Predictability of extreme precipitation and temperature months using the Met Office Seasonal Forecasting Atmosphere Global Circulation Model.

Ocean Applications Technical Note No. 29

A.W.Colman, P. Berrisford and M.K. Davey

September 2002

Produced by A.W.Colman
Met Office London Road Bracknell Berkshire RG12 2SZ United Kingdom.
Tel: +44 (0)1344 854509 Fax: +44 (0)1344 854499
Email: andrew.colman@metoffice.com www.metoffice.com

Predictability of extreme precipitation and temperature months using the Met Office Seasonal Forecasting Atmosphere Global Circulation Model

Andrew Colman$_1$, Paul Berrisford$_2$ and Mike Davey$_1$

$_1$Met Office OA-branch Seasonal Prediction group
$_2$ Reading University consultant

Abstract:

Trial seasonal-range forecasts of 850hPa temperature and surface rainfall, produced using the Met Office HadAM3 Atmosphere Global Circulation Model (AGCM), are verified against data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and from the Global Precipitation Climatology Project (GPCP). Predictability of noniles and terciles is assessed, focussing on the outer category 'extremes'. The extreme forecast skill is highest in the tropics and better for temperature than for precipitation. There is clear skill at 1 month lead times and marginal skill at longer lead times. For temperature, forecast skill in Europe is at its greatest north of about 45N, while for precipitation it is largest to the south west of the UK.

The skill of the HadAM3 model was confirmed by an assessment of the performance of long term (50 year) HadAM3 simulations forced with observed SST and the performance of persistence forecasts.

A probability forecasting system using discriminant analysis is described and assessed using Relative Operating Characteristic (ROC) scores.

This report describes exploratory research work carried out as a major part of the Corporate Investment Project CI107 on 'Predictability of Seasonal Extremes'. The work was contracted-out to Reading University, involving collaboration with the OA seasonal prediction group.

1. Introduction

It is extremes in climate that make the headlines. For example, the hot dry summer of 1995 or the persistent floods of Autumn 2000 remain in the public memory for many years. Extreme climate events like these impact on a wide variety of persons and can destroy livelihoods or even lives (Palutikof et al. 1997). In the light of this, The World Meteorological Organisation has made "Reducing Vulnerability to Weather and Extremes" its special theme for WMO day 2002 (www.wmo.ch). In addition to placing people and material objects in danger, extreme climatic events can place a large burden on public resources. For example, an extremely cold winter period drives up energy demand for heating, whereas a hot dry summer period can reduce the supply of water while increasing demand. The public utilities can find it very difficult supplying these demands, particularly when they have no prior knowledge of the event. Persistent anticyclonic conditions in wintertime can create pollution problems so that, in particular, people with chronic respiratory diseases such as asthma may have breathing difficulties. This in turn can overload hospital resources. For these reasons and many more, it is obvious that prediction of extreme events is highly desirable.

Studies of predictability of seasonal or longer timescale extremes have been largely confined to investigating El Nino and La Nina related teleconnections and climate change scenarios. Roplelewski and Halpert (1987,1989), Kiladis (1989) and others describe extreme anomalies related to ENSO. The prospect of long term changes in the nature and frequency of extremes in the 21st Century due to man made pollution and other causes using information from coupled models is discussed in the latest IPCC report (Houghton, 2001). Climatological mean annual extremes of temperature are quite well simulated by the models, whereas climatological means of extreme daily rainfall are somewhat less well simulated. Kharin and

Zwiers (2000) provide a relatively detailed assessment of the Canadian Global Coupled Model including its performance in predicting the geographical distribution of these extremes. Kiktev et al (2002) discuss how well changes in extremes observed over the last 50 years are simulated by long term runs of the HADAM3 model forced with observed SST and observed levels of greenhouse gases and aerosols. Palmer and Raisanen (2002) describe how the Global Coupled models used for the CMIP2 (???) exercise might be used to produce economically useful forecasts of extremes.

Seasonal forecasting is a young and rapidly developing science, and to date most predictability assessments have been in terms of, for example, above/below normal seasonal averages. As forecasting capacity advances, particularly with regard to atmosphere GCMs and the prospects for
large-ensemble multi-model systems, it is timely to explore the predictability of extremes.

The aim of this work is to make an initial assessment of such predictive skill for the Met Office system currently being used to produce The Monthly Outlook and regular seasonal forecasts, which is based on the HadAM3 atmospheric GCM. In doing so it has been necessary to devise a strategy for assessment, as there are no precedents. This assessment has focussed on monthly timescales, and temperature and rainfall variables, on the spatial scale of the AGCM grid. In the time available for this project (Nov 2001-Mar 2002), attention has been limited to deterministic forecasts.

In section 2, we describe the data used in this study and in section 3 we describe the method of assessing predictability. The results of the assessments of the SEMIC hindcasts are described in section 4, while in section 5 these results are evaluated and compared against persistence skill and the skill of long-term HadAM3 simulations. In Section 6 probability forecasts are described and assessed. Finally, in section 7, we present some conclusions and recommendations.

2. Data

In this study we confine ourselves to looking at global fields of monthly mean temperature and precipitation. We compare ensemble mean, monthly mean retrospective forecast data from the Met Office SEMIC (Seasonal Ensemble integrations at Monthly Intervals for model Climate and skill assessment) Project with monthly mean analysed 850hPa temperature from The European Centre for Medium-Range Weather Forecasts (ECMWF) and precipitation data from the Global Precipitation Climatology Project (GPCP) created by the NASA Goddard Space Flight Centre, USA. We also compare long term simulations of the Met Office HadAM3 model with gridded observed surface temperature and rainfall data provided by the Climate Research Unit (CRU) University of East Anglia, UK.

2.1 SEMIC forecasts

In the SEMIC Project, trial ensemble forecasts were produced using the Met Office AGCM for the period 1979 to 2000. The SEMIC forecasts start on the first day of each month and are run out to a range of 6 months. The model used is the 19 level version of the Met Office HadAM3 model. The grid resolution is 2.5 degrees North-South and 3.75 degrees East-West. The forecast ensembles consist of 9 members initialised respectively with each of the 6 hourly atmospheric analyses available within the 48 hours prior to the first day of the forecast. The sea surface temperature anomalies required for running this atmosphere only forecast model are persisted from the beginning of the forecast. In this work, we have SEMIC data available from forecasts with start dates in September to December for the 18 years 1982 to 1999, and with January and February start dates for the 18 years 1983 to 2000.

2.2 Long-term HadAM3 SST forced simulations

A disadvantage of the SEMIC data is that the assessment period (1982-2000) is rather short, particularly if one is assessing the predictability of extremes. For each forecast run, there are only two years in each nonile category (see below). We have tried to overcome this problem by repeating some of the assessments using longer datasets from climate simulations with the HadAM3 model. A 10 member ensemble of long term runs is available for the period 1948 to 1999 inclusive. The model is forced with observed sea surface temperatures from the Met Office HadISST1 dataset. However, unlike the SEMIC project, these are simulations rather than forecasts, and no initial-state atmospheric information is used.

2.3 Verification Data

The SEMIC 850hPa temperature forecasts are verified using ECMWF analyses. Where possible (i.e. from 1979 to 1993), data from the first ECMWF re-analysis (ERA15) is used, otherwise operational data is used. The ECMWF data were extracted from their database and interpolated to the Met Office Model grid. The SEMIC precipitation forecasts are verified using GPCP data, which were also interpolated to the Met Office model grid from a 2.5 by 2.5 degree grid. The monthly mean data required for verifying the forecast data cover the periods September to December, 1982 to 1999 and January to July, 1983 to 2000.

Neither ECMWF nor GPCP data is available prior to 1979 hence other data sources were required to verify the long-term HadAM3 simulations. Monthly gridded rainfall data from Hulme (1994) produced at CRU is used to verify the rainfall simulations from 1949-1998. Unlike GPCP data, these are not complete but do have a good land coverage. A gridded surface temperature database also produced at CRU by New is used to verify the temperature predictions. The temperature database was interpolated from a 0.5 degree square grid to the Met Office model grid (2.5 lat by 3.75 long) to verify the model simulations. These temperature data are nearly complete over land and available for 1948 to 1998.

3. Assessment method

A simple deterministic skill measure was derived to gain an idea of skill dependency on time of year, forecast lead, geographical location and forecast variable. For each of the 18 years considered we have 6 start dates, and forecasts to 6-month range (providing 6 ensemble-mean monthly-mean forecast values), yielding 36 verifying time/lead time combinations for both temperature and precipitation. At each grid point on the globe the forecast data in each of these 36 combinations is ranked from the coldest (driest) to the warmest (wettest) year. A similar ranking is performed for the verifying months in the analysed data. The rankings are split into 9 [3] bins (which we call noniles [terciles]), which contain 2 [6] year labels (identifying the years for which the data fall into the various bins).

The extreme forecast skill is then defined as the number of years common to the (outer) bin in question (e.g. the upper or lower nonile bin) in both the forecast and verifying data, regardless of the order of the years within the bin, divided by the total number of labels in the bin i.e. 2 for noniles or 6 for terciles. For example, consider a nonile bin of either extreme: if there is one year in common between the forecast and analysed data regardless of position within the bin then the (upper or lower) extreme nonile forecast skill is 0.5. If both year labels are the same for the forecast and observation then the skill is 1. If no years are common then skill is 0. It must be noted that a forecast skill of 1, as defined here, does not necessarily imply a perfectly ranked forecast, as the ranking of years within a bin is not considered.

4. Results (Assessment of SEMIC output)

The results will concentrate on the nonile extremes in four regions of the globe (Fig. 1): the entire globe; Europe (11.25W to 26.25E and 35N to 62.5N); North America (123.75W to 60W and 22.5N to 55.0N); and the tropics (20N to 20S). Ocean and land areas are both included. The results for the 36 verifying time/lead time combinations have been averaged in various ways to enable the results to be plotted as a function of space, lead-time, verifying month, forecast start date and time. When looking at the results it should be borne in mind that the random chance score for a nonile extreme is 1/9 (0.111) and for a tercile extreme it is 1/3 (0.333).

4.1 Mean skill of different variables

Table 1 summarises the results by presenting the average skill of the 36 verifying time/lead time combinations for temperature and precipitation for the nonile and tercile extremes in the four regions of the globe defined above. The average skill for both parameters for all the regions is greater than that expected from random chance. The extreme skill for temperature is greater than for precipitation in all four regions. Out of the 4 regions chosen for study, skill is highest in the tropics and lowest in Europe. The highest extreme nonile forecast skill is for tropical temperature where the value of 0.212 is nearly double that for random chance. Given the sample sizes are all quite large (at least 132 grid points x 18 years), all these average skill scores are significantly greater than chance (according to a binomial test for significance). Fig.2 shows the extreme nonile mean skill for temperature and precipitation over the whole globe. It is evident that skill is largest in the tropics and is greater for temperature than for precipitation. For temperature there is considerable skill (above 0.15) throughout much of the Pacific and Atlantic Oceans between 70S and 50N. The skill for precipitation, on the other hand, is much more confined to the tropical Pacific region.

4.2 Predictability v Lead-time

Fig.3 depicts the extreme nonile forecast skill as a function of forecast lead-time, in months, for temperature and precipitation in the four regions of the globe. It can clearly be seen that, for the most part, the skill level is superior to that gleaned from random chance. Furthermore, for both temperature and precipitation in all four regions, the skill at a lead-time of 1 month is greater than at other lead-times. For temperature, this skill is approximately 2.5 times that of random chance, whilst for precipitation it is nearly twice that for random chance. In addition, the skill for temperature over the entire globe, North America and the tropics remains above that for random chance at all lead times out to six months. In Europe the skill for temperature is above that of chance at lead times up to 5 months. In agreement with the previous sub-section, the skill in the tropics is nearly always greater than that elsewhere, whilst that in Europe is generally lowest.

4.3 Predictability v verifying month (month being forecast)

In the previous sub-section we saw that skill at a lead-time of 1 month is markedly greater than at other lead-times. For this reason, and given that we have forecasts commencing in only 6 months of the year from September to February thereby yielding an incomplete set of forecasts for verifying month against lead-time, it is not sensible to average results for various lead-times at particular verifying months. In the light of this, we show in Fig.4 the dependence of extreme nonile forecast skill on verifying month for a lead-time of 1 month only. Bearing in mind that these averages are constructed from one month only, so may not be as significant as other results presented in this section, Fig.4 shows that forecast skill is greater than chance for all months (September to January). For both temperature and precipitation, skill for North America is highest in November. The European skill for both temperature and precipitation is relatively low in October and December and generally high in January and February. The skill for the globe and the tropics varies only slightly with the month. However, for the global temperature at least, skill is greatest in wintertime (December, January and February). Again, the skill in the tropics is mostly greater than that elsewhere, while that in Europe is generally lowest, except in January and February when the European skill is as high, if not higher, than that in the tropics.

4.4 Predictability v forecast start date (date of start of the SEMIC run)

Fig.5 shows the dependence of extreme nonile forecast skill on the start month of the forecast. There appears to be no great dependence here. Again, skill in the tropics is generally greater than elsewhere, while that in Europe is generally lowest. European forecasts, for both temperature and precipitation, commencing in October had marginally lower skill than other start months.

4.5 Variation of predictability with time

Fig.6 shows the average number per month of global extreme nonile observations and correct (bin) forecasts for temperature and precipitation for the years 1982 to 2000. The trend in the observations of precipitation, where the number of extremes increases with time, is only hinted at in the forecasts. There is no discernible trend for temperature, in either the observations or the forecasts. Of the 11 years of above average numbers of extremes in temperature, 5 were captured by the forecasts, while the 8 years below average were captured by the forecasts in 6 of the years.

4.6 European skill

Fig.7 illustrates the geographical distribution of extreme nonile skill in the European sector for temperature and precipitation. For temperature, forecast skill is generally greatest in mid to Northern Europe with values for the lower extreme in excess of 0.2 over much of northern Europe from the Greenwich Meridian eastwards. For the upper extreme the skill approaches 0.2 over a similar longitudinal region, but confined to a smaller latitudinal range of about 5 degrees centred on 50N. For the upper extreme of temperature the skill over the UK is relatively poor, mostly being near the random chance level except for the far south, while for the lower extreme the UK skill is mostly above 0.125. The European skill is lower for precipitation than for temperature. As is the case for temperature, there appears to be more skill for precipitation at the lower extreme than the upper one. The most skilful area for the lower extreme is over and to the south west of the UK where values are mainly between 0.15 and 0.2. The forecast skill for precipitation over the UK is relatively poor in the upper extreme, as it is for temperature, where values are near the chance level. On a more local level, the forecast performance at one location can be illustrated by means of scatterplots. For example, Fig.8 displays such a plot for the temperature forecasts for January at a lead-time of 1 month at the location (0E, 52.5N), which is the grid point in the south east of the UK. It is clear that there is some skill in the forecasts as the points are distributed around the perfect forecast line. In this example we see that the lower nonile extreme would have a forecast skill of 1, as the two coldest years (1985 and 1987) are correctly forecast to be so, while the upper nonile extreme would have a forecast skill of 0.5 as the second warmest year (1989) is forecast to be in the correct category. In absolute terms the forecasts in the lower extreme are near perfect whereas those in the upper extreme are not so good.

5. Evaluation of Results

5.1 Assessments of HadAM3 climate runs

5.1.1 Skill v simulated month

Plots of mean simulation skill in predicting noniles for each calendar month are presented in Fig. 9. Skill is always above the chance level and the simulation skill for temperature is substantially higher than the skill for precipitation. The relatively high skill for temperature is not surprising given that sea surface temperature is specified. The skill for temperature is particularly high in tropical coastal regions. Variability between months is quite low particularly for the globe. In Europe, warm spring months and cold autumn months seem to be slightly more predictable than other extreme months.

5.1.2 Simulation skill time series 1949-1998

The number of correct extreme months simulated each year as a proportion of all simulations for the 50 years is displayed in Fig. 10. The number of extremes observed each year is also plotted

as the number of correct simulations will be dependent on this. The chance simulation skill is 2/(9x9) (=0.024) and the chance probability of an extreme of either sign is 2/9 (=0.22). The skill for temperature and rainfall is generally just above this level. Spikes follow the 1982-3 and the 1997-8 El Ninos but not other El Nino events.

5.2 Persistence skill

A common way of assessing seasonal forecasts is to compare against persistence. By assuming that the forthcoming month's rainfall or temperature anomaly will be the same as that for the preceding month, we can compare the resulting forecast with the AGCM prediction. Here, a persistence forecast is defined such that the extreme noniles contain the same year labels as those for the preceding month.

Fig. 11 is similar to Fig. 4 but is for persistence forecasts. Skill is clearly present but is not quite as high as the SEMIC skill for Europe or the Globe. Over the longer (1949-1998) period (Fig. 12), temperature and rainfall persistence skill is substantially higher than the AGCM simulation skill. This suggests that knowledge of the preceding month's atmospheric anomalies may be more important than knowledge of the SST anomalies in predicting extreme seasons.

5.3 Predictability of notable extremes over the UK

Table 2 contains a list of noteworthy climate extremes observed in the UK between 1982 and 2000. The extremes listed are months or seasons with an extreme Central England Temperature (CET) or England and Wales Precipitation (EWP) anomaly. Listed also is the position of the extreme season in the climatological probability distribution function (PDF) for that season. The criteria for selecting the dry, wet and cold extreme months and seasons was that they needed to be within the 5% most extreme (by rank) in the climatological record. For warm months and seasons a tighter criterion of 2.5% was set as in these recent years there has been a high occurrence of extreme warm seasons, probably due to global warming.

Alongside observed PDF positions are the locations of the SEMIC forecasts on the forecast PDF for the nearest model gridpoint to Central England (52.5N, 0W). The forecast PDF is evaluated from the model data, in this case the SEMIC years, excluding the year being predicted. Using the model PDF, rather than the observed PDF, prevents the forecast skill being affected by model bias. PDF positions for 1 to 6 month lead forecasts are shown where available. Note that at the time of writing there are no SEMIC runs starting in June, July or August so there are some gaps in Table 2. The performance of the forecasts is quite mixed but generally rather disappointing. In only two of the 27 cases (the extreme warm months of JFM 1990 and September 1999) is the correct anomaly sign consistently predicted at different lead times and in these cases there are only 3 forecasts. Also, In most cases where the correct sign is predicted, the prediction is in the a middle quartile (between 0.25 and 0.75) not indicating an extreme. In only a few (5 1 month lead cases and 1 2 month lead) cases are correct outer quartile predictions made. This suggest the AGCM is not good at picking up extreme signals.

LEPS (Linear Error in Probability Space, Potts et al 1996) is a skill measure assessing how near the forecast and verifying observation is on the climate PDF. If P_o is the position of the observation on the observed climatological PDF (somewhere between 0 and 1) and P_f is the position of the corresponding forecast on its PDF, then the LEPS score S is defined as

$$S=3(1 - |P_f - P_o| + P_f^2 - P_f + P_o^2 - P_o) - 1$$

LEPS skill can be evaluated from one or more LEPS scores

Skill = sum(S)/sum(S_{max}) if sum(S) >0 otherwise skill = sum(S)/(1-sum(S_{min}))

Where S_{max} is the highest LEPS score possible given P_o (i.e. when P_o= P_f) and S_{min} is the lowest LEPS score possible given P_o

The LEPS skill of these forecasts is presented in Table 3. LEPS is quite a harsh measure for assessing extreme predictions: if the forecast is close to the median and the observation is an extreme then a negative score will result. Only the 1 month lead forecasts have any skill according to the LEPS scores. The skill of the extreme cold month predictions at 1 month lead stands out. This reflects successful predictions of the cold in Feb 1986 and Jan 1987(Table 2b).The sample of cold UK winter months is small however and the skill is contributed to by the high climatological persistence of circulation patterns at this time of year. There is not much skill at longer lead times except from the September forecast.

5.4. Discussion & implications of SEMIC assessments

The main result from this work is that the extreme forecast skill from the SEMIC runs is greater than that that would be achieved from random chance or persistence. In fact, in the case of the average for the tropical temperature, the skill level is almost double that of chance.

The fact that forecast skill decreases with lead-time is not a great surprise. The model will obviously be closer to reality in the early stages of the forecast run. Skill levels for forecasts with a 1 month lead-time are about 2.5 times that of random chance. (NB here a lead-time of 1 month refers to the first month after the forecast start.) This skill does not appear so good when compared against persistence, which is around 80 per cent of model skill or twice that of random chance. According to the LEPS scores for notable extremes, only the 1 month lead-time forecasts have any skill.

We have seen that extreme forecast skill for temperature is greater than that for precipitation. This is not surprising, as temperature is one of the main prognostic variables in the model and is also an analysed variable in the data assimilation system, whereas precipitation is neither of these and hence maybe more subject to parameterisation error. Precipitation anomalies are generally more localised than temperature anomalies which makes it more difficult to estimate mean precipitation than mean temperature for a relatively large area around a global model grid point.

The dependence of forecast skill on the verifying month is not clear. For the European sector at least, there are indications that the skill for both temperature and precipitation is greater in January and February than in the other months. However, as already noted above, the significance of this result is open to question as we only include the month 1 forecasts in the averages. Whilst the month 1 forecasts are the most accurate they only provide one sample forecast (for each grid point) per month. There appears to be no clear dependence of skill on the month when the forecast commences. The assessments of the HadAM3 simulations do not show a strong dependence on the month of year and no evidence of the higher European skill in January and February as indicated in the SEMIC results.

As expected, out of the four regions we chose to study, the European sector generally has the lowest level of skill, except perhaps for the verifying months of January and February. This is not to say, however, that the European skill is lower than that in other regions of the mid latitudes. Within the European sector, the skill for the UK for the upper extremes of both temperature and precipitation is quite poor, being at or below the chance level. This is offset to some extent by the better skill at the lower extremes. It is not known why the forecast skill at the lower extreme is superior to that at the upper extreme.

6. Probability forecasts

Limited forecast accuracy and the chaotic nature of the climate system mean that it is not feasible to provide deterministic forecasts such as "next month will be extremely warm". A better way of expressing forecasts is in the form of probabilities, e.g. "there is a 20% chance of a warm extreme (nonile) next month, which is substantially higher than the chance probability of 11%".

Fig. 13 is an example probability forecast for March 2000 precipitation based on the February 1 SEMIC forecast output. Discriminant analysis (Afifi and Azen, 1979) is the statistical method used to predict the probability of the dry or wet nonile. Discriminant equations are used to produce probability forecasts for an event in a similar way to which regression equations are used to produce a point forecast. An estimate of the PDF of the predicted event as a function of predictor value is estimated from historical data and assumed to be normal in shape. In this case the event is the occurrence of an extreme (nonile) season and the predictor is the raw model forecast. The discriminant equations are calculated from a historical database of model hindcasts (the SEMIC hindcasts) and observed precipitation categories for a specified training period, 1982 to 1999 in this example. The training data in this example consists of 18 years each with 9 ensemble members, a total of 162 trial forecast cases to evaluate the equations. The 9 SEMIC model ensemble forecasts for 2000 are substituted into the discriminant equation to produce 9 probability forecasts. The probabilities are averaged to produce an overall forecast.

These probability forecasts have the advantage that they reflect the skill of the forecast system. The probabilities for an area with no track record of skill would be adjusted to be close to chance (0.111 for noniles).

Over many tropical areas the probabilities of an extreme nonile deviate substantially from chance (0.111). For example the probability of a wet nonile category is less than 0.05 over much of the tropical East Pacific. Over parts of SW Europe a dry nonile is substantially more likely to occur than by chance, but elsewhere in Europe probabilities are quite close to chance.

6.1 Assessments of Probability forecasts

The probability forecasts have been assessed using ROC (Relative Operating Characteristic) assessments. ROC is a WMO standard measure for assessing probability forecasts which is described by Stanski et al., (1989). ROC assesses the prediction of an event. To evaluate a ROC score, one categorises a series of verified forecasts into 4 categories:

H= Hits = Number of times event is predicted and occurs
M= Misses = Number of times an event is not predicted but does occur
FA = False Alarms = Number of times the event is predicted but does not occur
CR = Correct Rejections = Number of times the event does not occur and is not predicted

The total number of forecasts N = H+M+FA+CR

From this one can calculate the Hit Rate (HR = H/(H+M)) and the False Alarm Rate (FAR= FA/(FA+CR)).

If HR is greater than FAR then the forecast is skilful. ROC can be used to assess probability forecasts by setting probability thresholds. If the forecast probability exceeds the threshold, then the event is deemed to be predicted. To assess a probability forecasting system, a set of Hit Rates and False Alarm rates are calculated for a set of probability thresholds. In this study, probability thresholds are set at 11 10% intervals (i.e. 0.0,0.1,0.2,0.3 ...1.0). If one plots all the hit rates (y axis) against its false alarm rates pair (x axis) and joins up the plots, then the result is a ROC curve. The area under this curve is a general measure of skill and is sometimes called the ROC score (ROC).

If HR_n is the Hit rate for probability threshold 0.n and FAR_n is the False Alarm Rate for probability threshold 0.n

ROC = sum n=0 to 9 $(FAR_n-FAR_{n+1})(HR_n+HR_{n+1})/2$

The chance ROC score is 0.5, perfect is 1.0 and lowest is 0.0. The SEMIC predictions have been evaluated using the Jackknife method as follows. 18 of the 19 SEMIC years (1982-2000) are used to derive discriminant equations which enable one to predict the probability of an event from the model output. The discriminant equation is used to evaluate probabilities for the 19th year. This process is repeated with each of the 19 years to produce a full set of probabilities. ROC scores are evaluated from these probabilities.

Monte Carlo tests were used to determine the 95% significance level of the ROC scores. Sets of probabilities for the 19 years were selected at random and verified against random observations to produce 500 000 ROC scores. Means of random ROC scores for the same number of scores as is used to evaluate European and global averages (132 and 7008 respectively) were evaluated which consequently provide estimates of what the mean ROC scores for Europe and the globe would be if the model had no skill. The ROC averages were sorted and the 95% level ROC scores are listed in Table 4.

ROC scores have been evaluated for Dry and Wet extreme noniles of monthly and 3 monthly mean rainfall for each model grid point. The ROC scores have been averaged over the globe and over the Europe region (13W-28E, 34-64N). ROC scores are plotted against forecast month in Fig. 14. Each line represents a different lead-time. Forecasts of individual monthly means are plotted in Fig. 14a. There is clearly skill at the 1 month lead time for Europe and the globe but at longer leads there is only very slight skill for the globe and none for Europe.

Forecasts of 3 month means are potentially more skilful as noise is filtered out (Fig. 14b). Tropical seasonal rainfall has been found to be more predictable on the 3 month scale than on the 1 month timescale (e.g. Folland et al, 2001). Global mean skills are nearly all just above the 0.5 chance level and above the 95% significance level (0.502) with less variability between months than in the 1 month case. European forecasts starting between January and March have skill on this timescale but forecasts from earlier months fail to beat chance.

An indication of the geographical distribution of skill is shown in Fig. 15. The mean skill of all the 3 month mean forecasts with start dates from September though to May and lead times of 1,2,3 and 4 months is presented. Skill is clearly strong (extended areas above the 95% significance level of 0.54) for the Tropical Pacific and around Indonesia and for a few smaller tropical locations including NE Brazil, parts of the Caribbean, near the Guinea coast of East Africa and parts of SW Asia. Elsewhere including all extra-tropical regions, skill is patchy or below chance. The skill patterns for dry and wet extremes are quite similar but with higher skill for wet extremes in the tropical Pacific which is consistent with the sign skill results in fig. 2.

7. Conclusions and Recommendations

We have demonstrated that for the SEMIC forecasts there is extreme (outer) nonile forecast skill for 850hPa temperature and precipitation. Although this skill is highest in the tropics, the average skill over the entire globe for both parameters is still above the chance level, even at a forecast lead-time of 6 months. The skill is higher for temperature than for precipitation and is greatest at a forecast lead-time of 1 month. Unfortunately, the skill appears to be only significant at a lead-time of 1 month. Even though the skill in the European sector is low compared with the other sectors chosen here, there is still considerable skill for temperature in mid to North Europe. The skill for most of the UK, apart from in the south east, is relatively poor for temperature. By contrast, the skill for precipitation in the European sector is not far above chance level, though it is relatively high to the south west of the UK.

The results indicate that the AGCM seasonal forecast system has the potential to provide 'extreme' forecast products. At this stage it is not known whether the levels of skill are sufficient to attract commercial interest, particularly within Europe. The levels of skill in tropical regions suggest that a global-coverage seasonal extremes product should be developed (under the Government Meteorological Research programme) for the purpose of advising government departments and of linking to the environmental stresses area of work being developed within the Met Office.

Some considerations for future work:

- There is reason to expect that extreme skill will be higher for multiple-month averages, particularly at longer lead times, and this aspect should be investigated. Using calendar month averages may be convenient but can miss anomalies that straddle two months. Future work could try defining extremes from anomalies lasting more than a specified time, rather than using seasonal means.

- Extreme forecast skill is highest in the first month of the prediction. This is probably due to the information contained in the initial atmospheric conditions of the forecast. Future skill assessments should be compared against forecasts made by persisting the initial atmospheric conditions of the forecast for the first few days.

- The ability to assess extreme forecast skill is limited by the relatively small number of years of retrospective forecasts available. This situation will be improved by the availability of data from the EU DEMETER project that will cover 30 years.

- A framework for the assessment of extremes has been established in this project. This framework can be used as a basis for further assessments, e.g. of the coupled GCM data from the Met Office and ECMWF seasonal forecast systems, of multi-model systems, and of the 1-month range higher-resolution ECMWF system. Assessments could also be made of forecasts using various statistical methods and from other centre's GCMs (e.g. IRI or NECEP forecasts).

REFERENCES

Folland, C.K., Colman, A.W., Rowell, D.P & Davey M.K. Predictability of northeast Brazil rainfall and real-time forecast skill, 1987-98. J.Climate 14 1937-1958 (2001).

Houghton, J.T. Climate change 2001: the scientific basis. IPCC report Cambridge Univ. Press (2001)

Hulme, M., 1994: Validation of large-scale precipitation fields in general circulation models. In: *Global Precipitation and Climate Change,* Eds M. Desbois and F. Desalmand. NATO Advanced Research Workshop, Springer-Verlag.

Kharin, V.V and Zwiers, F.W. Changes in the Extremes in an Ensemble of Transient Climate Simulations with a coupled Atmosphere-Ocean GCM J.Climate **13** 3760-3788 (2000)

Kiktev, D., Sexton, D., Alexander, L., and Folland, C.K. Comparison of modelled and observed trends in indices of daily extremes. Submitted to J Climate (2002)

Kiladis, G.N. & Diaz, H.F. Global climatic anomalies associated with extremes in the Southern oscillation. J.Climate **2** 1069-1090 (1989)

Palmer, T.N. and Raisanen, J. "Quantifying the risk of extreme seasonal precipitation events in a changing climate" Nature **415**, 512 - 514 (2002)

Palutikof , J.P. , Subak,S. and Agnew, M.D. Economic impacts of the hot summer and unusually warm year of 1995. Dept. Environment Report, Univ E Anglia (1997)

Ropelewski, C.F and Halpert, M.S. “Global and regional scale precipitation patterns associated with the El Nino/Southern Oscillation.” Mon Weather Rev **115** 1606-1626 (1987)

Ropelewski, C.F and Halpert, M.S. “Precipitation patterns associated with the high index phase of the Southern Oscillation.” J.Climate **2** 268-284 (1989)

Stanski, H.R., Wilson, L.J. and Burrows, W.R. “Survey of common verification methods in Meteorology” World Weather Watch Technical Report No * WMO/TD 358, WMO, Geneva, Switzerland.(1989)

Figure Captions

Fig.1 Map of regions being assessed.

Fig.2 Extreme nonile mean skill for 850hPa temperature and precipitation over the globe.

Fig.3 The dependence of extreme nonile mean skill on the forecast lead-time in months for a) 850hPa temperature and b) precipitation, over three regions of the globe:

Thick line: the entire globe
Solid line: Europe
Dashed line: North America
Dotted line: the tropics
Green line: random chance

Fig.4 As Fig.3 but showing the dependence of skill on the verifying month.

Fig.5 As Fig.3 but showing the dependence of skill on the forecast start month.

Fig.6 The average number per month of extreme nonile observations (solid lines) and forecasts (dashed lines) for 850 hPa temperature and precipitation.

Fig.7 The upper and lower extreme nonile mean skill in the European sector for a) 850hPa temperature and b) precipitation.

Fig.8 Scatter plot for the month 1 January forecasts for 850hPa temperature at a grid point in the South East of the UK (0E, 52.5N). The dashed lines indicate the nonile extreme limits and the dotted lines indicate the tercile extreme limits.

Fig.9 Dependence on simulated calendar month of extreme nonile mean skill of SST forced longterm HadAM3 simulations for 850hPa temperature and precipitation over Europe and the globe.

Fig.10 Dependence on simulated year of extreme nonile mean skill of SST forced longterm HadAM3 simulations for 850hPa temperature and precipitation over Europe and the globe.

Fig.11 Extreme nonile mean skill of persisting previous months anomaly of surface temperature (1981-1998) and precipitation (1982-1999) over (a) the globe and (b) Europe.

Fig.12. Same as Fig 10 but for 1949-1998 for both variables.

Fig.13 Example probability forecast for extreme nonile precipitation in March 2000 from SEMIC ensemble forecast. Discriminant equations are calculated using SEMIC forecasts for 1982-1999.

Fig 14 Dependence on forecast start month of ROC skill of precipitation probability forecasts from SEMIC output 1982-2000. Skill in predicting 1 month means (a) and 3 month means (b)

Fig 15. ROC skill of precipitation probability forecasts from SEMIC for March from early February
Forecast (2 month lead).

Tables

Table 1. Average Forecast Skill

	Globe	Europe	North America	The Tropics	Random Chance
T nonile	0.185	0.150	0.186	0.212	0.111
P nonile	0.157	0.126	0.153	0.186	0.111
T tercile	0.412	0.357	0.409	0.446	0.333
P tercile	0.380	0.350	0.385	0.408	0.333

Table 2 PERFORMANCE OF SEMIC HINDCASTS IN PREDICTING EXCEPTIONAL SEASONS EVALUATED BY THE POSITION OF FORECAST AND OBSERVATION ON THE MODEL AND OBSERVED PROBABILITY DISTRIBUTION FUNCTION.

Table 2a EXCEPTIONAL WARM SEASONS

	Month	Year	Obs	PDF for n month lead forecast (52.5N,0E).					
			PDF	n=1	2	3	4	5	6
	July	1983	0.997	-	-	0.428	0.448	0.409	0.582
	JFM	1990	0.997	0.641	0.715	0.273	0.616	-	-
	Aug	1990	0.997	-	-	-	0.672	0.519	0.614
	Nov	1994	0.997	0.876	0.327	0.387	-	-	-
	JulAug	1995	0.997			0.452	0.491	0.539	
	Oct	1995	0.991	0.452	0.700				0.456
	Mar	1997	0.991	0.556	0.521	0.638	0.480	0.491	0.558
	Aug	1997	0.994				0.642	0.394	0.569
	Feb	1998	0.991	0.369	0.539	0.460	0.677	0.335	0.430
	Sep	1999	0.981	0.801				0.536	0.657

Table 2b EXCEPTIONAL COLD SEASONS

	Month	Year	Obs	PDF for n month lead forecast					
			PDF	n=1	2	3	4	5	6
	Feb	1986	0.015	0.063	0.428	0.532	0.524	0.480	0.315
	Jan	1987	0.030	0.234	0.536	0.469	0.483	0.275	
	June	1991	0.015		0.582	0.355	0.461	0.438	0.689
	May	1996	0.048	0.568	0.550	0.505	0.485	0.599	0.581

Table 2c EXCEPTIONAL WET SEASONS

	Month	Year	Obs	PDF for n month lead forecast					
			PDF	n=1	2	3	4	5	6
	AM	1983	.992	.461	.636	.574	.671	.525	
	Jan	1984	.960	.634	.431	.608	.513	.661	
	Oct	1987	.992	.804	.468			.575	
	Jan	1988	.986	.589	.803	.484	.552	.581	
	DJF	89-90	.987	.411	.249	.391	.561		
	JF	1995	.995	.444	.565	.500	.491	.337	
	Apr	1998	.987	.815	.464	.511	.515	.438	.526.
	SON	2000	.996	.616					

TABLE 2d EXCEPTIONAL DRY SEASONS

	Month	Year	Obs	PDF for n month lead forecast					
			PDF	n=1	2	3	4	5	6
	Apr	1984	.030	.304	.481	.640	.650	.563	.413
	May	1991	.014	.470	.588	.588	.496	.616	.676
	Aug	1991	.030				.410	.410	.344
	Jan	1997	.013	.335	.612	.398	.438	.530	
	Jul	1999	.035			.718	.445	.681	.325

TABLE 3 LEPS SKILL OF FORECASTS FOR 52.5N, 0E

	LEPS for n month lead forecast					
	n=1	2	3	4	5	6
WARM	0.120	0.001	-0.367	-0.049	-0.470	-0.262
COLD	0.642	-0.252	-0.126	-0.211	-0.043	-0.256
WET	0.107	-0.120	-0.200	-0.080	-0.186	-0.723
DRY	0.208	-0.410	-0.337	-0.138	-0.282	-0.015
ALL	0.184	-0.156	-0.246	-0.115	-0.276	-0.283

TABLE 4 95% LEVEL ROC SCORES

	Globe	Europe	All seasons 1 Grid point	I Season 1 Grid poi
1 month means	0.503	0.520	0.537	0.735
3 month means	0.502	0.520	0.541	0.705

FIGURE 1

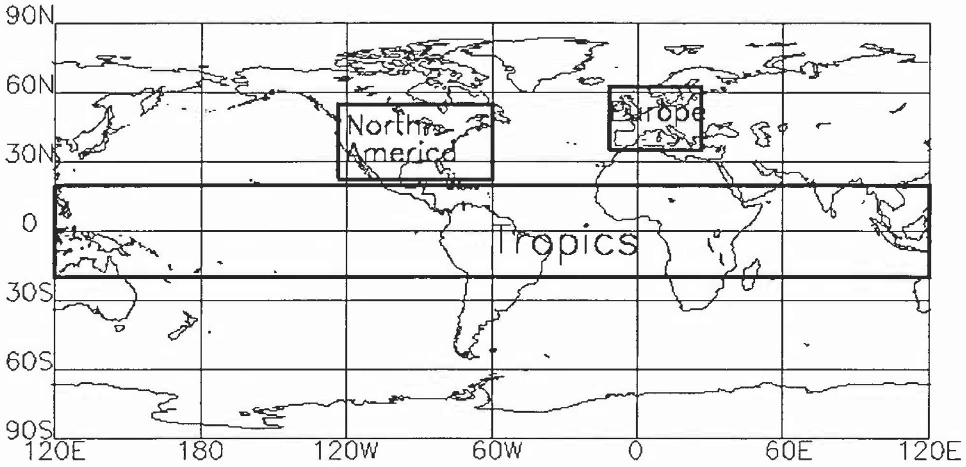
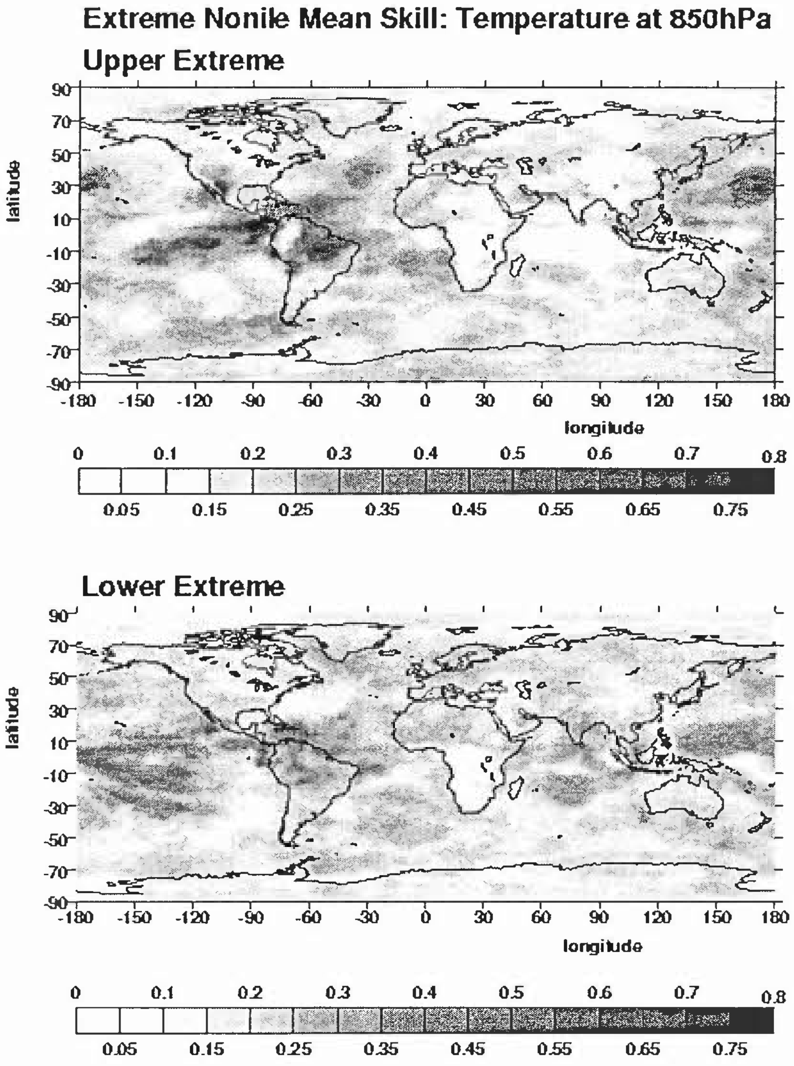

North America
Europe
Tropics
90N
60N
30N
0
30S
60S
90S
120E
180
120W
60W
0
60E
120E

FIGURE 2a



Extreme Nonile Mean Skill: Temperature at 850hPa

Upper Extreme

Lower Extreme

FIGURE 2b

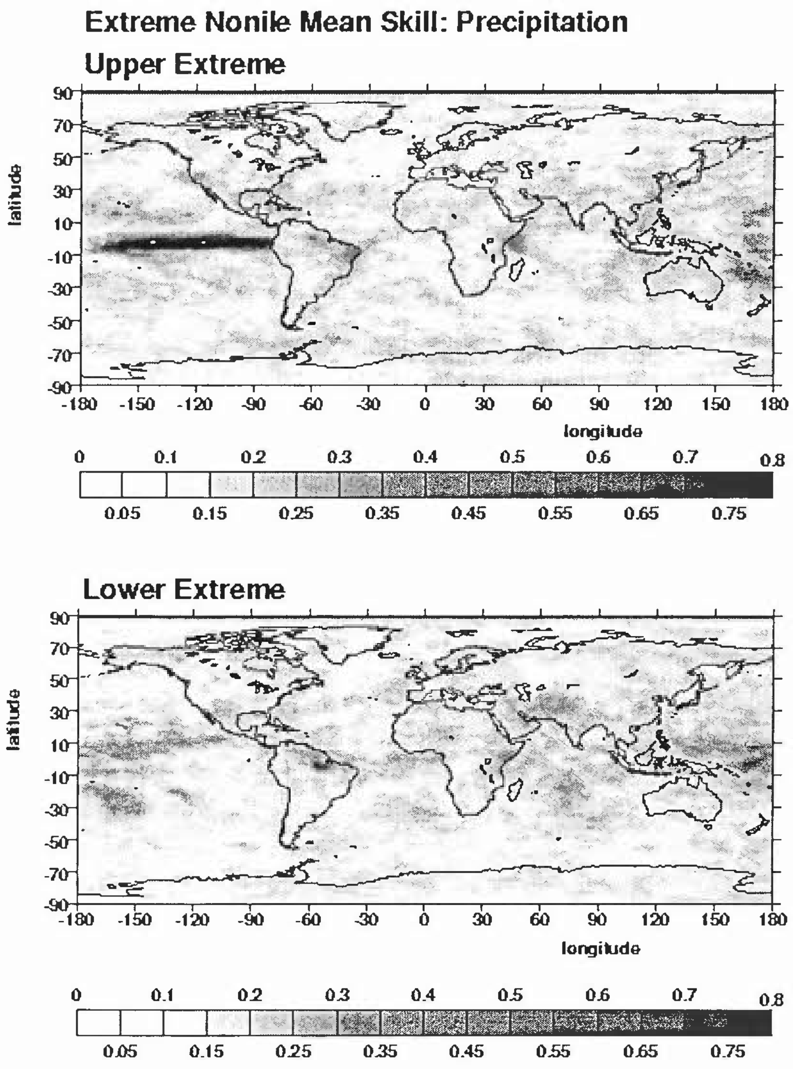

Extreme Nonile Mean Skill: Precipitation
Upper Extreme
latitude
90
70
50
30
10
-10
-30
-50
-70
-90
-180
-150
-120
-90
-60
-30
0
30
60
90
120
150
180
longitude
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
Lower Extreme
latitude
longitude

FIGURE 3a

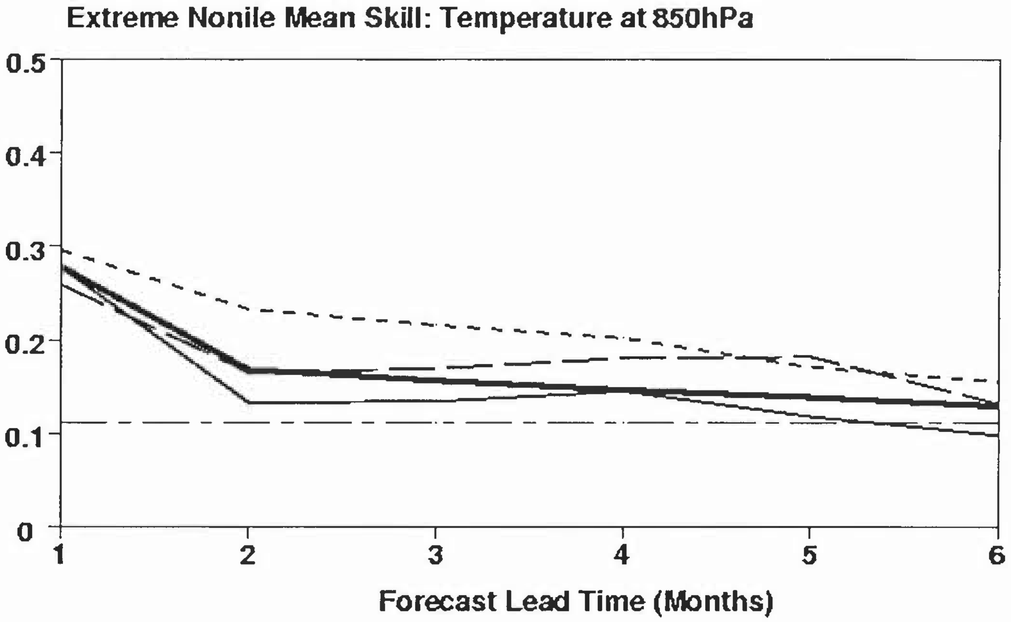

Extreme Nonile Mean Skill: Temperature at 850hPa
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Forecast Lead Time (Months)

FIGURE 3b

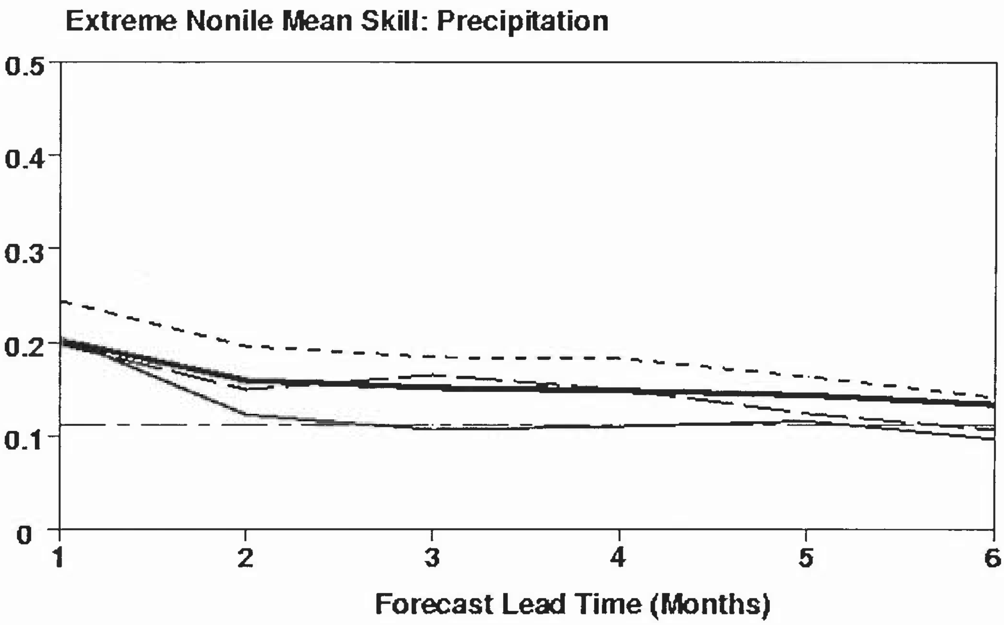

Extreme Nonile Mean Skill: Precipitation
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
Forecast Lead Time (Months)

FIGURE 4a

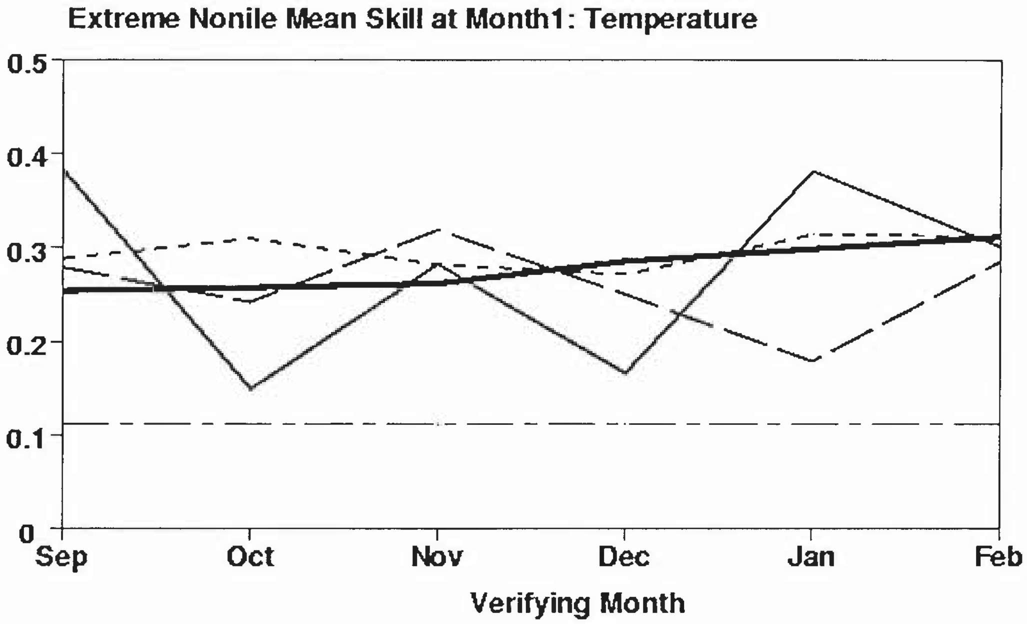

Extreme Nonile Mean Skill at Month1: Temperature
0.5
0.4
0.3
0.2
0.1
0
Sep
Oct
Nov
Dec
Jan
Feb
Verifying Month

FIGURE 4b

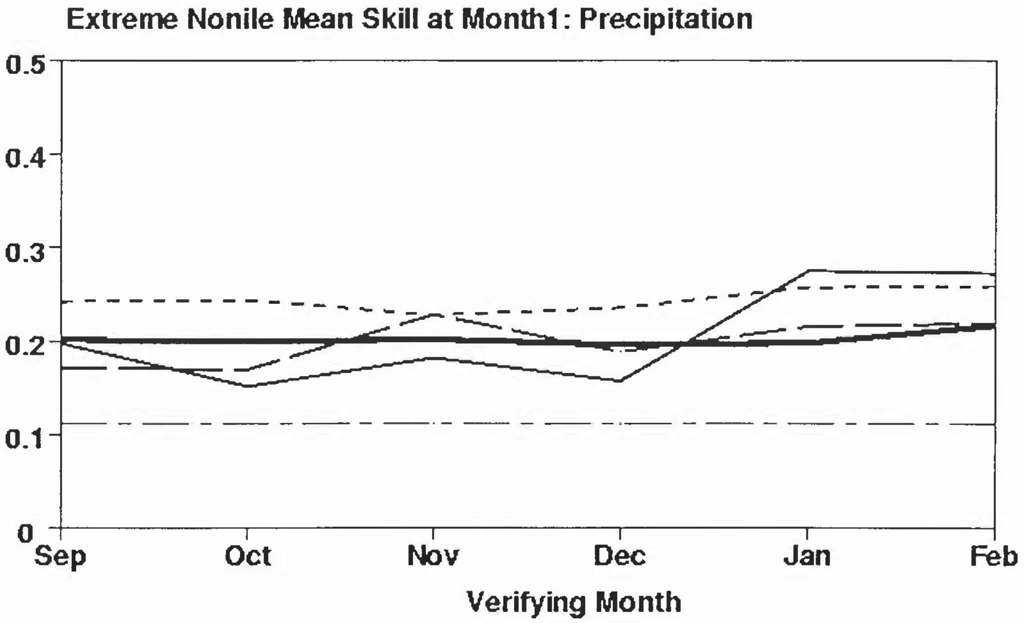

Extreme Nonile Mean Skill at Month1: Precipitation
0.5
0.4
0.3
0.2
0.1
0
Sep
Oct
Nov
Dec
Jan
Feb
Verifying Month

FIGURE 5a

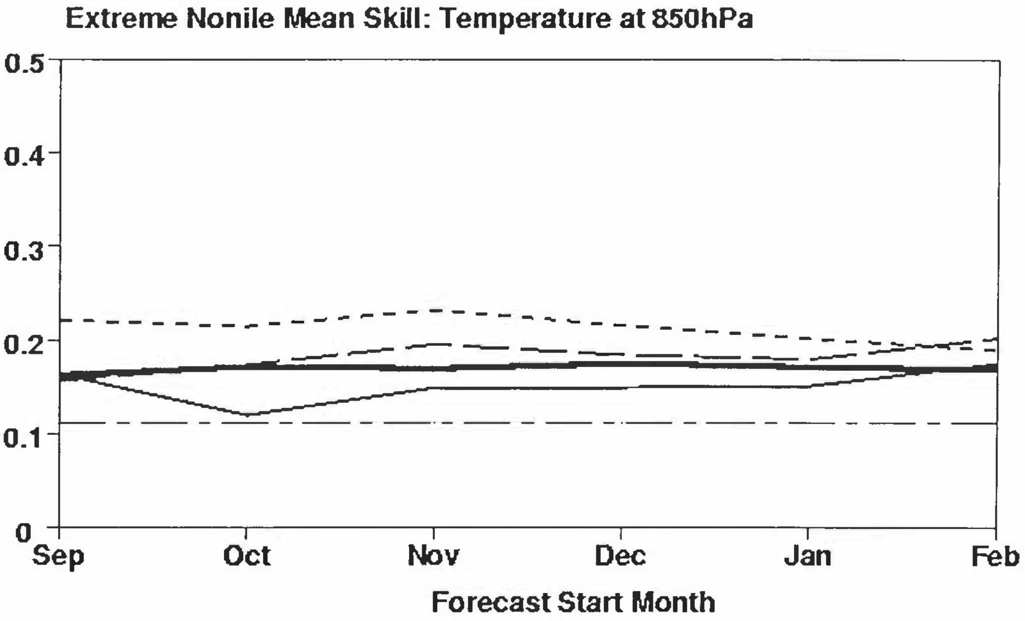

Extreme Nonile Mean Skill: Temperature at 850hPa
0.5
0.4
0.3
0.2
0.1
0
Sep
Oct
Nov
Dec
Jan
Feb
Forecast Start Month

FIGURE 5b

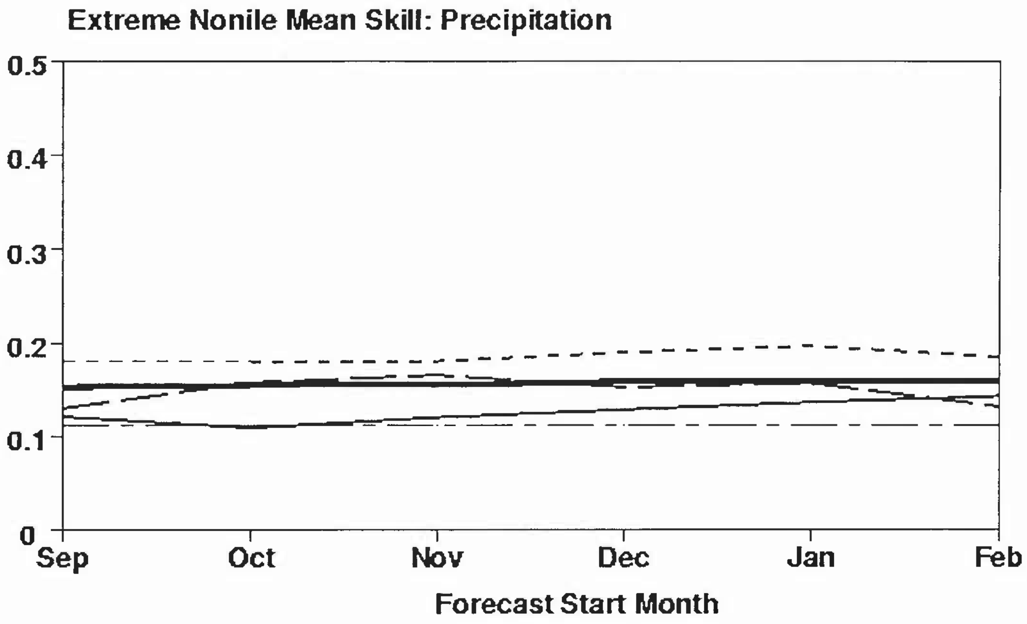
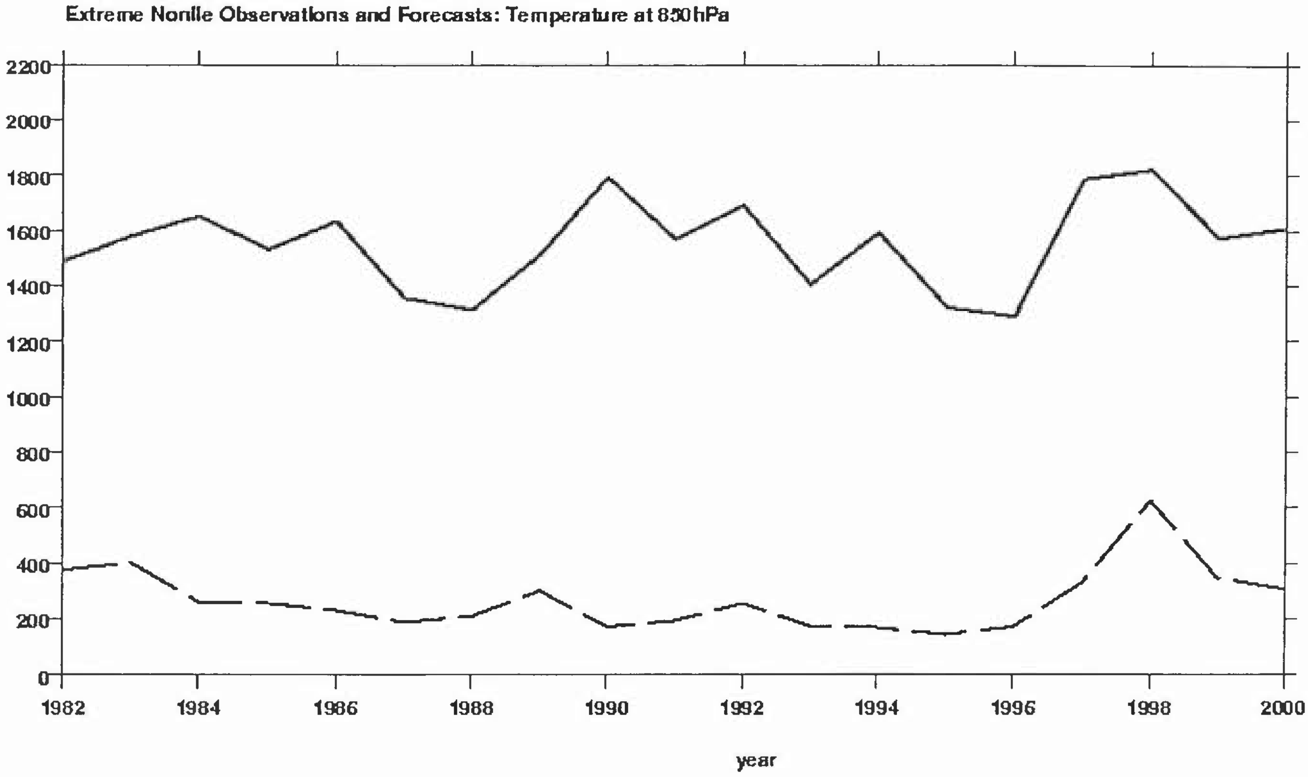

Extreme Nonile Mean Skill: Precipitation
0.5
0.4
0.3
0.2
0.1
0
Sep
Oct
Nov
Dec
Jan
Feb
Forecast Start Month

FIGURE 6a



Extreme Nonile Observations and Forecasts: Temperature at 850hPa

FIGURE 6b

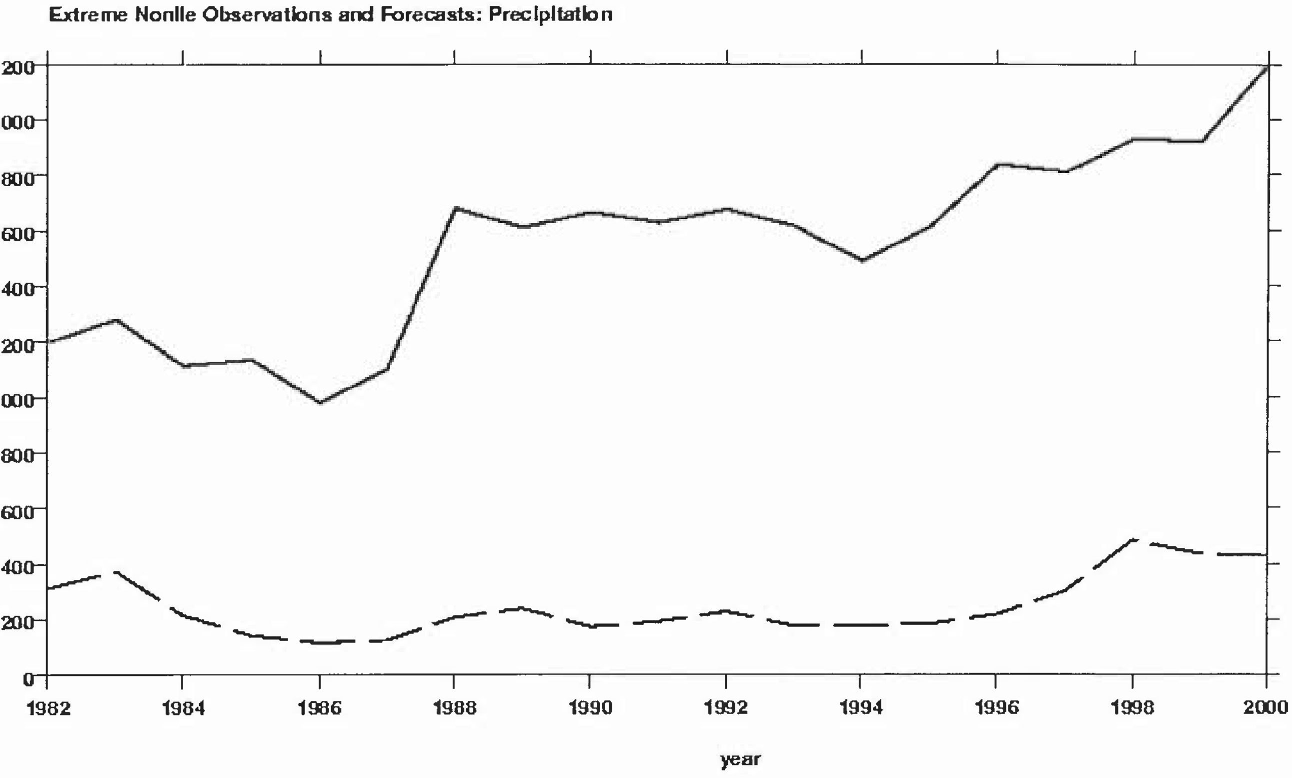

Extreme Nonlle Observatlons and Forecasts: Preclpltatlon
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
year

FIGURE 7a

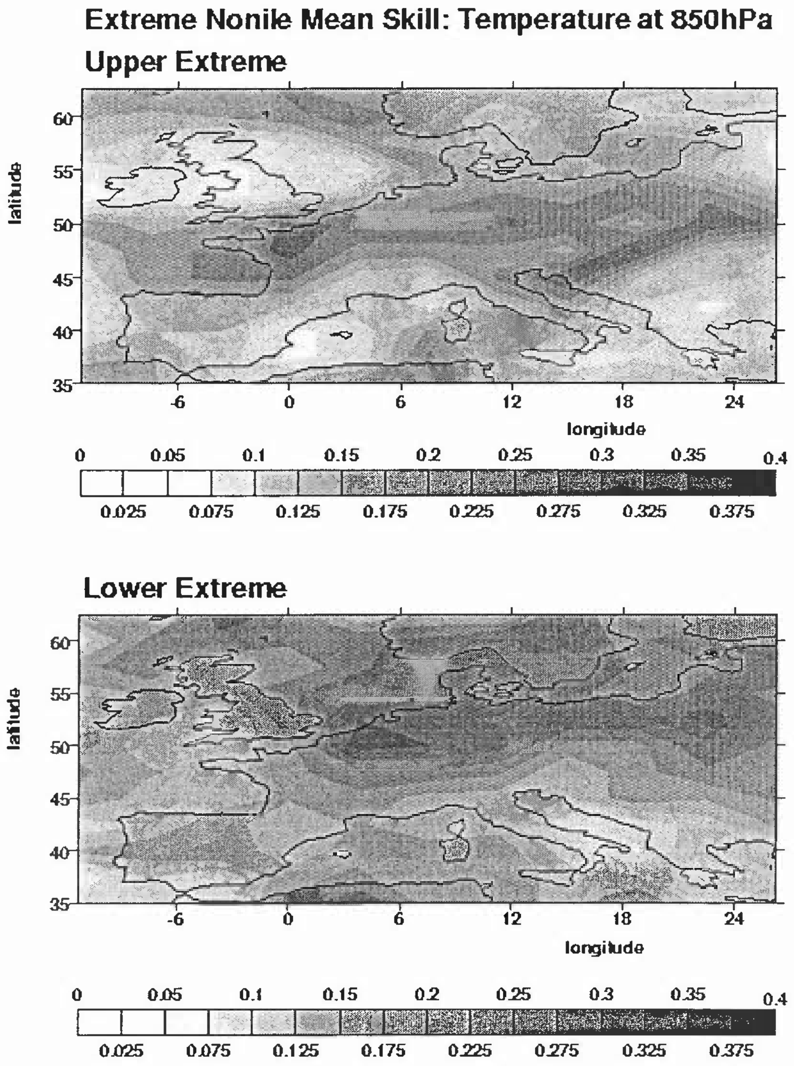
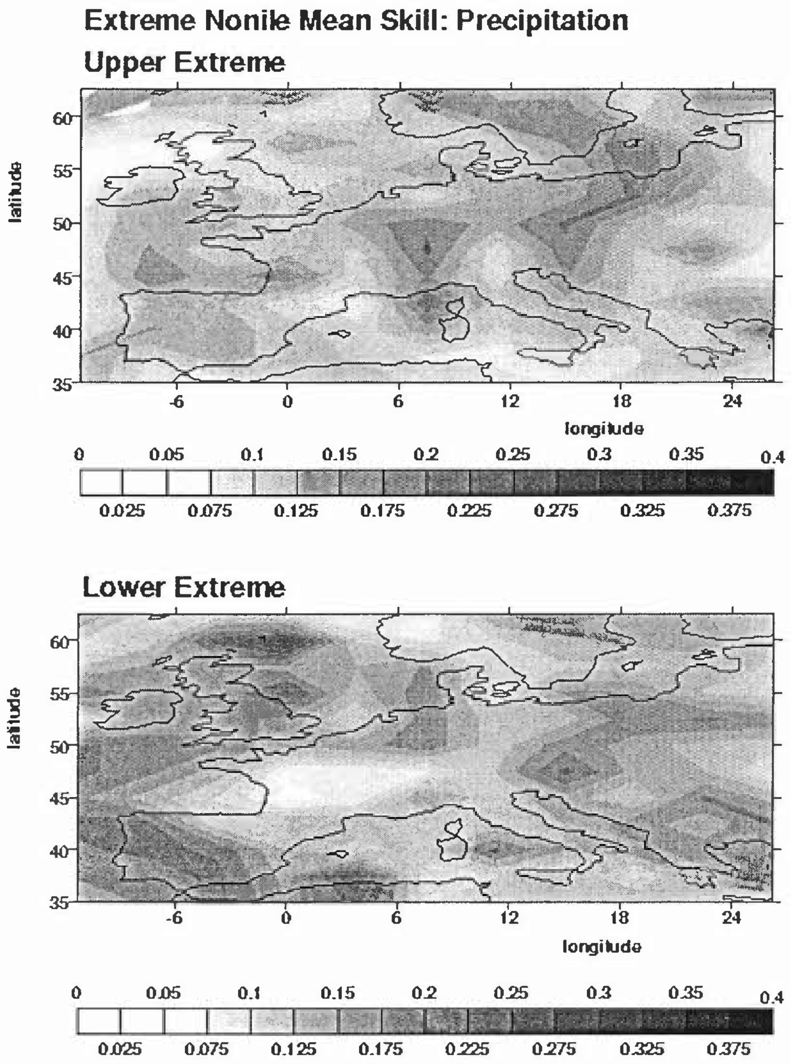

Extreme Nonile Mean Skill: Temperature at 850hPa
Upper Extreme
latitude
60
55
50
45
40
35
-6
0
6
12
18
24
longitude
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.025
0.075
0.125
0.175
0.225
0.275
0.325
0.375
Lower Extreme
latitude
60
55
50
45
40
35
-6
0
6
12
18
24
longitude
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.025
0.075
0.125
0.175
0.225
0.275
0.325
0.375

FIGURE 7b



Extreme Nonile Mean Skill: Precipitation

Upper Extreme

Lower Extreme

FIGURE 8

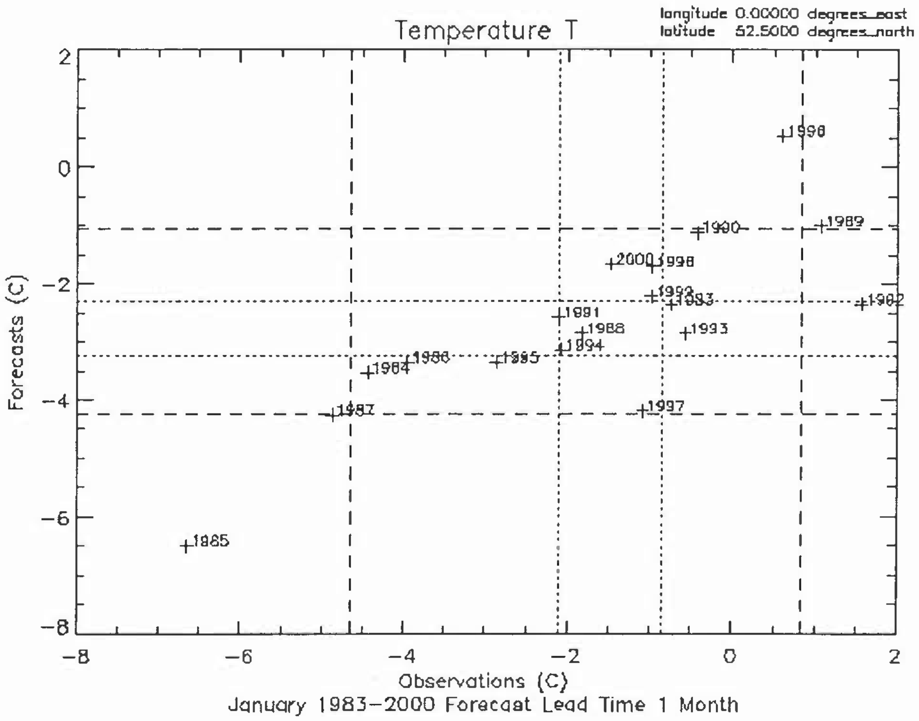
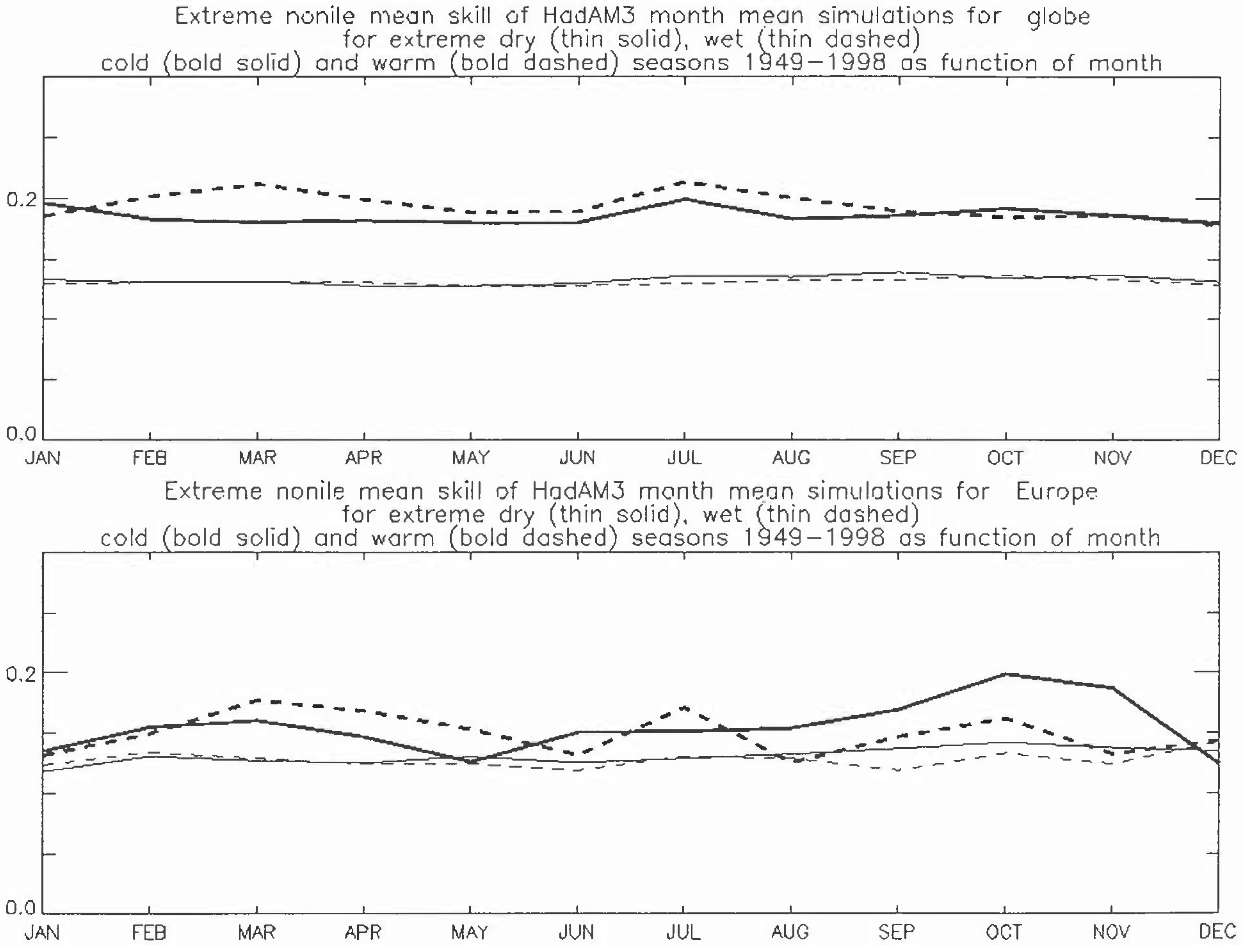

Temperature T
longitude 0.00000 degrees_east
latitude 52.5000 degrees_north
Forecasts (C)
Observations (C)
January 1983–2000 Forecast Lead Time 1 Month
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
1996
1990
1989
2000
1998
1992
1983
1999
1991
1988
1993
1994
1986
1995
1984
1987
1997
1985

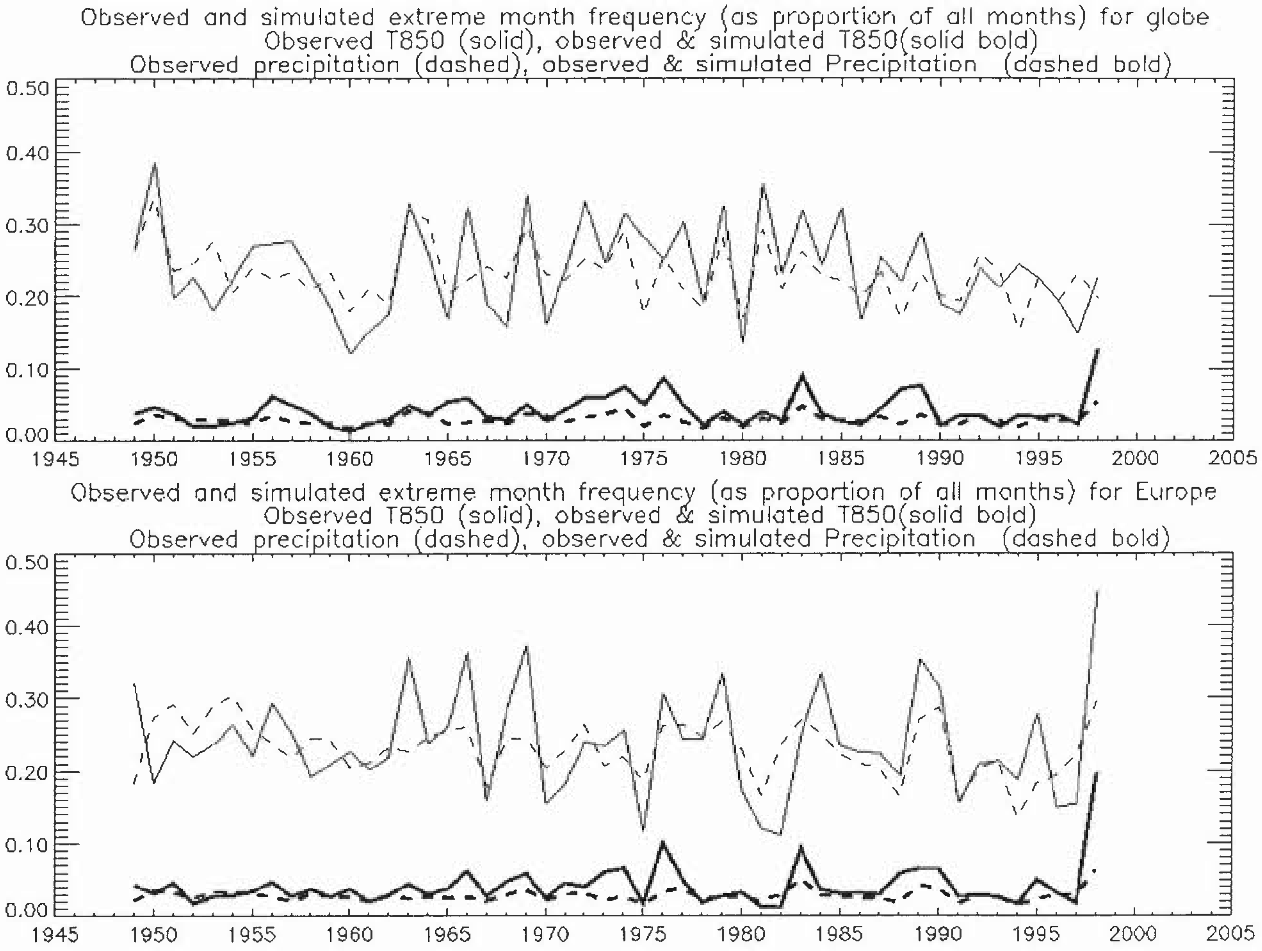
FIGURE 9



Extreme nonile mean skill of HadAM3 month mean simulations for globe
for extreme dry (thin solid), wet (thin dashed)
cold (bold solid) and warm (bold dashed) seasons 1949–1998 as function of month

Extreme nonile mean skill of HadAM3 month mean simulations for Europe
for extreme dry (thin solid), wet (thin dashed)
cold (bold solid) and warm (bold dashed) seasons 1949–1998 as function of month

FIGURE 10



Observed and simulated extreme month frequency (as proportion of all months) for globe
Observed T850 (solid), observed & simulated T850(solid bold)
Observed precipitation (dashed), observed & simulated Precipitation (dashed bold)

Observed and simulated extreme month frequency (as proportion of all months) for Europe
Observed T850 (solid), observed & simulated T850(solid bold)
Observed precipitation (dashed), observed & simulated Precipitation (dashed bold)

FIGURE 11a

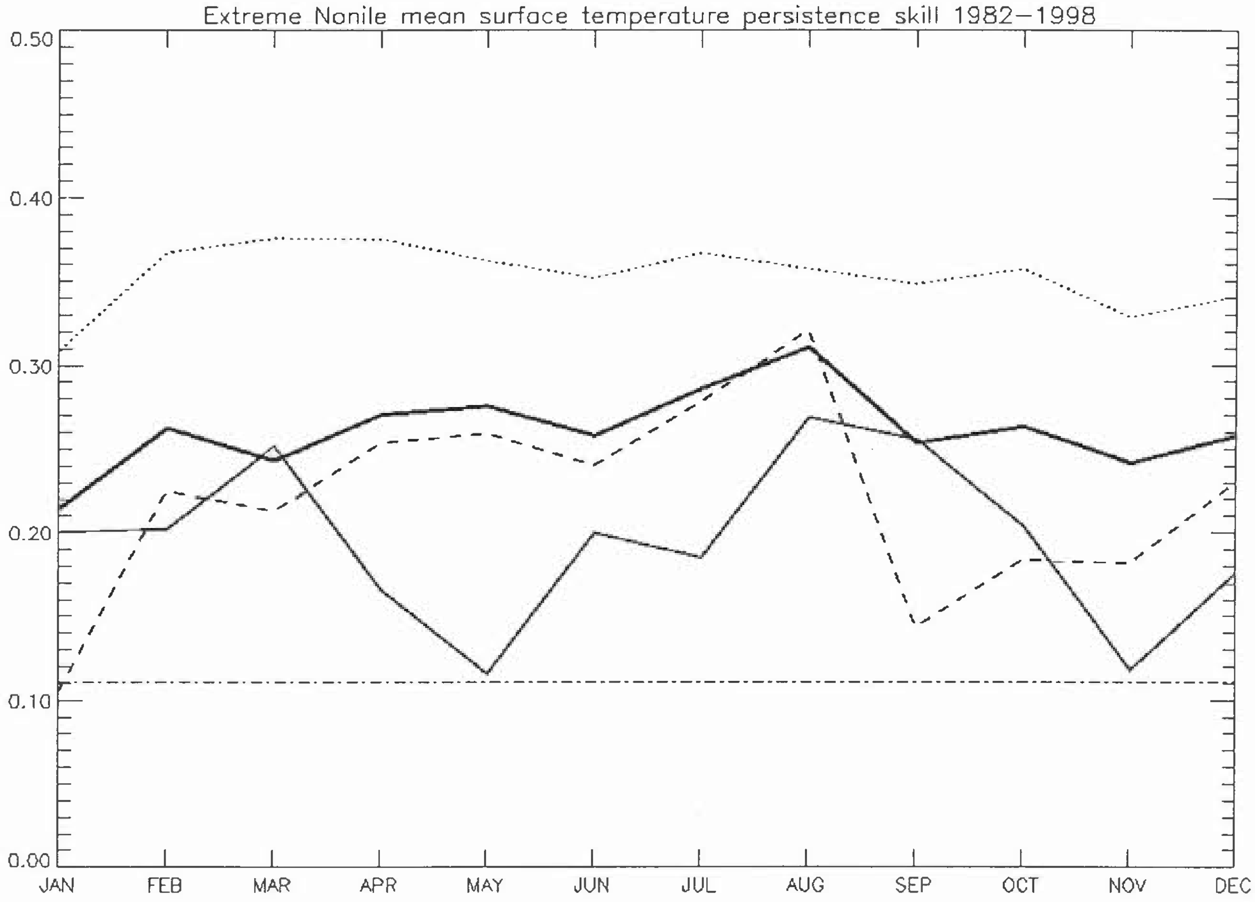

Extreme Nonile mean surface temperature persistence skill 1982–1998
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

FIGURE 11b

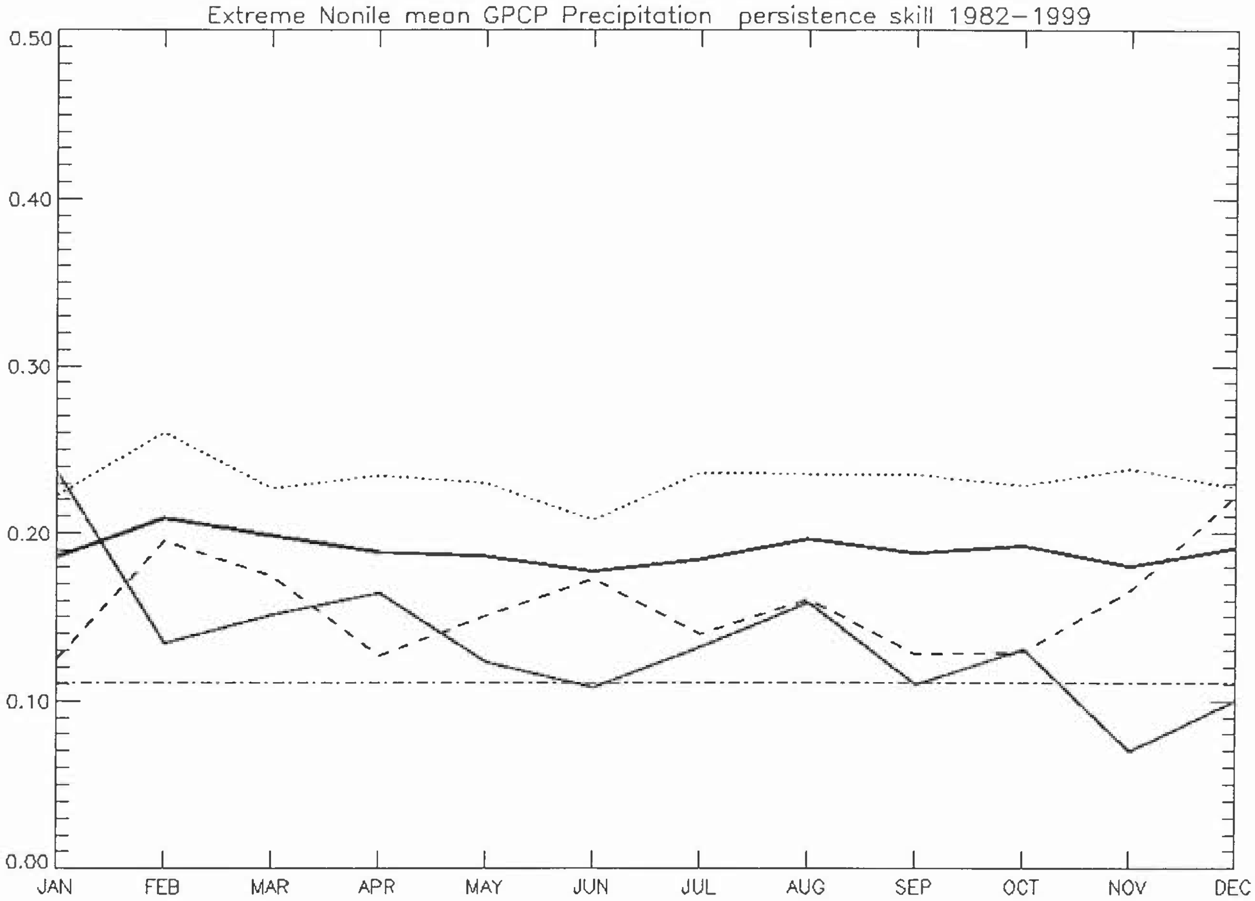

Extreme Nonile mean GPCP Precipitation persistence skill 1982–1999
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

FIGURE 12a

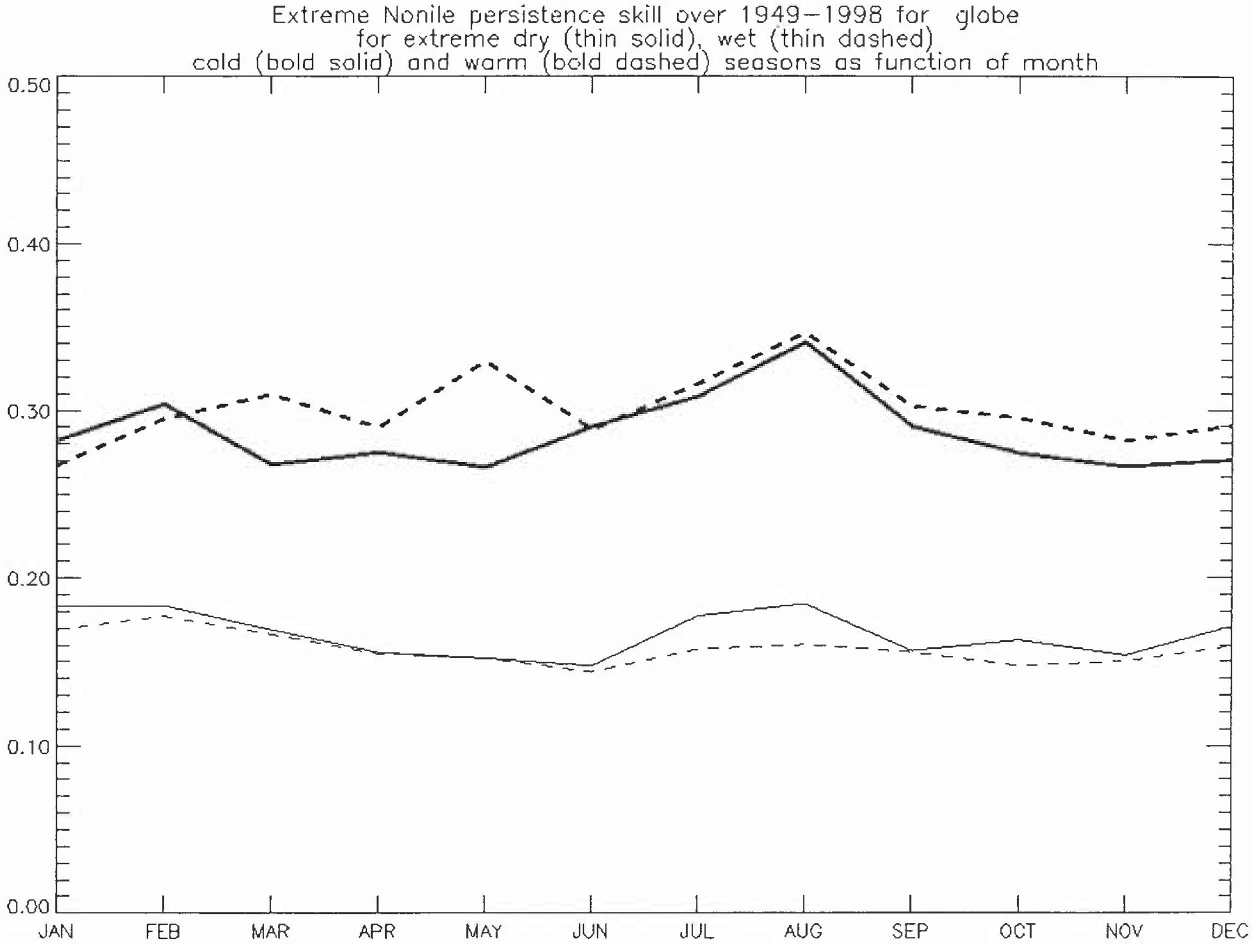

Extreme Nonile persistence skill over 1949–1998 for globe
for extreme dry (thin solid), wet (thin dashed)
cold (bold solid) and warm (bold dashed) seasons as function of month
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

FIGURE 12b

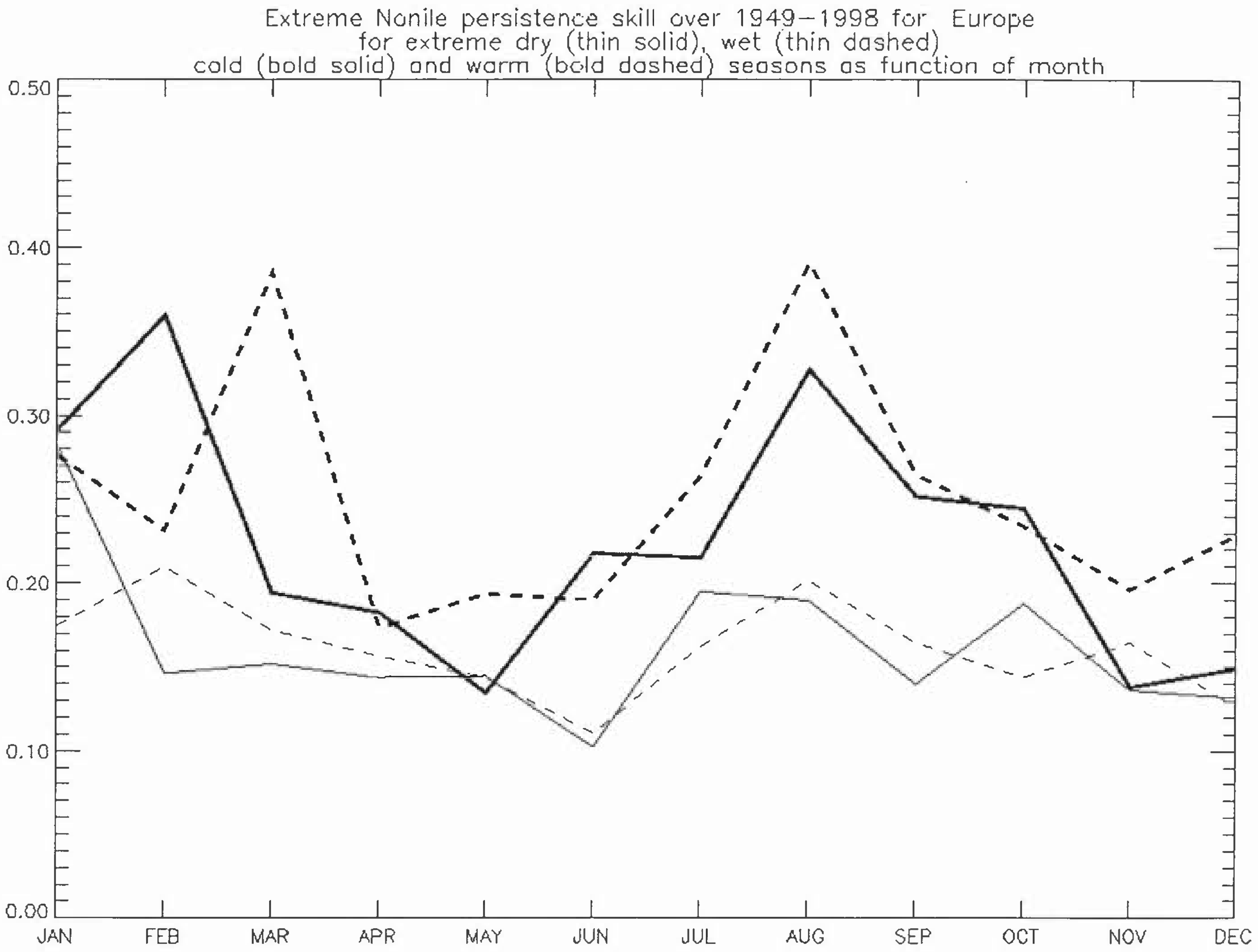

Extreme Nonile persistence skill over 1949–1998 for Europe
for extreme dry (thin solid), wet (thin dashed)
cold (bold solid) and warm (bold dashed) seasons as function of month
0.50
0.40
0.30
0.20
0.10
0.00
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

FIGURE 13

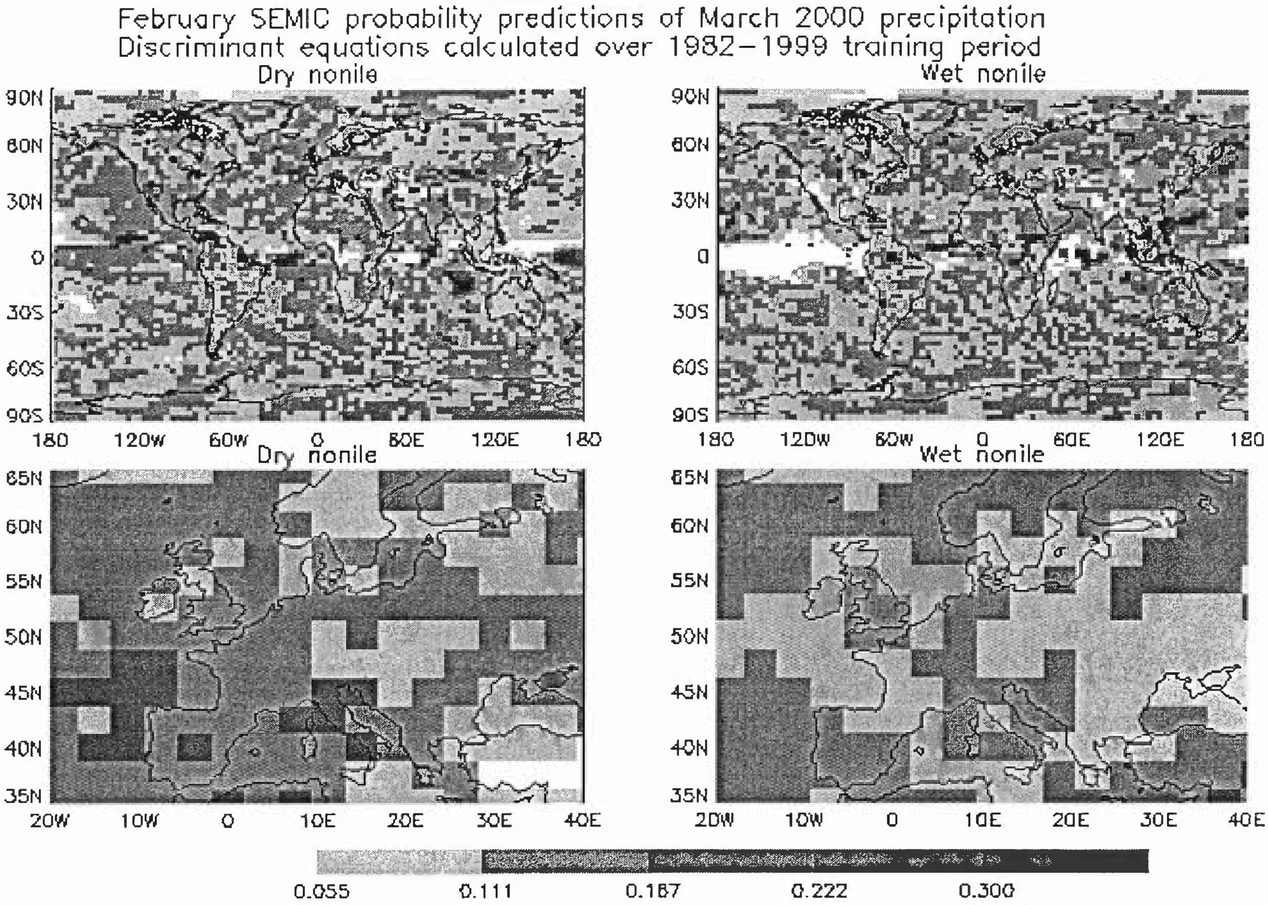

February SEMIC probability predictions of March 2000 precipitation
Discriminant equations calculated over 1982–1999 training period
Dry nonile
Wet nonile
90N
60N
30N
0
30S
60S
90S
180
120W
60W
0
60E
120E
180
Dry nonile
Wet nonile
65N
60N
55N
50N
45N
40N
35N
20W
10W
0
10E
20E
30E
40E
0.055
0.111
0.167
0.222
0.300

FIGURE 14a

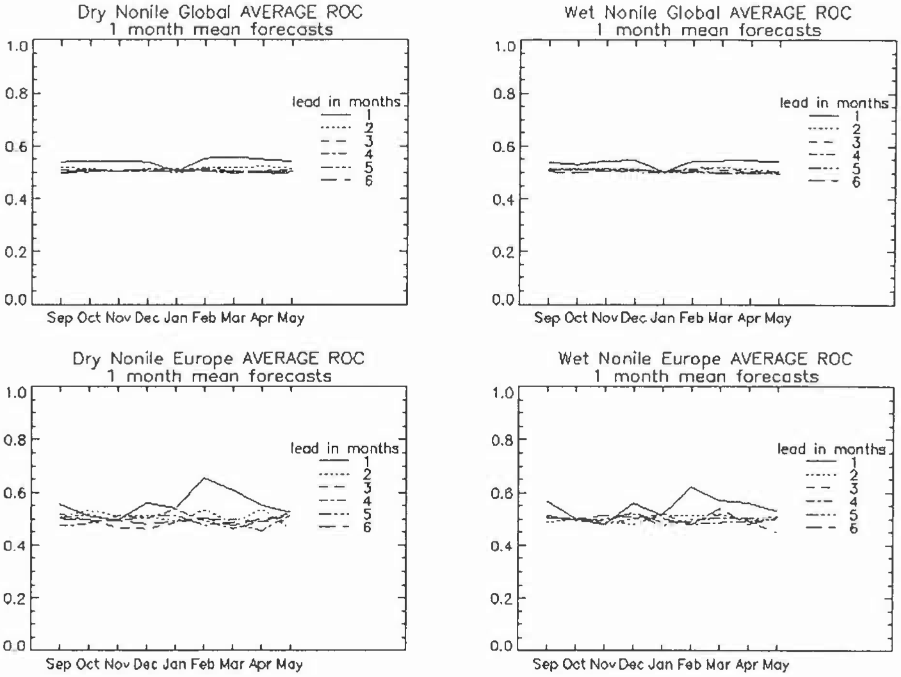
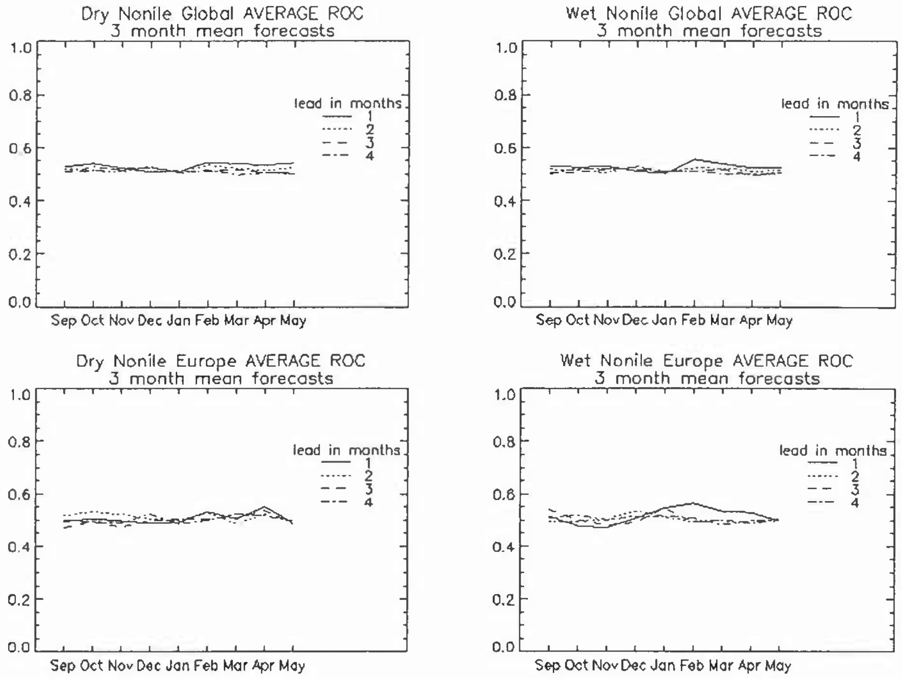

Dry Nonile Global AVERAGE ROC
1 month mean forecasts
Wet Nonile Global AVERAGE ROC
1 month mean forecasts
Dry Nonile Europe AVERAGE ROC
1 month mean forecasts
Wet Nonile Europe AVERAGE ROC
1 month mean forecasts
1.0
0.8
0.6
0.4
0.2
0.0
lead in months
1
2
3
4
5
6
Sep Oct Nov Dec Jan Feb Mar Apr May

FIGURE 14b



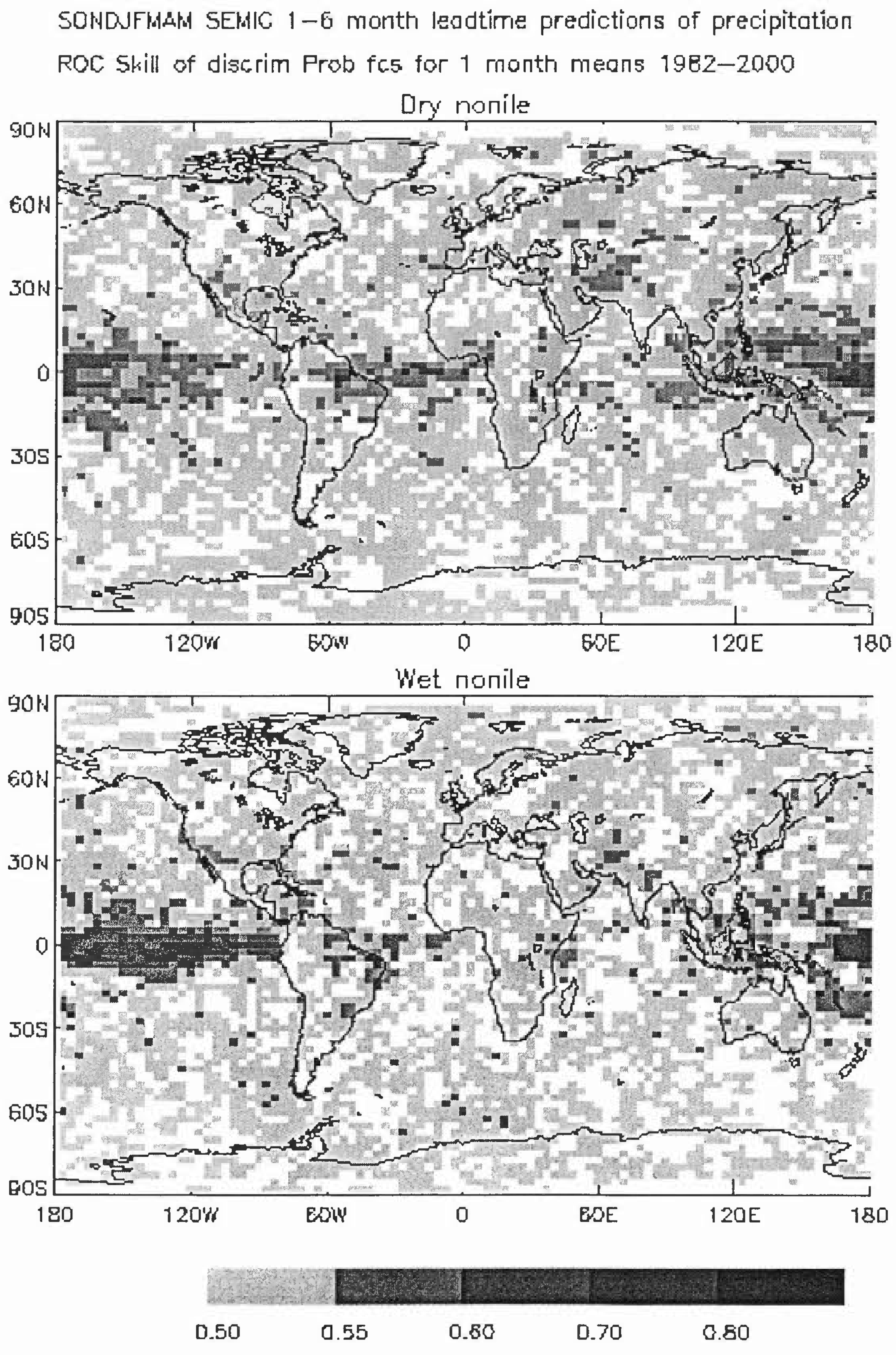
Dry Nonile Global AVERAGE ROC
3 month mean forecasts

Wet Nonile Global AVERAGE ROC
3 month mean forecasts

Dry Nonile Europe AVERAGE ROC
3 month mean forecasts

Wet Nonile Europe AVERAGE ROC
3 month mean forecasts

FIGURE 15


SONDJFMAM SEMIC 1-6 month leadtime predictions of precipitation

ROC Skill of discrim Prob fcs for 1 month means 1982-2000

Dry nonile

Wet nonile