SEASONAL AND ATMOSPHERIC PREDICTABILITY - THE ECMWF PROVOST EXPERIENCE

Tim Palmer and Cedo Brankovic

ECMWF, Reading, United Kingdom

1. Introduction

A series of ensemble integrations with the ECMWF model has been made to assess the predictability of the atmosphere on seasonal time-scales over the 15-year period (1979-1993) of ECMWF reanalysis (ERA). All integrations were about 120 days long, and the model resolution was T63L31. The observed sea surface temperature (SST) was specified at the lower boundary. Both SSTs and the initial conditions for the integrations are taken from ERA. For each season, a 9-member ensemble is created by running from consecutive initial conditions separated by 24 hours. The verification fields are also taken from ERA archives. In total, this series of experiments took the equivalent of about 185 years of integration time.

The experiments have been designed and coordinated with similar sets of experiments performed at Météo-France, Electricité de France, and the United Kingdom Meteorological Office. This coordinated activity has been supported through the European Union PROVOST (PRediction Of climate Variations On Seasonal Time-scale) programme. It is planned to assess all the integrations from the four centres in multi-model ensemble context.

During the experimental period, the largest SST anomalies were observed in the tropical Pacific: the warm anomaly, El Niño, in 1982/83 and cold anomaly, La Niña, in 1988/89. Somewhat weaker warm events occurred in 1986/87 and 1991/92, and cold anomaly in 1984/85.

2. Skill scores

Fig.1 shows the extratropical northern hemisphere (NH) anomaly correlation coefficients for the ensemble mean 500 mb height field, averaged over months 1-3 of the integrations. Results for the four seasons are shown separately. The verification anomalies are taken relative to the 15 year ERA climatology, whilst the model anomalies are derived using the 15-year averaged ensemble mean field (for the appropriate season).

Results show an overall positive level of skill for the NH. In spring and summer, the level of skill drops considerably between the first and second 3 months (months 1-3 and months 2-4). The level of skill is significantly enhanced for the NH if only years when El Niño was active are considered: for example, the El Niño winter of 1982/83, and the La Niña winter of 1988/89 stand out as having particularly high skill (see also anomaly correlation coefficients shown in brackets in Table 1).

Table 1 shows the 15-year average anomaly correlation skill scores of the 500 mb height for various regions, and for the months 1-3 and months 2-4 (shaded). The skill scores are averaged using the Fisher z-transform technique. For the NH, the highest skill is in spring (MAM), despite its sharp drop in the late spring (AMJ). This is consistent with the results from an earlier ECMWF model version, where the highest skill for three sectors in the NH was found in the MAM season (Brankovic and Palmer 1997).

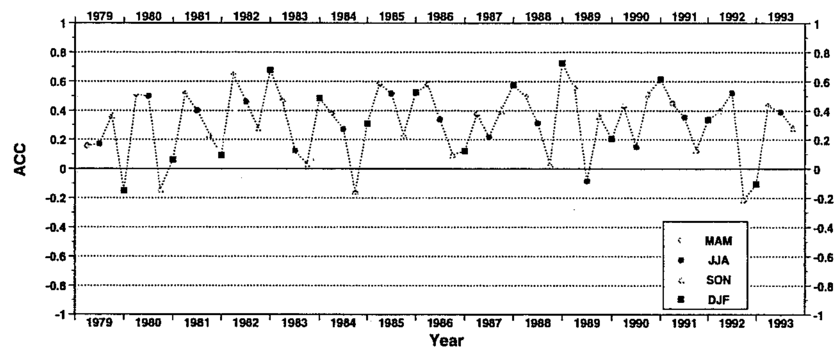

Fig.1 The northern hemisphere 500 mb height anomaly correlation coefficients for months 1-3.

The second highest skill score is in winter (DJF), and is maintained into the late winter (JFM). The contrasting behaviour of skill scores in the late spring and late winter seasons suggests the weakening of dynamical teleconnections between the tropics and the NH extratropics as the warm season approaches.

For Europe, the mean skill is lower than for the NH, which reflects the relatively large interannual variability of skill over regional domains. There is no clear winner between the winter, spring and summer seasons. In autumn, the ECMWF model performs rather poorly, having a negative anomaly correlation. For North America, the highest skill is in winter (DJF) and is maintained at a relatively high level into the late winter (JFM). Skill in summer and autumn clearly lags behind the colder part of the year. When only strong and moderately strong ENSO events considered, skill is much enhanced for North America, though not for Europe. However, due to the small sample (5 cases) these values may not be considered definite.

Season	NH	Europe	N. America
MAM	0.47 (0.48)	0.31 (0.12)	0.51 (0.59)
AMJ	0.22 (0.17)	0.16 (0.21)	0.22 (0.24)
JJA	0.35 (0.40)	0.28 (0.13)	0.34 (0.44)
JAS	0.21 (0.30)	0.22 (0.31)	0.15 (0.32)
SON	0.20 (0.30)	-0.31 (-0.17)	0.37 (0.50)
OND	0.16 (0.35)	-0.11 (0.02)	0.25 (0.47)
DJF	0.36 (0.52)	0.31 (0.36)	0.55 (0.75)
JFM	0.34 (0.53)	0.21 (0.23)	0.44 (0.76)

Table 1: The 500 mb height mean anomaly correlation for months 1-3 and months 2-4 (shaded). Figures in brackets are for 5 strong and moderately strong ENSO events only.

3. Model systematic errors

When assessing the quality of the model simulations by comparing with observations, whatever intrinsic predictability may arise from the underlying SSTs, is compromised by model systematic error. Fig.2 shows the systematic error in 500 mb height from the last 3 months (months 2-4) of the winter integrations. Largest values occur over the north Pacific and north Atlantic, and indicate a westerly bias over these oceanic areas. An error pattern similar to that in Fig.2 is found for the late spring season (not shown), with almost equally large trough-to-ridge amplitude in error dipoles as in winter. For summer, the NH error is largely reduced apart from the very high latitudes.

The error amplitude (locally up to 10 dam) compares with a typical magnitude of interannual variability. These model errors will therefore have a deleterious effect on skill when model fields are compared directly with observations. It is also possible that these errors may have a serious influence on estimates of the intrinsic level of atmospheric seasonal predictability associated with the extratropical SST anomalies.

Given the intrinsic nonlinearity of the atmosphere, it is clear that further work is required to reduce the level of model error on seasonal timescales. The progress that has been made reducing model error in the medium-range forecasts, suggests that the level of systematic error shown in Fig.2 can be significantly reduced over the coming years.

4. Measures of predictability

In order to assess intrinsic seasonal predictability of a certain parameter for some region, a ratio between the signal imposed from the boundary forcing (SSTs in our integrations) and the climatological "noise" must be evaluated. We do not estimate some gross (general) measure of seasonal predictability that should be based on sufficiently long climate record. Gross measures can significantly underestimate potential predictability for individual years. Therefore, our estimates focus on interannual variation of intrinsic seasonal predictability. We feel that the ability to discriminate years with the larger potential predictability is of far greater practical importance than general estimates of predictability.

4.1 Estimates of intrinsic predictability

Our measures of predictability are evaluated for seasonal mean 850 mb temperature by using two types of simple statistics: the *t*-test (grid point by grid point), measuring whether the ensemble mean for a given year was in a statistical sense significantly different from its climatological state, and the ratio of variance (or the *F*-test), measuring the size of ensemble variance for a given year compared with the total ensemble variance taken over all years. The predictions will be significantly different from the climatological mean when the *t*-value is large (and *F*-value is small).

For brevity, we focus on the late winter (JFM) area averages for two European sub-regions: northern Europe, 50-70N, 10W-30E and southern Europe, 35-50N, 10W-30E.

In Fig.3, heavy solid line represents the interannual variation of the *t*-value. For reference, the null hypothesis that the ensemble mean is not different from climate can be rejected with at least 80% or 99% confidence levels, if the *t*-variable exceeds value of 1.3 or 2.6 respectively (given as thin solid horizontal lines). The most

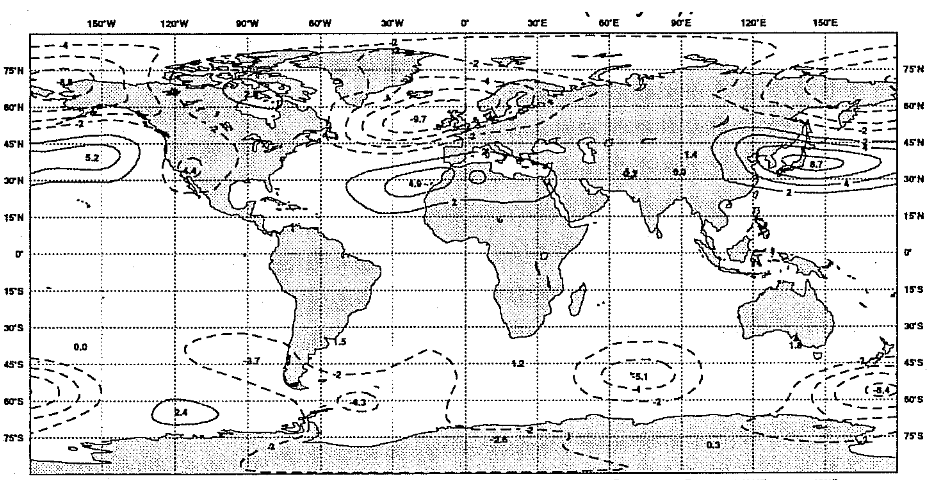

Fig.2 The 500 mb height mean error for the JFM season. Contours every 2 dam. Negative errors dashed.

predictable winter for northern Europe is 1984/85, whilst the most predictable winter for southern Europe is 1982/83. Interestingly, whilst there was a strong El Niño event in 1982/83 (giving rise to extremely large *t*-values over most of the tropics), there was only a weak La Niña event in 1984/85. It is possible that the predictability over northern Europe in winter 1984/85 was enhanced by, for example, the north Atlantic SST anomalies.

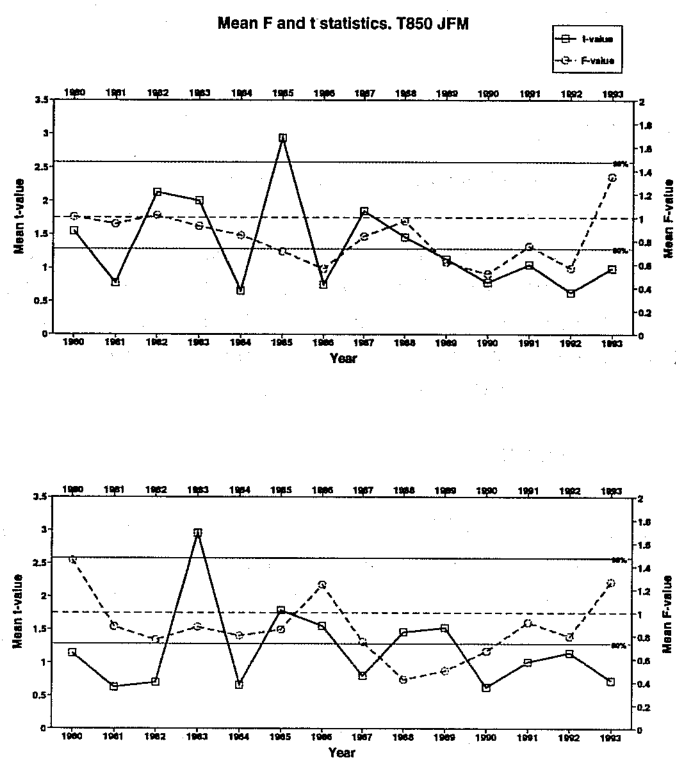

Fig.3 The 850 mb temperature area mean *t*-values (solid curve) and *F*-values (dashed) for the northern Europe (top) and for the southern Europe (bottom).

There is certainly evidence of seasonal predictability for these two subregions of Europe, but the level of predictability varies considerably from year to year. Moreover, a year which shows predictability signal over the northern Europe, does not necessarily show predictability for southern Europe. In winter, there are more predictable years for northern Europe than for southern Europe, but vice versa in summer (not shown). It should be noted that, unlike the skill scores, there is no substantial drop in the level of predictability between the first and second 3-month periods.

The other variable shown in Fig.3 (referred to the right-hand y-axis) is the area averaged ratio of ensemble variance for a given year divided by the total variance from the set of all ensemble fields taken over all the years considered (*F*-value). For regions where there is high intrinsic predictability of interannual fluctuations, the ensemble spread should be much less than the interannual variability of the ensemble mean, and the ratio of ensemble variance to the total variance will be small.

The area averaged *F*-values of seasonal mean 850 mb temperature is shown in Fig.3 as the heavy dashed line, and the reference level of F=1 is shown by the thin dashed horizontal line. Years with relatively large *t*-values and small *F*-values, such as winter 1984/85 for the northern Europe and winter 1982/83 for the southern Europe can be picked out as intrinsically predictable. For years when *F*-value was small but the *t*-value was not significantly different than climatology, one could have forecasted a climatological mean temperature with confidence!

4.2 Simulation of tropical precipitation

From a practical point of view, prediction of seasonal mean precipitation anomalies is of prime interest. Fig.4 shows simulation of such anomalies for two tropical regions. The solid line in Fig.4 shows the rainfall anomaly based on the ensemble mean field. The dashed line shows the anomaly in terms of the 0-24 hour accumulated field in the ERA archives. The dotted line shows the extreme (wet and dry) members of the ensemble. In all cases the curves have been standardized with respect to their own climatologies (i.e. the mean is subtracted, and the anomaly divided by the standard deviation, based on values from the appropriate 15-year dataset).

It can be seen that the simulations for the African Sahel (JJA: Fig.4a) and the Brazilian Nordeste (MAM; Fig.4b) are highly skilful. For the former region, it is interesting to note that the model has successfully captured the decadal trend in precipitation. These results highlight the fact that considerable skill can be expected from seasonal forecasts in many parts of the tropics.

In the headers in Fig.4, the rainfall rates associated with the 15-year ensemble mean and ensemble standard deviation (for ensemble mean **Ens** and individual integrations **Exp**) rainfall is given along with the equivalent values from the reanalysis. For example, for the Nordeste region, the model is about 2 mm/day drier than the reanalysis values (4.3 versus 6.3). On the other hand, the ensemble and reanalysis interannual variability is about the same. For the Sahel, the ensemble is wetter than the reanalysis, but the reanalysis has much more interannual variability than the ensemble. Overall, there is no clear over- or under-estimation of the 15-year mean rainfall rates, though the model does appear to underestimate interannual variability of rainfall rates somewhat.

5. Probabilistic verification

The use of ensembles for seasonal prediction suggests that seasonal forecast products will essentially be probabilistic in nature. We apply some probabilistic statistics to our seasonal forecasts. All the forecasts of an event *E* are stratified according to the forecast probabilities and observed frequency of the event. The event, *E*, considered here is the forecast of 500 mb height anomaly less than -3 dam over the northern hemisphere. For details about reliability diagrams and relative operating characteristic used here, the reader is referred to Wilks (1995) and Stanski et al. (1989).

Fig.5a shows the reliability diagram for the DJF season. The heavy solid line shows the relationship between ensemble forecast probabilities and the frequency that the event *E* verified. In a well-tuned system, this line should lie along the diagonal. The results show that probability forecasts of *E* clearly do have some reliability.

We calculated the Brier skill score (SSBS), which measures the model performance relative to a forecast of climatology, for the winter ensembles from ECMWF and UK Met Office. For these individual centres SSBS were 0.08 and 0.04 respectively. If the forecast score is no better than climate, SSBS=0, and SSBS=1 for a perfect forecast. When the two ensembles are combined, the Brier skill score has increased to 0.12, indicating that both forecasts have benefited from the larger ensemble. In a given example no weighting of ensembles has been applied, but ideally more weight should be attached to the model that has smaller systematic error.

Another diagnostic of probabilistic forecasting is the relative operating characteristic (ROC), that essentially gives some measure of forecast hit and false alarm rates (Stanski et al 1989). As for reliability diagrams, the ROC values are stratified according to probability categories. In Fig.5b, the heavy solid curve should lie above the diagonal if the forecast is to be considered useful (or the normalized area under the curve should exceed 0.5). Clearly, from Fig.5b, the forecasts from the winter ensemble appear to be useful.

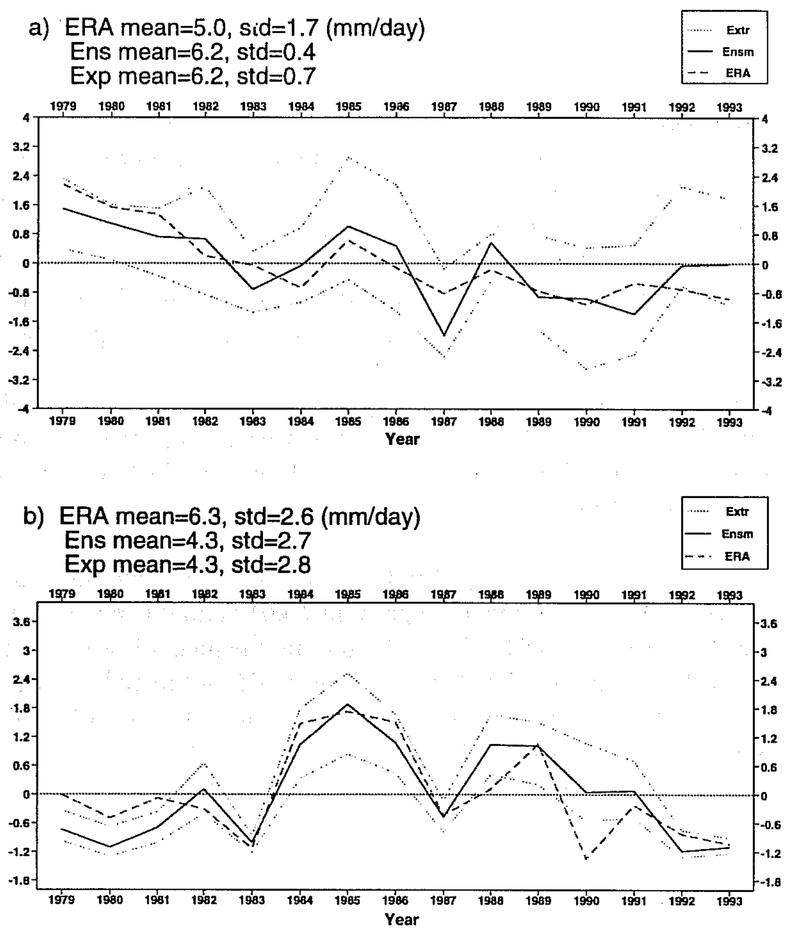

Fig.4 Standardized precipitation for a) the African Sahel in JJA, and b) the Brazilian Nordeste in MAM. Ensemble mean solid line, ERA dashed and extremes of ensembles dotted.

6. Relative impact of initial conditions and boundary forcing

Additional ensemble experiments have been performed to address the issue of the extent to which seasonal predictability is due to the underlying SST anomalies, or the initial conditions (which include land surface initial conditions). For the springs and summers of 1987 and 1988, two further (6-member) 120-day ensembles were made with a) initial conditions for 1987, and SSTs for 1988, and b) initial conditions for 1988 and SSTs

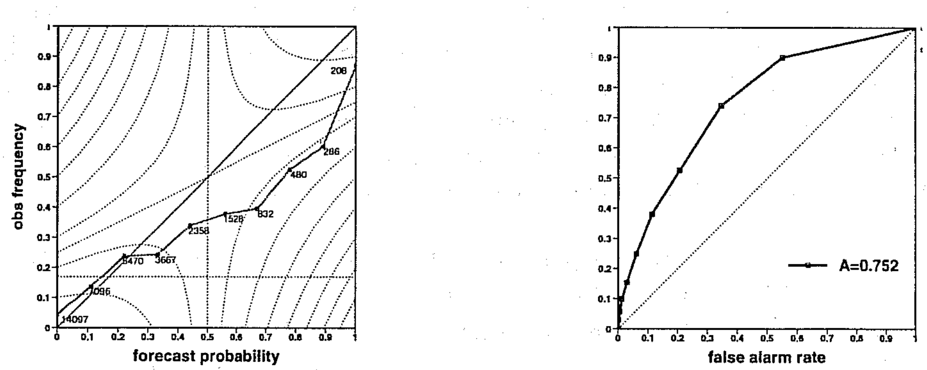

Fig. 5 Reliability diagram (left) and relative operating characteristic curve (right) for the DJF season.

for 1987. The ensembles in the group a) are referred as to hybrid ensembles with "wrong" SSTs; the ensembles in the group b) are referred as to hybrid ensembles with "wrong" initial conditions.

Results are shown in Fig.6 for precipitation in the two regions: the Sahel (JJA) and the eastern USA (AMJ). For the Sahel, 1987 was a drought year; for the eastern USA, 1988 was a drought year. The left-hand panels give the seasonally averaged rainfall amounts for 1987 and 1988 from the individual integrations that comprise the control ensembles. (Note that years are in the reverse order for the USA when compared with the Sahel.) In both cases, the ensemble precipitation is biased in the correct way (drier over the Sahel, wetter over the USA, in 1987).

The middle panels show the impact of the different ("wrong") SSTs between the two years, using the same initial conditions. For the Sahel, it is clear that SSTs played a preeminent role in explaining interannual variability. For the eastern USA, the role of the SSTs is much less clear cut - in three of the experiments there is an increase in rainfall, whilst in the other three there is a decrease. Overall, in terms of the ensemble mean, the 1987 SSTs give a weak increase in precipitation.

The right-hand panels in Fig.6 show the impact of the initial conditions from the two years while keeping SSTs the same ("wrong" initial conditions). Whilst the different initial conditions do not appear to play a significant role in explaining the observed interannual variability over the Sahel, they do appear to have been more important than SSTs in explaining the US rainfall interannual variation. As mentioned above, the initial conditions include the land-surface values.

7. Conclusions

Objective predictability measures and skill scores have been shown from ensembles of 120-day integrations of the ECMWF NWP model, run with observed SSTs. Overall, the levels of skill were positive, both for the tropics, the extratropical northern hemisphere as a whole, and for the European and North American regions. However, the level of skill for the European region was especially variable from one year to the next. This is also true, though to a lesser extent, for North America.

Systematic errors were not negligible in these integrations. Indeed, over the north Atlantic area (and north Pacific), the magnitude of systematic error was comparable with the magnitude of interannual variability. In order to quantify predictability independently of the model systematic error, internal estimates of whether the ensemble distribution for a particular year was significantly different from climatology were assessed. In fact, such estimates of internal statistical significance are not completely independent of model error. In particular, it is likely that with model error, the simulated response to the observed SSTs will be non-optimal.

Despite overall positive levels of skill and internal statistical significance, these two estimates do not address the question of whether they constitute "useful" levels of skill. Some estimates of probabilistic verification, like reliability curves and relative operating characteristic curves may be used to assess usefulness of seasonal ensemble predictions. However, it is likely that the potential "usefulness" of seasonal forecasts will be seriously compromised by the current level of model systematic error.

On the other hand, it is likely that these errors can be significantly reduced over the coming years. As such, the current level of model error is likely to be transient, and should not be factored into any assessment of the potential economic value of seasonal forecasts. One possible methodology for this would be to analyze economic value in terms of the perfect-model ensemble, i.e. when one ensemble member is deemed to be the "truth".

A second way of minimizing the effect of model error is to correct for it *a posteriori* using statistical post-processing techniques. For example, the singular value decomposition (SVD) technique is now well

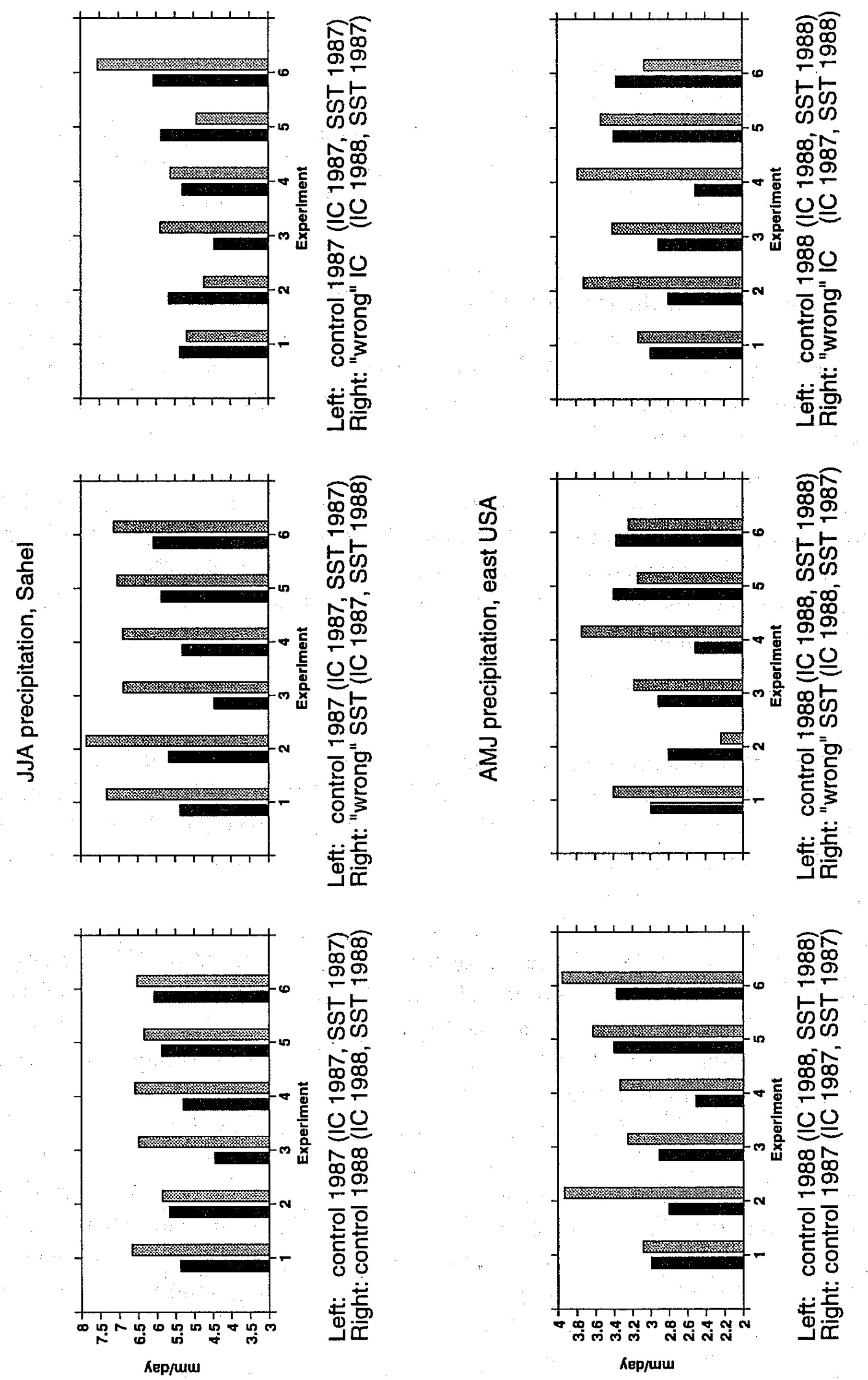

Fig.6 Relative impact of initial conditions (IC) and SST on precipitation over the Sahel (top) and eastern USA (bottom). See text for explanation. Left-hand bars refer to the control ensemble and are identical in each row.

established for such studies, and is being applied to the current integrations as part of the Provost programme. However, it is unlikely that reliable statistics can be obtained from only 15 years of integrations.

Based on these remarks, it would be worth considering a second, more extensive set of ensemble integrations in a few years time. Such a set could take advantage of proposed ECMWF 40-year reanalysis. Ensemble integrations from this set could be made in both coupled and uncoupled mode. The "usefulness" of these ensembles should be assessed through collaborative research with potential users of seasonal forecasts. It is expected that the level of model error for these second set of integrations will be substantially lower than those used in the current set. Nevertheless, it is likely that both SVD post-processing and perfect-model analysis will prove useful in the overall assessment of the results.

One of the major objectives of the Provost programme is the analysis of the uncoupled integrations from the four modelling groups, as a single multi-model ensemble. The major work to achieve this objective has been the archiving of all model results in a consistent fashion at ECMWF. The next stage of post-processing the multi-model is about to take place and an important component of this analysis will be in terms of probability forecast products.

References:

Brankovic, C. and T.N. Palmer, 1997: Atmospheric seasonal predictability and estimates of ensemble size. *Mon. Wea. Rev.*, **125**, 859-874.

Stanski, H.R., L.J.Wilson, W.R.Burrows, 1989: Survey of common verification methods in meteorology. 2nd edition, WMO WWW Tech. Rep No. 8 (WMO/TD. No. 358).

Wilks, D.S., 1995: *Statistical methods in the atmospheric sciences.* Academic Press, 464 pp.