and predictability as well as improvements in prediction tools. Our use of the tools has greatly improved in the past decade with the implementation of robust model bias correction and multi-modeling strategies. Furthermore, validation measures have become more sophisticated, rating the performance of forecast systems in a manner more consistent with the probabilistic world they describe. Still, further room for improvement exists. This article outlines the current practices of seasonal-to-interannual climate prediction: current understanding of the sources of variability, the tools used to predict it, common methodologies applied to those tools to produce forecasts, and relevant verification analyses with which to judge the performance of the forecasts. These are forecasts of opportunity, which if used prudently have potential to benefit decision-making.

Background

Before discussing current prediction practices and their accuracies, it is important to distinguish between prediction and predictability itself. The latter is a physical characteristic of the natural system, and is not altered by forecasting methodologies. The tools used to make forecasts are often employed in determining the theoretical limit of predictability, judging the model against itself, and as such predictability estimates can indeed change (for non-physical reasons) as models evolve. Nonetheless, it is often of interest to know how the current skill levels differ from the existing theoretical limits because such knowledge guides expectations for the skill impacts of improved practices. However, given the indeterminate nature of predictability estimates, this report focuses on skill estimates obtained by comparing model-derived forecasts with the observed climate, emphasizing seasonal mean surface temperature and precipitation variations over the US.

Attributable causes of US seasonal climate variability

Understanding US seasonal climate variability is essential for exposing the sources of its predictability. Seasonal forecasting (when done at the minimal 15-day lead times beyond which deterministic atmospheric predictions are skillful) is effectively the practice of predicting the climate signal due to external forcings. These forcings include anomalous sea surface temperature (SST), soil moisture, sea ice, and chemical constituents. The resulting climate predictability is known as predictability of the "second kind" or 2-tier, arising from the influence of specified boundary conditions on the atmosphere. For seasonal prediction practices using fully coupled Earth System models, the notion of such a 2-tiered system with external forcings vanishes, and predictability is of the "first kind" (ie 1-tier) arising solely from the initial Earth System conditions. It is important to note here that for seasonal prediction, longer-term changes of external forcing that *is* affecting the climate system, especially increasing greenhouse gasses, may be considered constant

over the season, although their changes from year to year should probably be included in dynamical models.

We will subsequently examine the skill of forecasts generated from both 1-tier and 2-tier systems. But, for purposes of discussing seasonal predictability, it is helpful to first consider the 2-tier system. The climate responses to the specified external forcings constitutes the "signal", whose probability of occurrence (i.e. verification) depends upon the signal strength relative to seasonal "noise" arising from internal atmospheric variability. Two approaches have been used to estimate such signals, and both focus on the contribution of SST anomalies to seasonal variability. One involves analysis of historically observed SST anomalies and the accompanying global circulation and surface climate impacts. This approach is illustrated in the studies by Barnett (1981), Horel and Wallace (1981), Ward and Folland (1991), Barnston and Smith (1996), to name only a few. An approximately correct atmospheric signal can be identified forced by the ENSO-related SST anomaly pattern, and to a lesser extent by one or two more localized tropical SST patterns (Hastenrath, 1995; Anderson et al. 1999). The period of globally adequate observational analyses is just long enough to resolve differences in the relationships between different "flavors" of ENSO SST forcing and climate over the US (Larkin and Harrison, 2005), but the record is not long enough to robustly connect presently unrecognized non-ENSO-related SST forcings and US climate. In a second approach, atmospheric models are used to simulate US seasonal climate variations during the past half century These find that ENSO SSTA is the primary source of forecast skill related to ocean influences, and that in ENSO's absence, skill is largely absent (e.g. Goddard & Dilley 2005; Quan et al. 2006) (Figure 1). Further research is required to better understand the role of non-ENSO ocean states in US climate variability.

Additional open questions concern the signals related to land boundary conditions, sea ice states, and the influence of anomalous atmospheric chemical compositions on US seasonal climate. Especially noteworthy is that no current dynamical practice for seasonal forecasting incorporates the direct effect of anomalous chemical composition, and it is unclear to what extent their implicit effect is already incorporated via ocean states. Among a suite of empirical tools employed by NCEP in their operational seasonal forecasts, the trend of surface temperature has been found to explain a large fraction of US seasonal temperature variations during the past decades (Huang et al. 1996), and this tool explains the majority of US temperature forecast skill at lead times greater than 1 season. Yet, neither the strength, seasonality, nor regionality of such trends have been distinguished from possible transient decadal variations. This leaves open the question on the best practice for including trends and their climatic forcings into seasonal prediction practices.

Current prediction tools and methodologies

The tools used for prediction, as mentioned above, include empirical models and dynamical models. Individually, empirical models continue to be competitive with dynamical models, which attests to the dominance of the linear ENSO signal as the primary skill source over the US. It is not clear if this will continue to be the case if

anthropogenically induced changes in the mean state impact the expression of climate variability, such as the teleconnection responses to El Niño conditions in the tropical Pacific or even the expression of ENSO itself. Conversely, the extrapolation of trends by the empirical models has kept pace with the recent increases in the strength and spatial coverage of above-normal temperatures over the US better than the dynamical models used for seasonal prediction (not shown). The most notable change in the armory of prediction tools has been the increasing use of coupled general circulation models (CGCMs) over atmospheric general circulation models (AGCMs). In theory CGCMs are superior to AGCMs because the two-way interaction between ocean and atmosphere can proceed realistically; whereas in an AGCM the ocean does not respond to the atmosphere, which leads to unrealistic air-sea heat fluxes over most regions. One exception is the ENSO region (i.e. near-equatorial Pacific) where the ocean largely forces the atmosphere interannually. US seasonal forecast skill obtained with AGCMs is expected to be comparable to that from CGCMs, because the currently realized skill in US terrestrial climate derives primarily from ENSO SSTA, To date, CGCMs still contain substantial biases in their representation of important boundary fields, such as SSTs. As a result, CGCMs currently do not out-perform AGCMs. That they have the potential to do so suggests possible future improvements to climate predictions as biases in CGCMs are diagnosed and minimized.

Given the existing biases in prediction tools, considerable effort has gone into methodologies that can identify and reduce them. The simplest of these removes the mean bias of the model climatology, and casts the prediction as anomalies relative to some base period. More complex, though less generalizable methods attempt to spatially correct patterns of anomalous climate due to inadequately resolved topography, or poorly captured teleconnection responses (Landman & Goddard 2002, Tippett et al. 2003). Recently, efforts have focused on attempting to recalibrate the probabilistic response of the model (Doblas-Reyes et al. 2005).

While these methodologies do improve individual model performance, one still finds that some climate signals are captured by some models and not others. This suggests that in addition to sampling the uncertainty arising from imperfect knowledge of initial conditions, the uncertainty arising from imperfect knowledge of the physical processes must also be sampled, specifically those represented through parameterizations. Substantial improvements in overall "predictionability" have been achieved through the combination of several models, so called multi-model ensembling (e.g. Robertson et al. 2004). As will be shown below, since all models do not always share the same strengths and weaknesses, by combining them into a single probabilistic forecast the spatial coverage of positive skill increases, and negative skill is reduced. This improvement in skill has been shown explicitly to result from the increase in model number rather than just the increase in realizations (Palmer et al. 2000). Another important result of multi-model ensembling is the dramatic improvement in the reliability in probabilistic forecasts.

One implicit criterion for combining multiple models is that they all perform 'adequately'. If one model were found to be measurably worse than the others, it should

be dropped. In some cases, the combination algorithm considers past performance of the models, assigning weights accordingly (Rajagopalan et al. 2001). Unfortunately performance weighting requires long histories (40+ years) of model forecasts in order to determine relative model performance robustly. This becomes a problem for most CGCMs used for seasonal prediction because the observational data required for their initialization does not exist prior to the 1980s. With only 20+ years of retrospective forecast data, it becomes difficult to assign meaningful weights to individual models. Methodologies for synthetically extending a retrospective forecast history or for combining models that could circumvent the limited model history and still allow for performance weighting could greatly improve the skill of the resulting forecasts.

Forecast system validation

Several measures of forecast validation exist, sometimes giving a different picture of where, when, and which prediction practice yields the most accurate forecasts. In general the use of more than one measure of validation is desirable, and in Fig. 1 we have already shown skill based on the rank probability skill score. In this section we highlight additional measures that provide valuable information about US prediction skill. The first is the area under the relative operating characteristic (ROC) curve. For a particular grid point or region, these curves indicate the percentage of hits and false alarms yielded by the forecast system for a given event (e.g. above-normal tercile category), under varying levels of confidence in the forecast. If the event were perfectly predictable by the forecast system, it would have a hit rate of 1.0 and no false alarms. The area under the curve would be 1.0. If the system were unable to distinguish between a hit and a false alarm, those rates would be equal, and the area under the curve would be 0.5, which is considered the level of no skill. Negative skill is indicated by values less than 0.5. What is particularly useful about ROC areas is that they can indicate condition skill, for instance, higher hit rates for the upper tercile category than the lower one. An example is shown in Figure 2, which illustrates that forecasts of above-normal temperature have witnessed higher skill than below-normal temperatures for the Dec-Jan-Feb season during the 1981-2001 period. Figure 2 also illustrates some of the other points raised in the previous section regarding AGCMs, CGCMs and multi-model ensembles. In particular, there is little difference in skill of the AGCM versus CGCM practices, and the multi-model combination of all dynamical systems exhibits the greatest skill.

The second validation diagnostic of forecast performance we demonstrate is reliability. This measure is particularly important as it indicates the extent to which forecast probabilities mean what they say. A striking characteristic of all dynamical models is that their probability forecasts are over-confident (Figure 3). There is no distinction between AGCMs and CGCMs in this shortcoming. Some improvement can be achieved by recalibrating the probability distributions of the individual models (not shown). The greatest improvements are obtained by combining the models, here accomplished by simply averaging the 3-category probabilities of the 8 CGCMs and the 3 AGCMs. There is a negative consequence of such a process, namely that the sharpness of the resulting forecasts is reduced (i.e., fewer high probability forecasts are indicated). Ideally, one

wishes to retain as sharp as possible a forecast while ensuring reliability. Work continues towards this goal.

Outstanding questions and room for improvement

What are core activities for improving climate forecasting practices? Developing new models of the atmosphere-ocean-cryosphere-land system, ensuring sustained long term observations, enhancing data assimilation techniques, and improving understanding of seasonal climate variability are essential. A commonly used metric for measuring the impact of such activities is the skill and reliability of forecasts. In this report the skill attributes of existing and emerging dynamical methods of seasonal predictions have been examined.

A relevant question concerns whether U.S. seasonal prediction skill is advancing with newer generation models. Considerable investment has been devoted towards improving climate models, in part for the purpose of advancing seasonal predictions. Recent examples include new efforts to implement an updated global coupled forecast system with increased resolution and improved atmospheric and oceanic components at NCEP (to be called the Coupled Forecast System (CFS03)), with similar efforts underway at NASA/GMAO including their plan to use a global 1° resolution atmospheric model. An implicit assumption behind such efforts is that newer generation dynamical models will lead to improved skill. We know, for example, that predictability exists in the extra-tropical climate that the current generation of models are not realizing (Anderson et al. 1999). Analogies may also be drawn from weather forecasting experience where steady improvements in models and data assimilation techniques resulted in progressively improved weather predictions. It may be that the seasonal prediction models are presently neglecting some important external forcings, such as the increasing greenhouse gasses in the atmosphere, which can affect the characterization (and bias corrections) of the model climate over periods of years. Poorly represented interactions of the atmosphere with the land surface and with the cryosphere may also hamper the skill of seasonal predictions over the US. Another aspect of the climate system that is typically not well represented in the seasonal prediction models is the interaction between the stratosphere and troposphere (Baldwin and Dunkerton, 1999), which has demonstrated occasions of predictable evolution and subsequent influence on the terrestrial climate over the northern mid-latitudes. Even if the model development improves simulations of seasonal climate variability, seasonal prediction skill will nonetheless be limited by inherent signal-to-noise considerations. The relevant question becomes whether the new generation of dynamical models yield signal-to-noise ratios that more accurately reproduce those in nature. It is therefore important to continually document and analyze the seasonal prediction skill from the improved dynamical prediction systems, and to cast those performances within improved knowledge of predictability limits.

References:

Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Am. Meteor. Soc*., **80**, 1349-1359.

Baldwin, M. P. and T. J. Dunkerton, 1999: Propagation of the Artic Oscillation from the stratosphere to the troposphere. *J. Geophys. Res.*, **104**, 30937-30946.

Barnett, T. P., 1981: Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Wea. Rev*., **109**, 1021-1041.

Barnston, A.G. and T.M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate*, **9**, 2660-2697.

Doblas-Reyes, F. J., Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A*, **57**, 234-252.

Goddard, L. and M. Dilley, 2005: El Niño: Catastrophe or opportunity. *J. Climate*, **18**, 651-665.

Hastenrath, S., 1995: Recent advances in tropical climate prediction. *J. Climate*, **8**, 1519-1532.

Higgins, R. W., A. Leetmaa, Y. Xue, and A. Barnston, 2000: Dominant factors influencing the seasonal predictability of U.S. precipitation and surface air temperature. *J. Climate*, **13**, 3994-4017.

Horel, J.D., and J. M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the southern oscillation. *Mon. Wea. Rev.,* **109**, 813-829.

Huang, J, H. M.van den Dool, and A. G. Barnston, 1996: Long-lead seasonal temperature prediction using Optimal Climate Normals. *J. Climate*, **9**, 809-817.

Landman, W. A. and L. Goddard, 2002: Statistical recalibration of GCM forecasts over Southern Africa using model output statistics. *J. Climate*, **15**, 2038-2055.

Larkin, N. K. and D. E. Harrison, 2005: On the definition of El Niño and associated seasonal average U.S. weather anomalies. *Geophys. Res. Lett.*, **32**, L13705, doi:10.1029/2005GL022738.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000. A probability and decision-model analysis of PROVIST seasonal multi-model ensemble integrations. *QJRMS*, **126**, 2013-2033.

Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing sources of U.S. seasonal forecast skill. *J. Climate*, **19**, 3279-3293.

Rajagopalan B., U. Lall, and S. E. Zebiak, 2002: Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles. *Mon. Wea. Rev.,* **130**, 1792-1811.

Robertson, A.W., Zebiak, S.E., U. Lall, and L. Goddard, 2004: Optimal combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, 132: 2732-2744, DOI: 10.1175/MWR2818.1.

Tippett, M. K., Barlow, M. and Lyon, B., 2003: Statistical correction of Central Southwest Asia winter precipitation simulations. *Int. J. Climotol.*, 23, 1421-1433.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711-743.

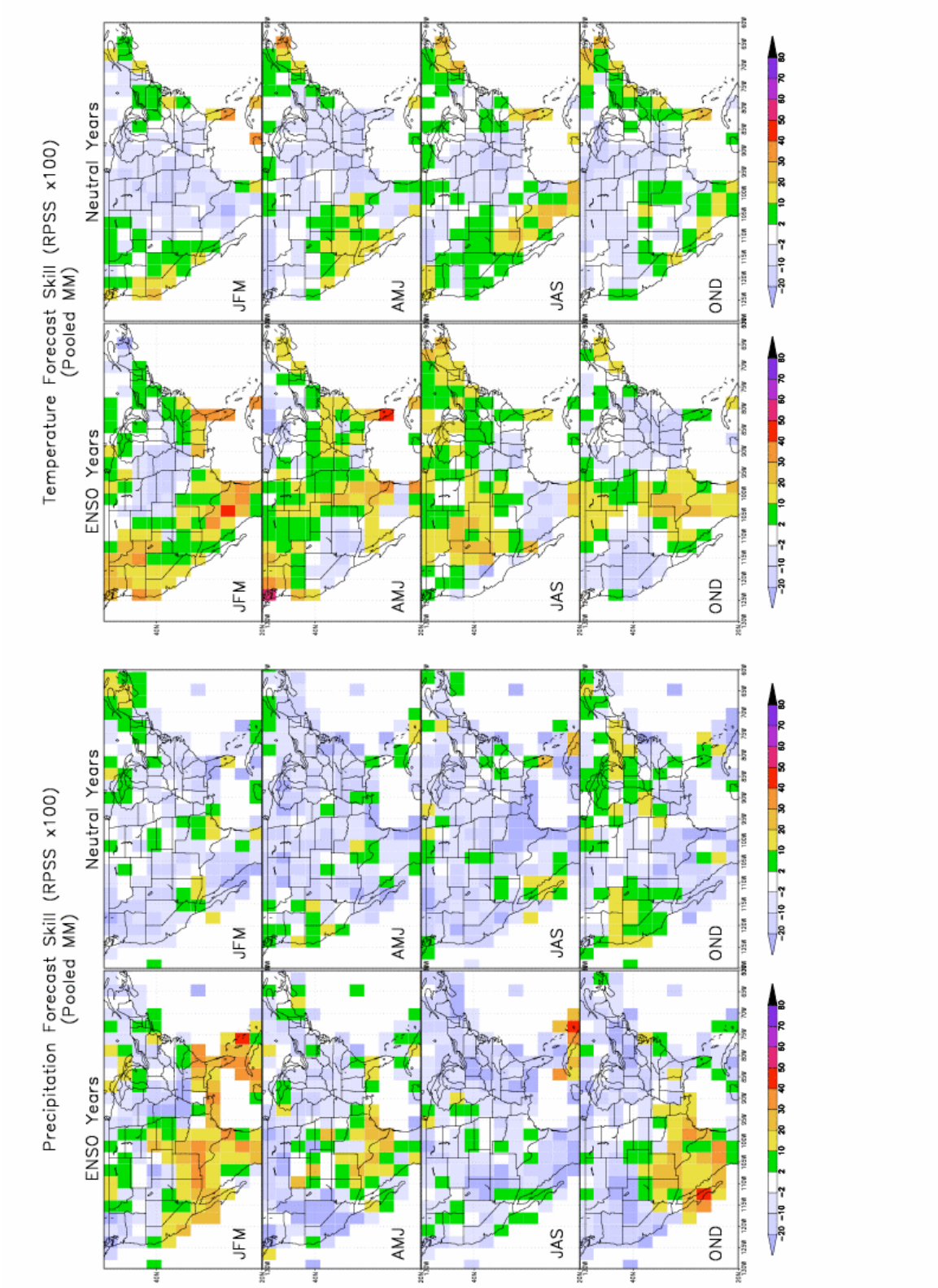

Figure 1. Differences in skill (RPSS) for 3-category seasonal rainfall forecasts between ENSO extremes and neutral conditions for the 1950-1995 period. Positive values indicate higher skill during ENSO extremes. (from Goddard and Dilley, 2005)

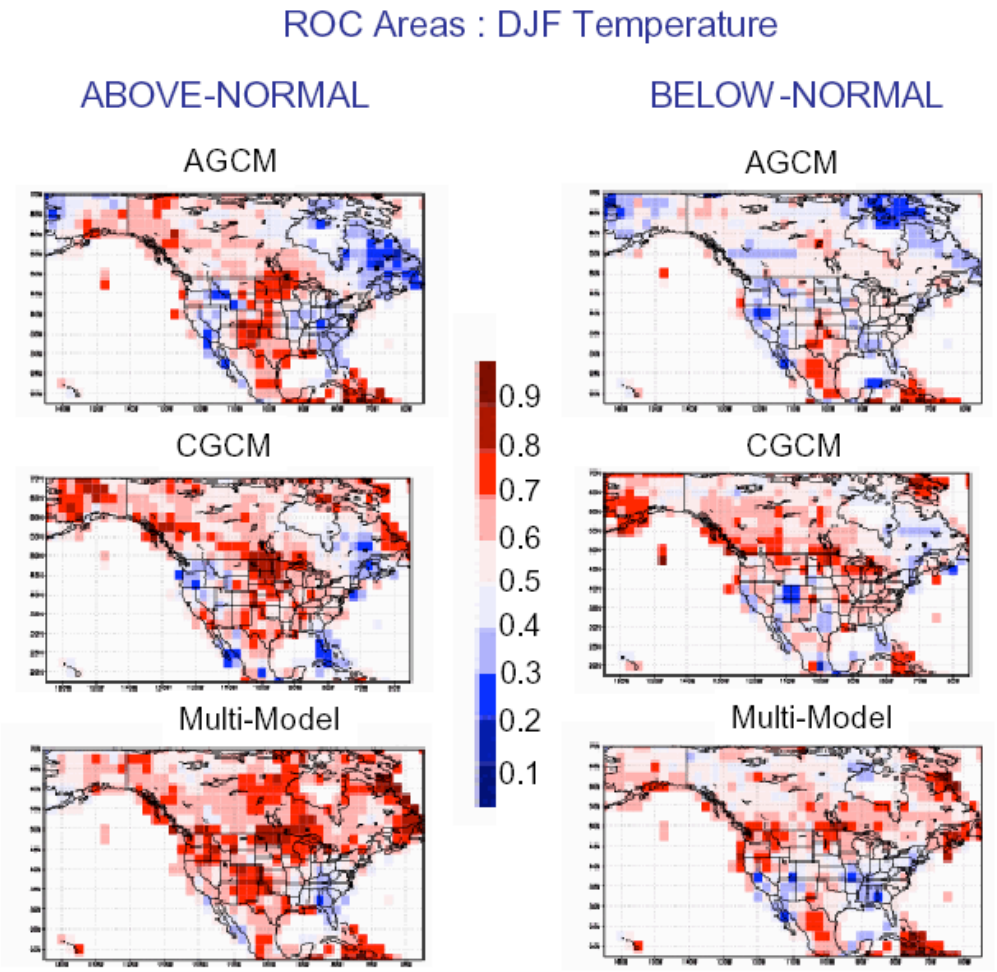

Figure 2. ROC areas for DJF temperature forecasts at 1-month lead for the period 1981-2001. The AGCM was forced by predicted SSTs. The multi-model forecast is based on equal weighting of 3-category probabilistic forecasts from 8 CGCMs and 3 AGCMs.

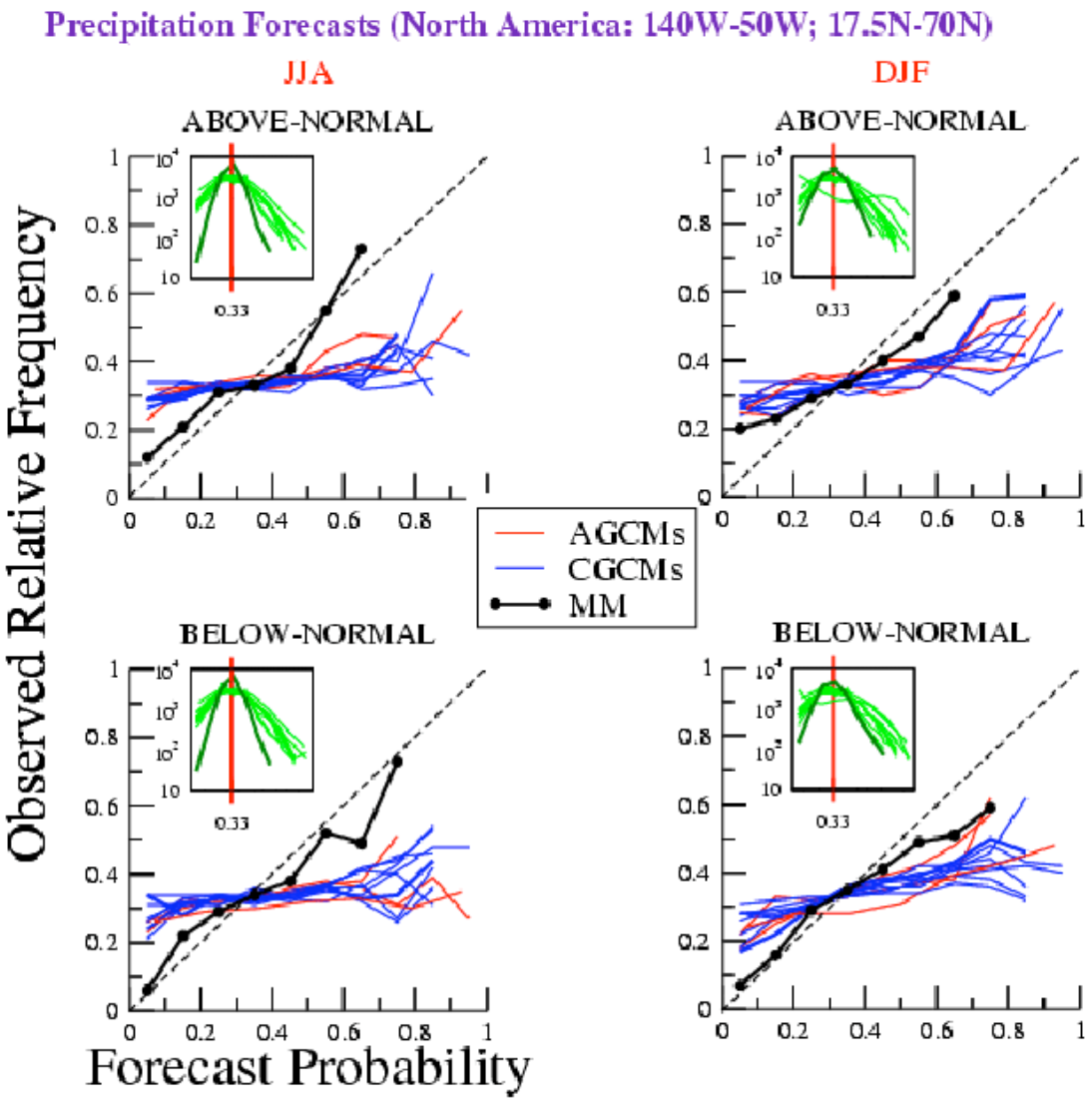

Figure 3. Reliability diagrams for the upper (Above-Normal) and lower (Below-Normal) category of 3-category forecasts for all terrestrial grid points over North America (140W-50W; 17.5N-70N). The colored lines show the reliability of the individual models; the light green lines in the inset boxes show the frequency with which forecasts of a certain confidence were issued for that category. The black line shows the reliability for the multi-model forecast, and the dark green line in the inset graph shows the confidence frequency.