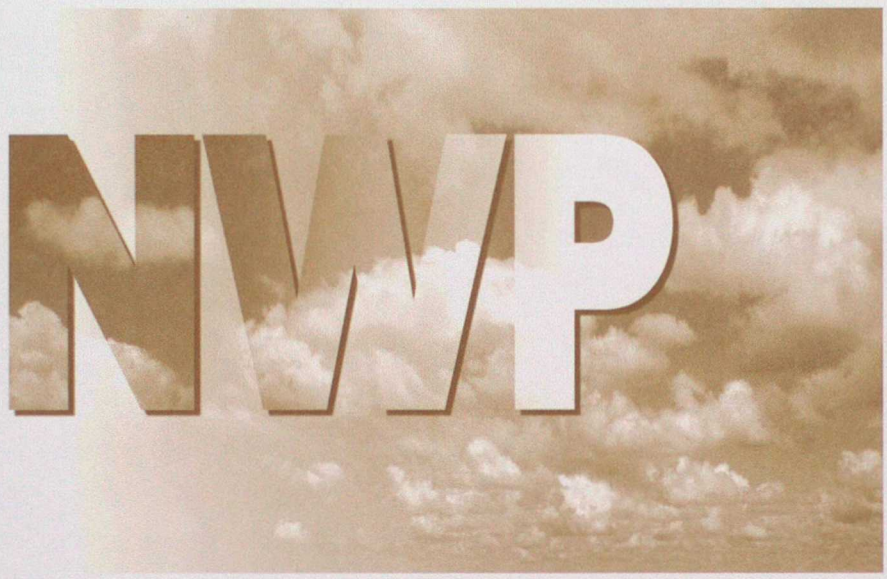
Numerical **Weather Prediction**



Forecasting Research
Scientific Paper No. 54

An assessment of seasonal predictability using Atmospheric General Circulation Models

by

R.J. Graham, A.D.L. Evans, K.R. Mylne, M.S.J. Harrison and K.B. Robertson

April 1999

The Met.Office

Excelling *in weather services*

Forecasting Research
Scientific Paper No. 54

An assessment of seasonal predictability using Atmospheric General Circulation Models

by

R.J. Graham, A.D.L. Evans, K.R. Mylne, M.S.J. Harrison and K.B. Robertson

April 1999

A version of this paper has been submitted for publication in the Quarterly Journal of the Royal Meteorological Society, and formed part of the UKMO contribution to the final report on the EC collaborative project, PROVOST (PRedictability Of climate Variations On Seasonal to inter-annual Timescales).

Numerical Weather Prediction
Room 344
Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

Please notify us if you change your address or no longer wish to receive these publications.

Tel: 44 (0)1344 856245 Fax: 44 (0)1344 854026 e-mail: jsarmstrong@meto.gov.uk

An assessment of seasonal predictability using Atmospheric General Circulation Models

R.J. Graham, A.D.L. Evans, K.R. Mylne, M.S.J. Harrison and K. B. Robertson
Meteorological Office, UK

Summary

Seasonal predictability is investigated using a 15-year set of four-month-range, 9-member ensemble integrations from AGCMs involved in the European project, PROVOST. The integrations were performed using prescribed ideal (observed) Sea Surface Temperatures (SST), and skill attained (referred to as "potential" skill) therefore represents an estimated upper bound on skill achievable with current models using predicted SST. Most analysis is presented for the UK Met. Office Unified Model (UM), the ECMWF T63 model (referred to as T63) and an 18-member multiple-model ensemble (JT2) constructed from these individual models. The benefits of higher-order multiple-models (employing all 4 participating PROVOST AGCMs) are also investigated. Evaluation is focused on four assessment regions; the tropics, the northern extratropics, Europe and North America. A variety of skill measures are used, with emphasis on assessment of probabilistic skill for the basic events: 3-month-mean 850 hPa temperature above/below normal; 3-month precipitation accumulation above/below normal. A summary of main results is provided below.

Potential skill: All skill measures for month 1-3 850 hPa temperature and precipitation calculated for the entire tropical and northern extratropical regions indicate that, while skill is highest in the tropics, skill is also available over the northern extratropics for all three models (UM, T63 and JT2) in all seasons. Scores for the northern extratropics are highest in spring (MAM). Scores for precipitation are generally lower than for 850 hPa temperature, however there is evidence of substantial potential for rainy season predictions in some tropical regions. Over Europe and North America skill scores for 850 hPa temperature are (for at least one of the UM, T63 and JT2 models) comparable to those of the northern extratropics in all seasons. Peak skill occurs over Europe in MAM (as found for the northern extratropics). In contrast peak skill over North America occurs in DJF, apparently as a result of enhanced predictability of the DJF PNA pattern during ENSO events. In non-ENSO years skill over Europe and North America is similar, suggesting that the greater predictability often attributed to the North American region relative to Europe may apply only during ENSO events. Skill for months 2-4 is generally lower than for months 1-3, though there is evidence that during ENSO events levels of skill in the first three months are maintained into the second three months. Potential skill is also found for precipitation over Europe and North America, with the exception of JJA and SON over Europe when scores appear similar to those available from climatology.

Skill prediction: ENSO forcing has a marked global impact on model predictability. Centres of predictability in the tropical Atlantic and Indian Ocean regions in non-ENSO years are found to transfer to the East Pacific during ENSO events. Largest ENSO-related skill enhancements over North America are found in DJF and over Europe in the following (post-ENSO peak) MAM season. Ensemble spread appears a useful indicator of ensemble-mean skill in some seasons over Europe and North America. Thus prospects for skill prediction appear promising, perhaps using strategies which combine information on both the state of ENSO and ensemble spread.

Benefits of multiple-model ensembles: Multiple-model ensembles enhance prediction capabilities, allowing the strengths of the individual AGCMs to be exploited without extensive *a priori* calibration of each model. The multiple-model ensembles frequently provide a filter for the more skilful individual model (the identity of which varies with season and region). The key factor determining the skill of the multiple-model appears to be the skill of the most skilful component ensemble and does not appear strongly connected with the increased ensemble size.

Use of persisted SST: Tests indicate that a substantial proportion of the skill achieved using observed SST is retained using persisted SST anomalies from the month preceding the initial date of the integration, indicating that use of persisted SST anomalies is a viable method for real-time seasonal prediction, at least for ranges of up to one season ahead.

User value: A methodology for linking technical forecast quality with financial value for users has been outlined using the Relative Operating Characteristic (ROC) and the user cost/loss matrix. Examples indicate promising potential for user value of seasonal predictions not only over tropical areas, but also in extratropical areas such as Europe. There is evidence that probabilistic predictions provide potential value for a greater range of applications than do deterministic predictions.

KEYWORDS: Seasonal predictability AGCM ensembles Skill prediction Multiple-models User value

1. Introduction

The scientific basis for seasonal prediction stems primarily from evidence that the atmosphere's lower boundary, particularly the Sea Surface Temperature (SST), influences weather regime frequency statistics (Palmer and Anderson, 1994) - and consequently the seasonal-mean weather conditions. The SST field itself evolves slowly compared with individual synoptic-scale weather systems, and is often relatively predictable (at least in the tropics) - thus representation of the SST evolution in Global Circulation Models, so-called dynamical seasonal prediction, provides a potential means of generating forecasts of seasonal-average weather. This paper describes investigations to evaluate that potential.

The work has been performed as part of the recent European project, PROVOST (PRediction Of climate Variations On Seasonal and inter-annual Timescales) which has provided an extensive simulation dataset that expands in a number of ways on those previously used. Briefly, the dataset (described in Section 2) comprises 4-month-range, 9-member ensemble AGCM integrations run over ideal (observed) Sea Surface Temperatures (SST), for all four seasons over the 15-year period 1979-93. The period contains 5 El Niño Southern Oscillation (ENSO) events, allowing assessment of the impact of ENSO on predictability. Moreover, integrations were repeated at four European centres using different AGCMs, allowing opportunities for comparing model performance and for assessing the potential performance benefits from combining predictions from different models. Here we describe research performed at the UK Meteorological Office (UKMO) as part of the PROVOST project. The key objectives and scope of the study are expanded below.

i) Assessment and comparison of AGCM "potential" skill.

A main objective is to obtain estimates of the upper-bound on skill (referred to as "potential" skill) achievable using coupled Global Circulation Models (CGCMs) by measuring the skill obtained when uncoupled AGCMs are forced with observed (i.e. perfect predicted) SST. Emphasis is placed on assessment and comparison of the UKMO Unified Model (Cullen, 1991) and the ECMWF T63 ensembles (hereafter referred to as the UM and T63 ensembles). The purpose of the intercomparisons is to gain insight into the potential enhanced capability available through combining the strengths of two or more AGCMs. Potential skill is evaluated on a global scale and for all four seasons with emphasis on the tropics, northern extratropics, Europe and North America.

Initial investigations were performed to assess the systematic AGCM errors in the PROVOST simulations. It was found that although, for all models, biases may be large relative to the climate variability in some regions/seasons, an *a posteriori* linear correction for the bias resulted in promising levels of overall skill. In this paper we therefore focus on verification of the bias-corrected simulations. A discussion of the model biases themselves is not presented.

ii) Assessment of the prospects for skill prediction

Large amplitude warm/cold SST events associated with the El Niño/La Niña phases of ENSO (El Niño Southern Oscillation) in the tropical Pacific may, through persistent thermal forcing on the atmosphere, give rise to enhanced predictability in some regions. Thus "windows of opportunity", when seasonal predictions might be expected to have relatively high skill, may occur for some regions/seasons - and may be identifiable in advance. To investigate the impact of ENSO forcing on predictability, the distribution of skill between the 5 ENSO years (El Niño; 1982/83, 1986/87, 1991/92 and La Niña; 1984/85, 1988/89) and 10 non-ENSO years is investigated. Hereafter in this paper we will refer to warm SST events as PW (Pacific Warm), cold events as PC (Pacific Cold) and both cold and warm events collectively as PC/W, to emphasise that it is specifically the presence (or otherwise) of SST anomalies in the tropical Pacific SST that is of interest.

The degree of spread in the ensemble solutions has been used with some success as a predictor of skill in medium-range ensemble integrations (see e.g. Molteni *et al.* 1996); low spread ideally associating

with high ensemble-mean skill. Here we investigate whether ensemble skill/spread relationships are also present at the seasonal range. The impact of PC/W events on the ensemble skill/spread relationship is also examined.

iii) Assessment of the benefits available from multiple-model ensembles
Studies of medium-range prediction have shown that multiple-model ensembles, comprising a combination of members run with different AGCMs, provide significant performance benefits relative to the component individual ensembles (e.g. Evans *et al.*, 1998). Here we investigate whether such benefits extend to the seasonal range. Analysis is focused on a joint 18-member ensemble (referred to as JT2) constructed by combining the UM and T63 ensembles; however higher order multiple-models, comprising three and four of the PROVOST AGCMs are also evaluated. Benefits from combining models in this way derive potentially from both the inclusion of complementary predictive information (Brown and Murphy, 1996; Evans *et al.*, 1998), and also, in the case of ensembles, from the increased ensemble size. The factor providing the greatest benefit, i.e. model combining or increased ensemble size, is discussed.

iv) Assessment of skill available using persisted SST anomalies for boundary forcing
Extensive hindcast validation of coupled Global Circulation Models (CGCMs) is currently restricted by the relatively short (3-D) ocean data record, making validation of such models for operational use difficult. In contrast, the relatively long-term archives of SST analysis (e.g. Rayner *et al.* 1996; Reynolds and Smith 1994) may be used to run hindcasts for validating prediction systems using persistence of SST anomalies (SSTA) for boundary forcing. Prescribed forcing from persisted SSTA is likely to be a competitive (and cheap) option, at least for one season ahead, because of the usual relatively slow evolution of the SST field. Here we test the viability of using persisted SSTA for operational real-time prediction by comparing skill from hindcasts against the skill benchmarks obtained from the "perfect-SST" PROVOST simulations.

v) Evaluation of the potential value of seasonal predictions to users
For operational seasonal prediction to be a viable concern, it will be necessary to establish that potential customers will be able to extract financial benefit from the forecasts, given the levels of technical skill that are available. To this end we outline a methodology for estimating user value. The method is based on the user cost/loss matrix associated with the outcome of probabilistic and deterministic predictions of specified weather events.

2. Experimental details

2.1 The PROVOST simulations

The PROVOST simulations are used to assess potential AGCM skill. They comprise 4-month ensemble integrations for each season in the 15-year period December 1979 to March 1994. The 4-month integration periods were specified as: March to June (MAMJ - northern spring); June to September (JJAS - northern summer); September to December (SOND - northern autumn); and December to March (DJFM - northern winter). Each season was simulated with 3 different AGCMs; the UKMO Unified Model (UM) run at climate resolution (3.75° long., 2.5° lat., 19 levels); the ECMWF T63 L31 model (referred to here as T63) and the Météo-France ARPÈGE T42 L31 model (referred to as AP1). A fourth set of integrations were run, for the winter season only, by Electricité de France (EDF) using the ARPÈGE model at T63 L31 truncation (referred to as AP2). The UKMO integrations were made using version 3.4 of the UM with HADAM2b physics; the performance of this version of the model has been discussed in Hall *et al.* (1995).

Integrations with each model were made in 9-member ensembles initialised with the 12 UTC analyses

from the ECMWF Re-Analysis (ERA - Gibson *et. al.* 1997) on the 9 consecutive days before each season. Observed values of Sea Surface Temperature (SST) and ice cover, corresponding to a "perfect" prediction of SST and ice cover, were used for lower boundary forcing, with values updated at 5-day intervals during the integration. The observed SSTs and ice cover were obtained from the UKMO Global sea-Ice and Sea Surface Temperature (GISST) analyses (Rayner *et al.* 1996) up to October 1981 and the Reynold's OI analyses (Reynolds and Smith 1994) for the remaining period.

The multiple-model configurations studied are an 18-member combination of the UM and T63 ensembles (referred to as JT2), a 27-member combination of the above two models and the AP1 (referred to as JT3) and a 36-member combination of all four models (referred to as JT4, and available for winter simulations only).

Simulated monthly and seasonal averages are derived from daily model values, valid at 12 UTC. In order to correct (*a posteriori*) for model bias, simulated anomalies are calculated by subtracting the model climate (defined over all 9 ensemble runs and all 15 years) from the individual ensemble fields, while the observed anomalies are derived from the ERA 15-year climatology.

2.2 AGCM hindcasts using persisted SST anomalies

Hindcasts experiments have been performed for twelve of the PROVOST DJF and MAM periods (1982-1993), for the UM only, using persistence forecasts of SST (rather than observed SST) to force the model lower boundary. The persistence SST forecasts are produced by adding one-month SSTA from November (DJF hindcasts) and February (MAM hindcasts) to the GISST or Reynolds OI climatological SST fields. The evolution of sea-ice is represented in the following way; where there is open sea initially, ice forms if SST_{clim}+SSTA falls below -1.8°C (personal communication, Davey). Thus a warm SST anomaly delays, while a cool anomaly hastens, the formation of sea-ice.

2.3 Verification

Assessments have been performed for 850 hPa temperature and precipitation using a variety of skill measures. However, emphasis is given to assessing potential probabilistic skill for the following events;

- 3-month mean 850 hPa temperature above/below normal,
- 3-month mean total precipitation above/below normal,

where we consider chiefly the 3-month mean over the first three months of the simulations.

Temperature and rainfall are selected for evaluation because of their interest to a wide range of users. Note that 850 hPa temperature anomalies may generally be considered a proxy for surface temperature anomalies. It is recognised that to be of benefit to many applications, skill at predicting higher threshold events on these variables (rather than just the sign of the anomaly) will need to be proven. However, evaluation for "above/below" events is considered a necessary first step - and skill at this level may be of practical use to some users. In addition to the event-based verification, conventional evaluations (e.g. anomaly correlation) of ensemble-mean 850 hPa temperature are also presented. Four main assessment areas are employed; the tropics, 30°N to 30°S; the northern extratropics, 20°N to 80°N; North America, 130°W to 60°W, 30°N to 70°N and Europe, 12.5°W to 42.5°E, 35°N to 75°N (Fig. 1).

Both the PROVOST simulations and the hindcasts are verified using the ERA dataset - both for 850 hPa temperature and precipitation. ERA precipitation is based on the accumulation over a 24-hr forecast run from the ERA analyses. The use of model-based precipitation analyses is not ideal, for

obvious reasons, but should be sufficient to provide large-scale estimates of potential skill for "above/below" events.

In this study, output from the T63, AP1 and AP2 simulations, archived at ECMWF, were interpolated onto the UM model grid (3.75° long., 2.5° lat.) prior to analysis.

3. Assessments of potential skill

3.1 Skill assessments for 850 hPa temperature

In this section we present verifications of the simulated anomalies in average 850 hPa temperature over the first three months of the PROVOST integrations. Results from a number of different skill diagnostics are compared with the aim of building confidence in generalised assessments of skill. Probabilistic skill is assessed using the Relative Operating Characteristic (Stanski *et. al.* 1989) and verification of the most probable anomaly sign (referred to as COMPAS scores - see Section 3.1.2). Deterministic (ensemble-mean) skill is evaluated using temporal correlation and spatial anomaly correlation scores. For brevity, discussion is focused on results from the UM simulations, with comparisons between the UM, T63, and multiple-models (JT2, JT3 and JT4) provided in summary form.

3.1.1 Probabilistic skill

The Relative Operating Characteristic (ROC) for a specific event is expressed in the form of a curve plotting hit rates against false alarm rates for the event over a range of forecast probability thresholds. The probability thresholds considered are, nominally, 0%, 20%, 40%, 60% and 80%. In practice the thresholds are defined according to the numbers of ensemble members which predict the event; threshold definitions for the 9-member ensembles are given in Table 1a, with an analogous procedure applied for the multiple-model ensembles. Note that the hit and false alarm rates, for each probability threshold, are defined as proportions of the observed frequencies of the event and non-event respectively (Table 1b). ROC evaluations of the PROVOST simulations for the four assessment regions have been constructed by calculating hit and false alarm rates over the spatial/temporal domain represented by all grid points in the region and all 15 PROVOST years.

For UM MAM simulations of the event 850 hPa temperature below normal, hit rates exceed false alarm rates for all threshold probabilities (20%, 40%, 60% and 80%) of the event in all four assessment regions (Figs 2a-d), indicating that the ensemble has skill in detecting the event both in tropical and extratropical regions. Skill is greatest in the tropics (Fig. 2a) where the hit rate/false alarm rate ratios are largest. Skill over Europe (Fig. 2c) is generally comparable to that of both the northern extratropics as a whole (Fig. 2b) and is somewhat greater than found for North American region (Fig. 2d).

Note that the greater the skill of the ensemble, the more the ROC curve must bow up towards the top left corner; the point (0,1) corresponding to perfect deterministic skill (i.e. all members correctly predict the event over all forecasts), and points on the diagonal corresponding to no skill (i.e. skill is no better than that available from a climate or random forecast). Thus the area under the ROC curve provides a useful overall index of skill; a value of 0.5 (the area under the diagonal) or less indicating no skill, and a value of 1 indicating perfect deterministic skill. The area under the ROC curve will be referred to hereafter as the ROC score. ROC scores for the tropics, northern extratropics, Europe and North America (Figs 3a-d) are 0.7, 0.61, 0.63 and 0.58 respectively. ROC scores for the event, 850 hPa temperature above normal, are identical to those for the below normal event (because the events are complementary) and are not shown.

Seasonal differences in ROC scores, and differences in performance of the UM, T63 and JT2 ensembles are compared in Figure 3. For all three ensembles, the ROC score exceeds the 0.5 threshold for skill in all four regions and for all seasons (except for the T63 SON simulations over Europe). Thus potential skill is available for all four seasons both in the tropics and in extratropical regions, including Europe and North America. Little skill variation with season is evident in the tropical region (Fig. 3a) with all three ensembles achieving ROC scores of order 0.7. In the northern extratropics (Fig. 3b) the ROC scores for all seasons/models are lower than in the tropics, at order 0.6. There is evidence, most notably in the T63 simulations, of a spring (MAM) maximum and autumn (SON) minimum in skill. Branković *et al.* (1994) have also found skill in the northern extratropics a maximum in spring (based on a 5 year set of 3-member ensemble mean simulations). Note that of the three ensemble configurations, the JT2 ensemble achieves the best overall performance, achieving in each season of both regions a ROC score equal to or better than that of the better individual model (the UM appears better in the tropics, while the T63 has the better performance in the northern extratropics).

Seasonal and model differences in ROC score are more pronounced in the regional areas of Europe and North America (Figs 3c&d). Over Europe the ROC score for all three ensembles is at a maximum in spring (MAM) and a minimum in autumn (SON) - as found for the northern extratropics. In contrast, peak scores for North America, in all three ensembles, occur in the winter (DJF) season. Enhanced skill over North America in winter may reflect higher predictability in winters with PC/W events (see Section 4). The season with lowest skill over North America differs between the models (summer for the UM; autumn for the T63).

Comparisons of individual model performance indicate that the UM performs better over Europe, while the T63 is more skilful over North America. The better model in each region achieves ROC scores comparable to those obtained for the northern extratropics (e.g. MAM over Europe and JJA over North America (Figs 3c&d)), indicating potential gains in capability from the use of more than one AGCM in an operational environment. Note that the JT2 ensemble appears to act as a filter for the more skilful individual model, and thus provides a means of exploiting the strengths of each model. Even where differences in the ROC score between models is relatively large (see again MAM over Europe, JJA over North America (Figs 3c&d)), the value achieved by the Joint JT2 system is similar to that obtained by the more skilful model. The main exception being the autumn season over Europe, where ROC score for the T63 simulation is below the 0.5 threshold for skill.

A corresponding analysis, including all 4 PROVOST individual models and the multiple-model combinations JT2, JT3 and JT4 is provided in Fig. 4 for simulations over Europe and North America. When individual model skill is at similar levels the multiple-models provide improved skill (e.g. JJA simulations over Europe, Fig. 4a), which may derive both from the presence of additional models and from the increased ensemble size. However, the most striking benefit provided by the multiple-model ensembles is the skill filtering property in regions/seasons when skill with the individual models varies widely, in this respect the benefits noted for JT2 are usefully extended by the JT3 and JT4 ensembles. Note, for example, the DJF simulations over North America (Fig. 4b) for which each multiple-model (JT2, JT3 and JT4) consistently matches the skill of its most skilful individual component model; T63 for JT2, AP1 for JT3 and AP2 for JT4 - resulting in best overall skill (out of the multiple-models) for JT4. Similar results are found for the DJF simulations over Europe (Fig. 4a), where the higher skill of the UM and AP1 models is matched by the JT4 ensemble, despite relatively lower skill from the T63 and AP2. The fact that JT4, a 36-member ensemble, provides similar skill to the best 9-member individual ensemble clearly indicates that the increased ensemble size plays only a small role in producing the relatively high skill of the JT4 - the key element in determining the skill of the multiple-model appears to be the skill of the most skilful component ensemble. Of course, it may be possible to duplicate the skill filtering effect with fewer than 9-members of each individual ensemble. These

two examples show that the relative skill of individual models may differ markedly with season and region and that the multiple-model technique can usefully exploit the strengths of all individual models. Moreover, because of its skill filtering properties, the multiple-model improves potential capability without the need for *a priori* identification of the strengths of the individual component models.

Global spatial plots of ROC score, generated by obtaining ROC curves for each model grid point from the 15 available simulations, are provided for the UM simulations in Figures 5a-d. The plots give further insight into the spatial variation of skill, though, because of the smaller sample size (15), individual values should be treated with caution. ROC scores exceed the 0.5 threshold for skill over many parts of the globe (shading breaks in at 0.55 on Fig. 5). Consistent with the regional analysis, high ROC scores are most widespread in the tropics where the ensemble simulations frequently reduce to a single deterministic solution (e.g. consistent indication of the event in all 9 ensemble members) and maximum deterministic skill over the 15 year period is approached or achieved in many areas. Marked regional variations in ROC score are evident both in the tropics and extratropics, with the regional assessment areas (particularly North America) encompassing regions of widely different skill. In both the tropics and extratropics ROC scores are generally highest over the oceans and lowest over continental interiors.

Although ROC scores in the tropics are generally relatively high, scores fall below 0.55 in some regions. Striking examples, when ROC scores are below 0.55 over a substantial area, may be seen over the West Pacific in MAM and JJA (see e.g. Fig. 5b (JJA) when the minimum extends into parts of Indonesia). The implication is that, in this region, the model is unable to resolve above and below normal seasonal average 850 hPa temperature events, despite provision of observed SST. A similar minimum is evident in ROC scores obtained with the T63 (not shown). Regions of relatively low score may be the result of local deficiencies in the model atmospheric response to SST - indicating deficiencies in the parameterisation of heat transfer processes (e.g. convection). In contrast to the West Pacific, the East Pacific region shows a much more consistent coverage of relatively high ROC scores, indicating a more satisfactory model response.

Correspondence of high ROC scores with regional weather regimes may be noted in some areas. For example highest ROC scores (exceeding 0.9) occur over India in JJA (Fig. 5b) - indicating that best skill for temperature coincides with the peak months of the south-west monsoon. In contrast highest ROC scores over Indochina and south-east Asia occur in DJF and MAM (Figs 5d&a, respectively), indicating that, for this region, best skill is found during the north-east monsoon.

Over Europe, highest scores tend to be concentrated in northern and western regions; in winter (Fig. 5d), for example, peak scores are found north of 60°N and also over the western Mediterranean area, with a distinct minimum over central areas. Over North America the higher scores are located over southern, western and northern continental fringes. Note that ROC scores over the North American interior are frequently below 0.55, indicating lower skill than over much of western Europe (notably in spring, Fig. 5a). Scores are relatively low over much of extratropical Asia, with scores above 0.55 most widespread east of 90°E and peak values occurring over the Pacific rim.

3.1.2 Verification of the most probable anomaly sign

A lower order verification (relative to the ROC score) of the probability distribution may be obtained by verifying the most probable anomaly sign, i.e. the sign indicated by the majority of ensemble members (e.g. 5 or more for the 9-member ensembles). The number of times the most probable anomaly sign is correct may then be calculated over all 15 PROVOST years for each season. This measure of skill will be referred to as the COMPAS (COrrect Most Probable Anomaly Sign) score. One advantage of the COMPAS score is that it provides an event-based verification that is readily

interpreted by non-specialists.

An example of this diagnostic is provided in Fig. 6 which shows the distribution of COMPAS scores obtained with the UM for MAM simulations of 850 hPa temperature anomalies. The distribution of COMPAS scores is well correlated with the corresponding ROC scores (Fig. 5a). Scores are highest and most widespread in tropical regions - where areas with frequencies exceeding 11 (out of 15) are substantial, and peak scores exceed 13. (Scores equal to or exceeding 11 may be considered significant at about the 95% level or greater, since the chance probability, assuming a binomial distribution, is ~6%.) Consistent with the ROC analysis, areas with significant COMPAS scores (i.e. 11 or more) are less widespread in the extratropics. However, scores exceed 11 over a number of regions including parts of western Europe and North America. COMPAS scores obtained with the T63 and JT2 ensembles (not shown) show similar large-scale features to those obtained with the UM.

The percentage geographical area for which the COMPAS scores exceed the 95% significance level is compared for all four seasons and three ensembles in Figs 7a-d. Highest coverage of significant skill, of order 35%, is found in the tropics (Fig. 7a). Peak coverage of skill over the northern extratropics (Fig. 7b) is found in spring (MAM) at order 25%, with a minimum in autumn (SON) at order 15%. Areal coverage of significant COMPAS skill appears relatively low over Europe in DJF (Fig. 7c) despite relatively high ROC scores (Fig. 3c). This latter result is consistent with the local concentration of higher ROC scores over north-western Europe and the Mediterranean (Fig. 5d) discussed previously. Over North America the areal coverage of significant COMPAS skill is highest in MAM (Fig. 7d), though the overall ROC score is highest in DJF (Fig. 3d).

Comparison of Figures 7c&d gives a further indication that useful complementary information exists in the UM and T63 performances over Europe and North America. Greater coverage of COMPAS skill is obtained with the UM over Europe and with the T63 over North America. Skill coverage is generally lower in the JT2 model; a result which arises because small displacements in the location of skill peaks between individuals models lead to a spatial smoothing of the JT2 COMPAS scores, and consequently to lower coverage above the specified threshold. Thus for some purposes, selection of the better individual model for each region is an alternative to model combining. Note however that, in contrast to the multiple-model method, strategies for exploiting complementary model information that are based on model selection would require *a priori* calibration of the strengths and weaknesses of the individual models.

3.1.3 Temporal and spatial correlation of the ensemble mean

The ensemble-mean field represents the first moment of the probability distribution and its correlation with the observed field gives an indication of basic model skill, which may be contrasted with the event-based probabilistic verification of the previous sub-Sections. The distribution of point correlations of UM ensemble-mean and observed 850 hPa temperature fields is provided in Figures 8a-d. The significance of the correlations has been estimated using a Monte-Carlo technique to estimate the probability of achieving equivalent correlations by chance (500 correlations were calculated, each after randomly scrambling the yearly order of the ensemble-mean values). Correlations are plotted only when they are significant at the 90% level or higher. As found for the ROC score and COMPAS score assessments, skill is best in the tropics where the correlation coefficient (CC) frequently exceeds a value of 0.6, with local peaks in excess of 0.8. Significant correlations are also present in the extratropics, though with lower CC values of order 0.4 with peaks to 0.6. In most regions/seasons significant correlations of the ensemble mean correspond well with regions of high ROC score (Figs 5a-d) and COMPAS score (Fig. 6 - MAM only). However there are some exceptions, notably in summer over Europe (Fig. 8b) where the CC indicates little significant skill, in contrast to the ROC score (Fig. 5b) which suggests a local peak in probabilistic skill in western regions (see end of this section for a discussion).

Average spatial anomaly correlation coefficients (AC scores) of ensemble mean values over the four assessment regions are provided for the UM, T63 and JT2 ensembles in Figures 9a-d. The averages are over the 15 PROVOST years (14 for T63 and JT2 in DJF) and are calculated using the Fisher z-transform method. AC scores are positive in all seasons and all regions, except for T63 autumn (SON) simulations over Europe. AC scores differ from zero with high levels of significance (from a t-test) in the tropics and northern extratropics in all seasons and with all three ensembles (when significance exceeds a threshold of 95% or 99%, the threshold value is plotted above the bars). Over North America significant AC scores are obtained in all seasons with all three ensembles, with best performance from the T63. Over Europe significant AC scores are obtained with at least one model in all seasons (all three ensembles in MAM; T63 only in JJA; UM and JT2 in SON and DJF). The latter result indicates that, as an alternative to multiple-model methods, selective use of two (or more) AGCMs can enhance prediction capabilities, provided the strengths and weaknesses of the individual models are established.

Seasonal and model differences in mean AC scores show broad similarities with the results of the ROC analysis. In particular skill scores are highest in the tropics (Fig. 9a - average AC score of order 0.4 - 0.5), where seasonal and model differences are small; in the northern extratropics (Fig. 9b) all three ensembles indicate best skill in MAM (AC scores are of order 0.2-0.3); over Europe skill is most consistent over the three ensembles in MAM and DJF, with scores similar in both periods, and comparable to the northern hemispheric values; over North America all three ensembles indicate best skill in DJF.

Benefits from the JT2 ensemble are evident in the tropics (Fig. 9a) and over Europe (Fig. 9c) where the AC scores obtained exceed or are similar to those of the better individual model (except for JJA over Europe). In particular the JT2 achieves a score in SON over Europe similar to that of the UM (and significant at the 99% level) despite the negative correlation obtained with the T63 ensemble. Similar benefits are not evident over the North American region (Fig. 9d) where the better individual model performance obtained with the T63 ensemble is not matched by the JT2.

Although, as described above, ROC and AC scores give similar results in most cases, different results are found for some seasons/regions. For example, the UM AC score over Europe in SON is comparable with that obtained for DJF and MAM (Fig. 9c), whereas the ROC score indicates a seasonal minimum in UM skill for SON (Fig. 3c). Another discrepancy occurs for the JJA period, for which a marked minimum in AC score occurs over Europe (Fig. 9c, and also for the point correlations in Fig. 8b) which is not matched with a corresponding minimum in the ROC score (Figs 3c&5b). Differences in AC and ROC skill assessments may be expected, since the former is a measure of the phase correspondence between two fields while the latter measures skill for threshold values in the field. Differences may also rise through the greater information content of the probabilistic format. Ensemble members with positive correlations are present in all seasons and all years, including years in which the ensemble mean is negatively correlated (Fig. 10). Thus years with zero or negative AC score may still contribute to positive probabilistic skill. A notable example is DJF 1989/90 (Fig. 10d), for which the majority of members are positively correlated (best member approaching a value of 0.8) while the ensemble mean is negatively correlated. For the case of average JJA skill discussed above, note that 5 of the 15 years have negative ensemble-mean AC scores (Fig. 10b), however in each of these years the ensemble contains members that are positively correlated, sometimes with relatively large coefficients (e.g. 1980, 1981 and 1984). Thus there is information regarding the potential for probabilistic prediction, which may contribute to a higher ROC score, that is not present in verifications of the average ensemble-mean anomaly correlation coefficient.

3.2 Skill assessments for precipitation

For reasons of brevity, we restrict diagnosis of skill for precipitation to results obtained with the ROC analysis. ROC curves for UM MAM simulations of the event 3-month mean total precipitation below normal are provided in Figures 11a-d (ROC scores for the complementary event "above normal precipitation" are identical and are not shown). Skill above that of a random or climate forecast is evident in all four regions (i.e. ROC scores exceed the 0.5 threshold). Highest scores are found for tropical regions (0.62). The score for Europe (0.55) is comparable to that of North America (0.54) and the northern extratropics as a whole (0.56). Hit rates exceed false alarm rates for all probability thresholds in the tropics; however, in the northern extratropics and two sub-regions hit rates and false alarm rates at the 20% and 80% thresholds are similar (intersections lie close to the "no-skill" diagonal). In all regions (and seasons, see below) the ROC score is lower than for 850 hPa temperature (compare Figs 2a-d). A lower level of predictability for precipitation is to be expected, since its production is sensitive to a greater range of "chaotic" processes than act on the 850 hPa temperature field.

Seasonal differences in ROC score for precipitation, and differences in performance of the UM, T63 and JT2 ensembles are compared in Figures 12a-d. Little skill variation with season is evident in the tropical region (Fig. 12a), with all three ensembles achieving ROC scores of order 0.6. In the northern extratropics (Fig. 12b) the ROC scores for all seasons/models are lower than in the tropics, at order 0.55, with an indication of enhanced skill in DJF and MAM relative to JJA and SON. As noted for the 850 hPa temperature simulations, the JT2 ensemble achieves the best overall performance, obtaining in most seasons of both regions a score similar to or better than that of the better individual model.

Differences in performance between seasons and models are more pronounced over Europe and North America (Figs 12c&d). In both regions ROC scores are highest in spring and winter, with values comparable to the northern hemispheric values. Over Europe (Fig. 12c) scores in winter and spring are generally similar, while over North America (Fig 12d) DJF appears the season of highest skill (as found for 850 hPa temperature). In both regions scores for JJA and SON are below the hemispheric values, notably over Europe where for both seasons scores fail to exceed the 0.5 threshold for skill in two of the three ensembles. On average there is little difference in the performance of the UM and T63 ensembles. Note, however that, with the exception of the relatively low skill seasons (JJA and SON) over Europe, the ROC score achieved by the JT2 ensemble exceeds or is similar to that obtained by the more skilful individual model.

Global spatial plots of ROC score are provided for the UM simulations in Figures 13a-d. To reduce "noise" in the observed and simulated precipitation fields, both have been smoothed over 3x3 grid-point boxes (7.5° lat. by 11.25° long.) prior to calculation of the ROC curves. The distribution of skill is broadly similar to that found for 850 hPa temperature (c.f. Figs 5a-d), however skill coverage is generally lower, particularly at higher ROC scores (e.g. 0.8 and above). As for 850 hPa temperature, highest scores (exceeding 0.9 in places) occur in tropical regions. However, scores greater or equal to 0.55 (scores exceeding 0.5 indicate skill) are present over many parts of the extratropics with notable peaks in some regions (e.g. over north-western Europe in MAM).

In some tropical regions high ROC scores (order 0.9) correspond with the local wet season, indicating substantial potential for seasonal rainfall prediction. Examples are north-east Brazil in MAM (Fig. 13a, relatively high ROC score values are also present over parts of the tropical north-east coast in JJA (Fig. 13b) and DJF (Fig. 13d)); equatorial Africa (notably the Guinea coast) in JJA (Fig. 13b), implying model skill in simulating the average activity/location of the ITCZ; and the Indian sub-continent in JJA (ROC score ~0.7) - implying skill in simulating average monsoon rainfall in this region. Evidence of potential model skill in these regions has led to production of experimental real-

time predictions (see e.g. Harrison *et al.*, 1997a&b and Evans *et al.*, 1998a). Skill, albeit at a lower level (ROC scores ~0.6-0.7), also appears present for the East Asian south-west monsoon in JJA over Indochina and southern China (Fig. 13b). Over Malaysia and Indonesia highest ROC scores are found for JJA and SON, perhaps indicating better skill with rainfall associated with the south-west monsoon (May to September) than that of the north-east monsoon (November to April). Relatively high ROC scores are present over parts of Australia in all seasons. However, simulation of the wet season over tropical northern Australia (DJF) appears relatively unskilful (Fig. 13d). Other tropical regions with relatively high ROC scores include parts of Central America and the Caribbean, for which ROC scores are relatively high over much of the year.

Marked regional and seasonal variations in ROC score are also present in the extratropics. Over North America highest scores are found over southern and western regions, with the notable exception of DJF when relative high scores (order 0.7) are also present over central and north-western regions. Over Europe, as noted above, highest scores (order 0.8) are found in north-western regions (northern UK and Scandinavia) in MAM. The caveat that these scores are based on a small sample should be emphasised, however they serve to indicate where regional enhancements in skill may be expected.

4. Prospects for skill prediction

Although 15-year average ensemble mean AC scores (Fig. 9, 850 hPa temperature) are below values usually considered useful for medium-range NWP (a threshold of 0.6 is frequently quoted for instantaneous fields), particularly skilful years are present in the timeseries (Fig. 10). In MAM and DJF (Figs 10a&b), for example, AC scores exceed an (arbitrarily selected) value of 0.5 in 4 and 6 years respectively. However, the presence of such skilful years is only of value if the occurrence of relatively high skill can be predicted in advance. In this section we examine the prospects for skill prediction. We first consider the impact of large amplitude PC/W events on global-scale predictability, secondly we consider the value of both PC/W events and internal ensemble spread as predictors of skill over North America and Europe. Analysis is restricted to 850 hPa temperature.

4.1 The impact of PC/W events on predictability for SON, DJF and MAM

To assess the global impact of PC/W events on seasonal predictability we first use the COMPAS score introduced in Section 3.1.2. COMPAS scores have been calculated separately for the 5 DJF periods with PC/W events; 1982/3, 1986/7, 1991/2 (PW - El Niños) and 1984/5 and 1988/9 (PC - La Niñas), and for the 5 SON/MAM periods preceding/following these DJF periods when substantial SST anomalies are also present (Fig. 14). Note, however that SST anomalies are also substantial in two SON periods (1985 and 1987) that are not classed as pre El Niño or La Niña. It was found that removing these years from the non-PC/W subset made little difference to average UM anomaly correlation scores for non-PC/W years, and on this basis - although SST anomalies are relatively high - these two SON periods were retained as non-PC/W seasons. The JJA period was not included in the analysis, because SST anomalies in JJA are usually of relatively smaller magnitude (e.g. as in 1986 when there was a transition from PC to PW) and because association of a single JJA season with each PC/W event is less straight forward (i.e. SST anomalies may be of similar magnitude in JJA periods prior and post the DJF peak).

The response to PC/W events, as measured by the COMPAS score, was found to be similar for the UM, T63 and JT2 ensembles, thus for brevity we refer here only to the UM response. To standardise the COMPAS scores in the 5 PC/W and 10 non-PC/W periods, the score is assigned a significance value, defined as the complement of the probability of achieving an equivalent score with a random forecast (assuming a binomial distribution). For example, the probability of a random forecast achieving the correct anomaly sign in all 5 PC/W years is 0.03, thus a COMPAS score indicating all 5 PC/W years correct is assigned a significance value of 0.97. It should be recalled that although the

5 PC/W events covered (for all four seasons) by the PROVOST dataset represents an increase on the number available to previous numerical studies of predictability, the sample size is nevertheless relatively small and results are therefore tentative.

Marked differences in UM COMPAS score significance (hereafter COMPAS significance) are apparent between PC/W years and non-PC/W years in all three (SON, DJF and MAM) periods (compare upper and lower panels of Figs 15-17), indicating that PC/W forcing has a marked global impact on predictability. Considering first the non-PC/W years in the tropics, the main centres of predictability appear located over the equatorial tropical Atlantic in SON (Fig. 15, lower panel), over the Indian Ocean/South-east Asia/West Pacific region in DJF (Fig. 16, lower panel) and distributed more evenly across the tropical oceans in MAM (Fig. 17, lower panel). During PC/W events the above sequence appears disrupted, with the main centre of predictability located over the East Pacific/northern South America/northern tropical Atlantic region in all three seasons, with little evidence of the centres of predictability found over the equatorial tropical Atlantic (SON) and Indian Ocean (DJF) in non-PC/W years.

Examples of the impact of PC/W forcing on predictability in the extratropics may be seen for both the North American and European regions. COMPAS significance reaches the 95% level over substantial parts of north-western North America in PC/W DJF periods (Fig. 16), compared with non-PC/W years when significance values are rarely greater than 80% - suggesting enhanced predictability of the Pacific North American pattern in PC/W winters. Barnett *et al.* (1997), also find higher predictability of the winter PNA during strong tropical Pacific SST events. Enhanced predictability in PC/W DJF periods and following MAM periods is also suggested for Mexico and the southern states of the USA, notably in MAM (Fig. 17). Little difference in predictability between PC/W and non-PC/W years is evident over Europe in SON and DJF. However, enhanced predictability in PC/W years is indicated for MAM simulations over mainland Europe, for which COMPAS significance at the 95% level is most widespread following a PC/W peak (compare upper and lower panels of Fig. 17).

In comparing levels of skill over Europe and North America it is interesting to note from Figure 16 that predictability over North America and Europe in DJF appears similar in non-PC/W years (lower panel - coverage of COMPAS significance exceeding 80% is small in both regions), whereas predictability is somewhat greater over North America in PC/W years (upper panel). The enhanced skill over North America in PC/W DJF periods (which appears related to greater predictability of PNA) will impact on the skill evaluations over all 15 years presented in Section 3, and may in part explain why best skill for this region is found in DJF, in contrast with Europe (and the northern extratropics as a whole) for which best skill is found in MAM.

In SON (Fig. 15) the main PC/W impact indicated for the North American region is higher predictability over the extreme north-east of North America in PC/W years (upper panel). Predictability over Alaska in this season appears relatively high in both PC/W and non-PC/W years (consistent with high ROC scores in this region, Fig. 5c). For the European region the main suggested impact is enhanced predictability over the high latitude North Atlantic (over and to the north of Iceland) in non-PC/W years (Fig. 15).

It is interesting to consider Figures 15-17 in terms of the temporal changes in predictability across the three periods. In PC/W years (top panels) the region of highest predictability over the tropical east Pacific in SON spreads eastward by DJF to parts of northern South America and the tropical Atlantic. Between DJF and MAM there is a marked increase in predictability over parts of Europe and North Africa (upper panels) that is less evident in non-PC/W years (lower panels) - suggesting a further, north-eastward, translation of predictability. The region of predictability centred over the western Mediterranean in MAM is particularly suggestive of a north-westward extension of predictability

evident over the northern tropical Atlantic in DJF. Note also the increase in predictability over southern Africa across the three periods, reaching a maximum in MAM.

A further assessment of the impact of PC/W events on UM predictability over Europe and North America is provided using anomaly correlation scores (Figs. 18&19). Scores for months 2-4 of the simulations, in addition to months 1-3 are also considered. Consistent with results from the COMPAS significance analysis, enhancement of UM month 1-3 AC scores during PC/W years peaks in DJF over North America and in MAM over Europe (Figs 18a&b). Over Europe (Fig. 18a) the AC scores in PC/W years increase throughout the SON-DJF-MAM period (order 0.1, 0.3 and 0.4 respectively), while the score for non-PC/W years remains relatively constant (at order 0.2). Note that in MAM the average score in PC/W years (order 0.4) is significantly different from zero (at the 95% level) while the average for non-PC/W years is not significant. The only other AC score (over both PC/W and non-PC/W years) significantly different from zero over Europe occurs for non-PC/W SON seasons. Note that significance is not always preserved when the sample is divided into PC/W and non-PC/W years. For example, the 15-year average score for UM DJF simulations over Europe is significant at the 95% level (Fig. 9c), however averages over the PC/W and non-PC/W sub-samples are not found significant, Fig. 18a).

Over North America (Fig. 18b) the AC score in PC/W seasons increases from order 0.3 in SON to a maximum (of order 0.5) in DJF, in both periods the score is significantly greater than zero (at the 95% and 99% levels, respectively). In contrast the AC score for non-PC/W years is of order 0.1-0.2 (and not significant) in all three periods. The average anomaly correlation for PC/W seasons is lowest in MAM, when scores are similar to those of non-PC/W years (order 0.2). This is in direct contrast to the European area, where MAM appears the period of maximum enhancement from PC/W forcing (Fig. 18a). Note that (as indicated from the COMPAS analysis), skill over North America and Europe is similar in non-PC/W DJF periods (AC scores are in fact slightly better over Europe), but greater over North America in PC/W DJF periods (compare Figs 18a&b). The greater predictability often attributed to the North American region relative to Europe would thus appear to apply during PC/W DJF events only.

AC scores for months 2-4 of the UM simulations are provided for Europe and North America in Figs 19a&b respectively. There is an indication that while average skill for months 2-4 in non-PC/W years decreases relative to months 1-3, average skill in PC/W years is maintained into the later period.

Corresponding results from the T63 and Joint JT2 model are similar in most respects to those presented above for the UM and are therefore not shown. In particular all three ensembles indicate best skill enhancement in PC/W years over North America in DJF and over Europe in MAM.

4.2 Relationship between ensemble spread and ensemble-mean skill

Prediction of model skill may also be approached through the relationship between ensemble spread and the skill of the ensemble mean. For a given case, the spread of ensemble members about the ensemble mean is a measure of the sensitivity to initial conditions, and thus should allow assessment of the intrinsic predictability; low spread ideally associating with high ensemble-mean skill. In this section we evaluate ensemble skill/spread relationships for simulations over Europe and North America. In order to investigate possible links between ensemble spread and PC/W events, emphasis is given to analysis of the seasons for which PC/W events appear to have most impact, i.e. MAM over Europe and DJF over North America.

4.2.1 Europe

Scatterplots of ensemble-mean anomaly correlation and ensemble spread (defined as the average, Fisher-z transformed, anomaly correlation of the ensemble members with the ensemble mean) for

MAM and DJF simulations of 850 hPa temperature over Europe are provided in Figs 20&21 for the UM, T63 and JT2 ensembles. A measure of ability to distinguish relatively skilful from unskilful predictions is provided by comparing the total entries in the diagonal quadrants with the total in the off-diagonal quadrants (here the quadrants are constructed using the ensemble median values of AC skill and spread). A positive skill/spread correlation, and thus potential for skill prediction, is indicated if entries are maximised in the diagonal quadrants. The linear correlation is also provided; however, because of the sensitivity of linear correlation to outlying points, the non-linear measure described above is preferred. Years corresponding with PC/W events are indicated with symbols, and the identity of other individual years is denoted with letters.

For MAM simulations, the UM ensemble (Fig. 20a) shows clear potential for distinguishing relatively skilful and unskilful predictions. A total of 11 simulations fall in the diagonal quadrants representing correct assessments of ensemble-mean relative skill, and 2 fall in off-diagonal quadrants, representing incorrect assessments. Hereafter we express this measure of the skill/spread correlation as the ratio "diagonal entries/off-diagonal entries", i.e. in the above example, 11/2; when diagonal entries are greater/fewer than the off-diagonal entries we refer to positive/negative correlation. Note that the total of correct and incorrect assessments may vary because simulations with skill or spread equal to the median value (of which there may be more than one) are not counted.

Recall that for Europe MAM is the season for which AC scores are most enhanced, on average, during PC/W events. It is interesting, therefore, to compare AC scores for individual PC/W and non-PC/W years. Of the four UM simulations (Fig. 20a) with highest AC scores (1980, 1987, 1989 and 1990 all have scores of order 0.6 or above) two are in PC/W years (1987 and 1989). The simulation in PC/W year 1992 also has above median skill (and has the fifth highest AC score). However, the presence of a PC/W event does not guarantee skill, with below median scores obtained in the PC/W years 1983 and 1985, the score in 1985 being amongst the lowest (order -0.2) achieved in all years. Conversely, relatively high scores are achieved in two non-PC/W years (1980 and 1990). Note that the three PC/W simulations with above median skill are also associated with below median spread, while the two with below median skill are associated with above median spread. Thus there is an indication that ensemble spread may be useful for discriminating PC/W years with likelihood of either relatively high or low skill.

In contrast to the UM, the T63 MAM simulations (Fig. 20b) show a negative skill/spread correlation (4/8). However, the enhancement of AC scores in PC/W years appears somewhat more consistent than for the UM - with 4 of the 5 simulations in these years achieving above median skill (compared to 3 out of 5 for the UM). Note that the JT2 ensemble (Fig. 20c) retains much of the ability for skill prediction exhibited by the UM simulations, achieving a skill/spread correlation of 9/3.

Figures 20a-c may be used to indicate potential strategies for skill prediction in an operational environment. For illustration we assume a policy of issuing forecasts only in years when skill prediction strategies indicate a relatively skilful forecast; and assess the strategy on the number of forecasts correctly/wrongly identified as skilful and the number correctly/wrongly rejected as unskilful. Four potential strategies using PC/W impact and skill/spread correlation may be defined and are denoted here S1, S2, S3 and S4.

S1: PC/W impact, i.e. issue forecasts in all years with PC/W events.
S2: ensemble spread, i.e. issue forecasts only in years when the spread is lower than a threshold value. For illustration, the median value of spread is used as the threshold.
S3: the union of S1 and S2, i.e. issue forecasts in all PC/W years and also in any year when ensemble spread is below threshold.
S4: the intersection of S1 and S2, i.e. issue forecasts only in the subset of PC/W years when the

ensemble spread is below the threshold value.

For the UM, strategies S2 and S4 appear optimal for MAM (Fig. 20a). Adopting S2 (low spread only) would result in six issued forecasts. Of these six forecasts; five would have above median anomaly correlation (i.e. above ~0.2 in this case) with two particularly skilful forecasts (the PC/W years 1987 and 1989); one would be particularly unskilful (1991); and two skilful forecasts would be rejected (1980 and 1990). Strategy S4 (PC/W year *and* low spread) is a more cautious approach resulting in only 3 issued forecasts (1987, 1989 and 1992); however all three forecasts would have above median skill. Such a cautious approach may be optimum for some applications. In contrast, strategy S1 (PC/W years only) would result in 5 issued forecasts only 3 of which (the same 3 as for S4) would have above median skill, and strategy S3 (PC/W year *or* low spread) would result in the issue of 2 more forecasts than strategy S2 (in 1983 and 1985), both of which would have below median skill.

For the T63 (Fig. 20b), strategies S2 and S4 do not perform well because of the negative skill/spread correlation; S2 achieves only 2 skilful forecasts out of 7 issued, while S4 achieves 2 skilful predictions out of 3 issued). However, strategy S1 (PC/W years only) performs relatively well, achieving 4 relatively skilful forecasts out of the 5 issued. The above results confirm expectation that the optimum strategy for skill prediction will vary between AGCMs.

Positive skill/spread correlations are found in DJF simulations in all three ensembles (Figs 21a-c). For the UM (Fig. 21a), the skill/spread correlation obtained for DJF (8/6) is lower than that for MAM (11/2), while for the T63 the correlation for DJF (also 8/6) is higher than for MAM (4/8). Considerable benefits are obtained with the JT2 ensemble which achieves the highest skill/spread correlation of 10/4 (Fig. 21c). Note that although the median AC score is lower for JT2 compared with the UM and T63, the average score over all years is similar (Fig. 9c). Moreover, it is apparent from Figs 21a-c that the simulations correctly identified as skilful by the JT2 ensemble, have scores that are similar to or higher than the simulations correctly identified as skilful by the UM and T63 ensembles. Thus the average skill of the "issued" forecasts is enhanced in the JT2 ensemble.

Average skill enhancement in PC/W years over Europe is less evident in DJF than for MAM in the UM ensemble (Fig. 18a). The lower impact is evident in Figs 21a, with AC scores in PC/W years spread across the range of skill. There is no evidence of a strengthening of the skill/spread correlation in PC/W years as found for UM simulations in MAM.

A further result evident from Figs 21a-c is the distinct clustering of the DJF simulations into two groups, one with relatively high skill and one with relatively low skill (this effect is particularly evident in the JT2 simulations). Reasons for the apparent bimodal distribution of skill are not explored here but may, for example, be connected with possible lower/higher predictability associated with high/low frequency of blocking regimes (and thus might be related to the phases of the North Atlantic Oscillation). Note also that the predictable and unpredictable years are generally the same for each model; with performance in 10 of the 14 simulations in agreement; the exceptions being 1988 and 1992 (skilful UM, unskilful T63) and 1987 and 1991 (skilful T63, unskilful UM).

The skill/spread correlations found for all four seasons over Europe are summarised in Table 2. For the UM, correct skill assessments exceed incorrect assessments in all seasons except SON, while for the T63 a positive skill/spread correlation is found only for DJF. Note that benefits to the skill/spread correlations are provided by the JT2 ensemble; positive correlations obtained with the UM (the better overall individual model in this case) are improved in JJA and DJF, and a high correlation (9/3) is maintained in MAM despite the negative T63 correlation.

4.2.2 North America

Skill/spread correlations for MAM over the North American region are negative with all three ensembles (Figs 22a-c and Table 3), with values of 6/7, 6/7 and 4/7 for the UM, T63 and JT2 respectively. This is in contrast to the promising skill/spread correlations found (with the UM and JT2) for MAM over Europe. The simulations in PC/W years appear to be predominantly of low spread (notably in the T63 and JT2 simulations, Figs 22b&c), but there is a wide variance in AC scores in these years. For DJF (Figs 23a-c), however, a positive skill/spread correlation of 8/6 is achieved by both the UM and T63 models, with a notable improvement to 10/4 achieved with the JT2 ensemble. Over North America average skill enhancements during PC/W years are a maximum in DJF (see e.g. Fig. 18b), and indeed AC scores in all PC/W years are close to or above the median value in all three ensembles (Figs 23a-c). The consistently high AC scores found in PC/W years for DJF results in good performance of the S1 skill prediction strategy (issue forecasts in PC/W years only). For all three ensembles, the S1 strategy would capture 5 of the 11 predictions with positive AC scores and correctly reject the 3 years with particularly poor predictions (1979, 1980 and 1981 - correlations of order −0.2 or below). Note, however, for the T63 the S3 strategy (PC/W years *or* low spread) gives better performance than S1, capturing 8 of the 11 simulations with AC score better than 0.2 while still successfully rejecting the remaining 3 years with negative AC scores. The S3 strategy with the UM and JT2 models is less successful; the number of skilful forecasts issued is increased by 2 (1987 and 1989) but at the expense of not rejecting all of the simulations with negative AC scores (e.g. low skill UM simulations for 1979 and 1981 are not rejected (Fig. 23a)).

The skill/spread correlations found for all four seasons over North America are summarised in Table 3. Positive skill/spread correlations are obtained with all three ensembles in DJF and JJA with the best correlation achieved in both periods by the JT2 ensemble. Note that MAM is the only season in which all three ensembles have negative correlations.

Comparison of Tables 2&3 indicates that the UM achieves better skill/spread correlations over Europe, while the T63 performs better over North America. In both regions benefits to the skill/spread relationship are available from the JT2 ensemble. The JT2 improves the skill/spread correlation in 4 of the 6 cases (considering Tables 2&3 collectively) for which at least one individual ensemble achieves a positive skill/spread correlation, and achieves a similar positive correlation in one of the remaining 2 cases (MAM over Europe).

5. Use of persisted SST anomalies for lower boundary forcing

Real-time seasonal forecasting requires a forward projection of the SST evolution. The SST prediction may be supplied either through a coupled ocean/atmosphere model or by a "two-tier" system in which independent forecasts of SST are used as prescribed forcing to AGCMs. One relatively cheap option for the two-tier system is to use a persistence forecast of SST, in which SST anomalies (SSTA) over a period preceding the initial time of the forecast are persisted throughout the integration. In this section we compare skill obtained using persisted SSTA from the month preceding the initial time, with the skill obtained using observed SST (i.e. as evaluated for the PROVOST simulations). Details of the method used to produce the persistence SST forecasts are given in Section 2.2. The comparison has been performed for 12 of the 15 PROVOST MAM and DJF periods for 850 hPa temperature and precipitation using the ROC score discussed in Section 3.

In the tropics ROC scores obtained using persisted SST anomalies are consistently lower than, but nevertheless comparable with, the "perfect" SST skill bound provided by the observed SST runs (Table 4a). Scores with persisted SST are depressed by no more than 5%, and are above the 0.5 threshold for skill for both 850 hPa temperature and precipitation. For the northern hemisphere (Table 4a) and the European and North American regions (Table 4b), results are mixed, with scores obtained

using persisted SST equivalent or even higher in about half the observed/persisted pairs (note that in DJF persisted SST scores are equivalent or higher in 5 out of 6 pairs). Differences (positive or negative) are mainly small, except for Europe in MAM for which the ROC score for 850 hPa temperature drops from 0.59 (observed SST) to 0.46 (persisted SST). Case study investigations of individual JJA and SON seasons indicate that skill with persisted and observed SST are comparable regardless of the season.

The skill comparisons of Tables 4a&b are for a 12 year sample (1982-1993) which includes 5 MAM and DJF seasons with PC/W events (see Section 4.1). In the tropics, where SST forcing dominates predictability (see e.g. Branković and Palmer 1999 - this volume), we may expect greater than average percentage skill reduction in PC/W years if anomalies in the SST field develop rapidly after the forecast is initialised. However, lower than "average" skill reduction may be expected in non-PC/W years. In the extratropics Branković and Palmer (1999) note that the attribution of predictability to either SST forcing or atmospheric initial conditions is difficult, the dominant factor varying with region. Thus the lack of a consistent "winner" between observed and persisted SST in the northern extratropics and European and American sub-regions may be part due to lower sensitivity to SST, and part due to the sufficient accuracy of persistence SST forecasts in most years.

With the above caveats, the accuracy of persisted SST anomalies at least up to one season ahead appears, on average, sufficient to achieve skill approaching the potential estimated upper bound (with current AGCMs). We can conclude that use of persistence forecasts of SST as forcing to AGCM ensemble integrations appears a competitive (and cost effective) method for real-time seasonal forecasting, at least for a range of up to one season ahead.

6. User value of seasonal predictions

In this Section we develop a methodology for assessing the user value of forecasts, following and extending the work of Murphy (1994, 1997), and apply it to investigate the potential value of dynamical seasonal predictions over Europe. Table 5 gives the cost/loss matrix for a user wishing to act on predictions of an adverse weather event. Each of the four contingencies in Table 5, hit, miss, false alarm or correct rejection is associated with a financial impact, or loss; denoted here by L_h, L_m, L_f and L_c respectively, and explained in Table 5. For convenience the losses are measured relative to the 'normal' loss associated with a correct rejection (i.e. the event is not forecast and is not observed, so that $L_c = 0$), however results obtained without this assumption are identical. The frequency of the four contingencies must be established (through "track record" validation of the prediction system) for a range of forecast probability thresholds on the event. For any one probability threshold, these frequencies are denoted here by h (hit), m (miss), f (false alarm) and c (correct rejection). The expected mean expense (ME_{fx}) of taking action when the forecast probability exceeds a given threshold is thus,

$$ME_{fx} = hL_h + mL_m + fL_f \qquad (1)$$

Note that h, m, f and c are related to the hit rates and false alarm rates of the ROC verification method (Table 1b) by,

$$h = oHR;\ m = o(1-HR);\ f = (1-o)FAR;\ c = (1-o)(1-FAR),$$

where o is the overall frequency of the event. Thus the ROC method provides an ideal technical validation of the prediction system for use in estimating forecast value. Note also that the definition of the mean expense (1) is an extension of previous definitions in that it allows the loss associated with a hit to differ from that associated with a false alarm, rather than assigning a single loss value (equal to the cost of protection) to both contingencies. If $L_h = L_f$, then (1) reduces to the simpler form (see e.g. Richardson (1998)). The extended definition allows representation of financial benefits that may result from actions taken on advance warnings of an event which later occurs (i.e. a hit). Such benefits may offset expenditure on protection. For example, protective action for a farmer might include switching part or all of a planned seed crop to one more suited to the probable expected seasonal conditions. Resulting high yields from the new crop if the event occurs may offset the additional expenditure incurred.

The value of the forecast system may be defined as the saving the user can expect compared to the expected mean expense incurred when no forecast information is used. With no forecast the user has two options, either always take protective action or never protect. Which of these options is preferable depends on the climatological frequency of the event, o. Never protecting will incur a mean expense of oL_m , while always protecting will incur $oL_h+(1-o)L_f$. Thus the mean expense based on climatological information only is:

$$ME_{cl}=\min(oL_m\ ,\ oL_h+(1-o)L_f)$$

In financial terms the value of the forecast system is thus given by:

$$v=ME_{cl} - ME_{fx}$$

When value is used as a verification tool it is convenient to scale it relative to the maximum possible value which may be obtained by following a perfect forecast system, in which case protective action is only taken when the event occurs, with no misses or false alarms. The mean expense for a perfect system is $ME_p=oL_h$, and the scaled value is given by:

$$V=\frac{ME_{cl}-ME_{fx}}{ME_{cl}-ME_{p}}$$

This scaled value thus has a maximum value of 1 for a perfect system. For a forecast system that is no better than climate V=0; note however that there is no lower bound, and for forecasts with low hit rates or high false alarm rates (large m or f), or when large losses are associated with misses and false alarms (L_m or L_f), V may be negative.

Forecast value may be evaluated in this way for a range of forecast probability thresholds of the event and value plotted as a function of the threshold. In this way the user may select the probability threshold that provides the greatest value. Two examples based on the ROC verifications of the PROVOST simulations presented in Section 3 are provided in Figure 24. The weather event considered is spring (MAM) 850 hPa temperature above normal over Europe, each example shows potential value obtained with different sets of user losses as might pertain to two differing user applications. Value obtained from both probability forecasts and from deterministic forecasts based on the ensemble mean are considered for the three ensembles UM, T63 and JT2. The cost/loss matrix for the example of Figure 24a is defined as $L_h=1$, $L_m=4$ and $L_f=4$. The derivation of the loss values will in general be complex and depend on details of the user sensitivity and planned response to each of the four contingencies. For a simple interpretation, however, we may assume in this example that, on forecasts

of an "adverse" weather event, the user spends 4 units on protection, a sum equivalent to his loss in the event of a miss, and that financial benefits in the event of a hit offset expenditure on protection by 3 units. Figure 24a shows that, with these loss estimates, potential model skill is indeed sufficient to obtain value. Highest value in this case, equal to approximately 12.5% of the value of a perfect forecast system, is achieved by the ensemble means of both the UM and JT2 ensembles, and also the JT2 with a probability threshold of 60%. In this case the T63 ensemble achieves about 7.5% with a 60% threshold, but the ensemble mean only about 3%.

Note that for the example in Figure 24a the cheapest option based on climatology is to never protect - thus the value curves approach zero as the threshold probability for protection approaches unity and the forecast advice tends to "never protect" (i.e. the same guidance provided by climatology). For the example in Figure 24b, the loss associated with false alarms (L_f) has been reduced to 2 units. The reduction in the cost of false alarms means that the cheapest climatological option changes from never protecting to always protecting, and the overall structure of the probability value curves changes. The curves are now constrained to zero at probability threshold 0% (i.e. forecast advice is to always protect - the same guidance obtained from climatology). Comparison of Figures 24a&b shows the strong sensitivity of potential forecast value to the user losses, for example the threshold of 60% probability which gave maximum value in the first example (Fig. 24a) now gives negative value, while maximum value in the second example (Fig. 24b) is achieved with a threshold of 20%. Note that while in the first example maximum value is similar or greater with probabilistic information than with the ensemble mean, in the second example positive value is available only from the probability forecasts. This result demonstrates that issuing predictions probabilistically is likely to extend the range of applications that can extract value from the forecasts.

The above examples demonstrate that there is potential for user value from seasonal predictions over Europe. For optimum value close cooperation between forecast supplier and user appears necessary. For example the supplier will need to investigate model performance for meteorological events which are both relevant to the user and feasible on the seasonal timescale. While the user may need to revise and adapt the cost/loss estimates (e.g. through changed responses to the four contingencies) to make the most of the forecast system performance.

The above examples were deliberately chosen to demonstrate potential value over seasonal predictions in the extratropics. Clearly, potential value will be at greater levels in tropical areas where model skill is at a higher level. Making use of a similar method, with user supplied values for the cost/loss table, Harrison and Graham (personal communication) have estimated the value of model seasonal rainfall predictions over the southern African region to be at least of order US10^8 - 10^9, corresponding to a cost/benefit ratio over regional investment of order 20-200.

7. Summary and conclusions

Seasonal simulations from the 15-year PROVOST database have been analysed to assess the potential skill of 9-member ensemble integrations of the UK Met. Office Unified Model (UM) and the ECMWF T63 model for seasonal prediction. A joint 18-member ensemble (the JT2 ensemble), produced by combining all members of the UM and T63 ensembles, and higher-order multiple-model ensembles employing all four participating PROVOST models, have also been assessed. The ensembles were integrated using observed ("perfect" predicted) SST to force the lower boundary, and the skill attained therefore represents an upper bound with the AGCMs employed. For the UM ensembles, skill achieved using observed SSTs was compared to that obtained when observed SSTs are replaced with a persistence forecast of SST based on SST anomalies from the month preceding the forecast initial time. The comparisons provide an assessment of the viability of persisted SSTs for real-time prediction.

All skill measures calculated for the entire tropical and northern extratropical regions indicate that, while skill is highest in the tropics, skill is also available over the northern extratropics with all models in all seasons. Skill is present for both 850 hPa temperature and precipitation, though skill for the latter is generally at a lower level than for the former. Nevertheless, there is evidence of substantial potential for rainy season predictions in some tropical regions. In both the tropics and extratropics, highest skill tends to be centred over oceanic regions and lowest skill over continental interiors. Best skill in the northern extratropics is found in the spring (MAM) season for both 850 hPa temperature and precipitation.

Over Europe and North America skill is highest in DJF and MAM, when skill scores over both regions are comparable (apart from T63 AC scores which are considerably higher over North America than over Europe), and similar to northern hemispheric values. Peak skill occurs over Europe in MAM, as found for the northern extratropics. In contrast peak skill over North America is found in DJF, apparently as a result of enhanced winter predictability of the PNA mode during PC/W events. Skill in JJA and SON is generally below hemispheric values for both regions, most notably for precipitation over Europe (when ROC scores are close to the 0.5 threshold for skill). Spatial variations of skill within the assessment regions, both for 850 hPa temperature and precipitation, indicate that over Europe highest skill tends to be concentrated in northern and western regions; while over North America highest levels of skill are located over southern, western and northern regions. ROC scores indicate that skill over the North American interior is frequently lower than over much of western Europe.

Performance differences between the UM and T63 models are more pronounced over the regional areas of North America and Europe, the UM generally achieving better scores over Europe and the T63 achieving better scores of North America. Thus useful complementary skill is present between the individual models which could be exploited in an operational environment. In these regions the skill filtering property of the JT2 ensemble provides substantial benefits, achieving ROC scores similar to or better than the more skilful individual model, even when skill differentials between the individual models are relatively large. Benefits from the JT2 ensemble are less apparent for the AC skill scores, but notable improvements are found in some seasons. Higher order multiple-models (JT3 and JT4) show further improvements over JT2. The skill of the multiple-models appears to be mainly a function of the skill of the most skilful component ensemble, rather than being principally related to the increased ensemble size. An important benefit of the multiple-model method is that it allows improved potential capability without the need for *a priori* identification of the strengths of the individual component models, as would be needed, for example, if a strategy of choosing the best model for each region were adopted.

PC/W events in the tropical east Pacific are found to have a marked global impact on model predictability. Results with the UM indicate a transfer of centres of predictability located over the tropical Atlantic and Indian Ocean in, respectively, the SON and DJF periods of non-PC/W years, to the tropical east Pacific in both the SON and DJF periods of PC/W years. In the northern extratropics an eastward transfer of enhanced predictability throughout the SON, DJF and MAM period is suggested, with the largest enhancement occurring over North America in DJF, apparently as a result of increased predictability of the PNA pattern, and over Europe in MAM. In non-PC/W DJF periods skill over Europe and North America is similar, suggesting that the greater predictability frequently attributed to the North American region applies mainly to DJF periods with PC/W events. In non-PC/W years skill over North America and Europe is generally lower for months 2-4 than in months 1-3, however in PC/W years skill for months 1-3 appears maintained into the month 2-4 period.

Ensemble consistency also provides useful information for skill prediction in most seasons. Of the individual models best skill/spread correlations for Europe are obtained with the UM which achieves

positive (non-linear) skill/spread correlations in all seasons except SON. In contrast, the T63 has the better skill/spread performance over North America, achieving positive correlations in all seasons except MAM. In most seasons of both regions the JT2 ensemble brings substantial improvements to the skill/spread correlations attained with the individual models.

Skill prediction strategies using combinations of PC/W impact and skill/spread correlation were illustrated using the 15 year sample of the PROVOST dataset. Prospects for the development of both cautious and adventurous strategies appear promising. However, choice of the optimum strategy appears dependent on the prediction model used, the region and the season. Application of the strategies will therefore require calibration of the model skill/spread characteristics and response to PC/W events for all regions of interest.

Comparisons of integrations using persisted SST anomalies and observed SST as boundary forcing indicate that, on average, a substantial proportion of the skill achieved using observed SST is retained using persisted SST anomalies, both in the tropics and in the extratropics (though there may be local variations in the proportion of skill retained). Thus the use of persistence forecasts of SST appears a viable method for real-time seasonal forecasting, at least for a range of one season ahead.

For future development of operational seasonal prediction it will be crucial to establish the levels of technical skill (i.e. as measured using skill scores) required in order for seasonal predictions to be of value to users. A methodology for linking technical forecast quality with financial value for users has been outlined using the Relative Operating Characteristic (ROC) and the user cost/loss matrix. Examples employing an assumed user cost/loss matrix indicate promising potential for user value of seasonal predictions not only over tropical areas, but also in extratropical areas such as Europe.

Acknowledgements

The help of all other collaborators in the PROVOST project is acknowledged; specifically Tim Palmer (ECMWF and PROVOST coordinator), Michel Déqué (Météo-France) and Jean-Yves Canneill (Eléctricité de France). Discussions with Tim Palmer and Čedo Branković are also gratefully acknowledged. Thanks are extended to Ruth Evans who performed many of the UM simulations. The persisted SSTA fields referred to in Section 5 were provided by Sarah Ineson and Michael Davey of the Ocean Applications Group, UK Meteorological Office.

References

Barnett, T.P., K. Arpe, L. Bengtsson, M. Ji and A. Kumar, 1997: Potential predictability and AMIP implications of midlatitude climate variability in two general circulation models. J. Climate, 10, 2321-2329.

Brown, B.H. and A.H. Murphy, 1996: Improving forecasting performance by combining forecasts: the example of road-surface temperature forecasts. Met. Apps, 3, 257-265.

Branković, Č., T.N. Palmer and L. Ferranti, 1994: Predictability of seasonal atmospheric variations. J. Climate, 6, 217-237.

Branković, Č. and T.N. Palmer, 1999: Seasonal skill and predictability of ECMWF PROVOST ensembles. To be submitted to Q.J.R. Meteorol. Soc.

Cullen, M.J.P., 1991: Introduction to the unified forecast/climate model. UK Met. Office FR Division Scientific Paper No. 1. Available from the National Meteorological Library, London Road, Bracknell, Berks. RG12 2SZ, UK.

Evans, T., R. Graham, M. Harrison, M. Davey and A. Colman, 1998a: A dynamic one-month lead seasonal rainfall prediction for July to September 1998 for North Africa from 20°N to the equator. Experimental Long-lead Forecast Bulletin, Vol 7, No. 2, June 1998, 38-42.

Evans, R.E., M.S.J. Harrison and R.J. Graham, 1998: Joint medium range ensembles from the UKMO and ECMWF models. UK Meteorological Office Forecasting Research Division Technical Report No. 243. Available from the National Meteorological Library, London Road, Bracknell, Berks. RG12 2SZ, UK.

Gibson, J.K., P. Kållberg, S. Uppala, A. Hernandez, A. Nomura and E. Serano, 1997: ERA description. ECMWF Re-Analysis project Report Series.

Hall, C.D., R.A. Stratton and M.L. Gallani, 1995: Climate simulations with the Unified Model: AMIP runs. UK Meteorological Office Climate Research Technical Note 61. Available from the National Meteorological Library, London Road, Bracknell, Berks. RG12 2SZ, UK.

Harrison, M., T. Evans, R. Evans, M. Davey and A. Colman, 1997a: A dynamical one-month lead seasonal rainfall prediction for March to May 1997 for the north-eastern area of South America. Experimental Long-lead Forecast Bulletin, Vol 6, No. 1, March 1997, 25-28.

Harrison, M., M.K. Soman, M. Davey, T. Evans, K. Robertson, K. and S. Ineson, 1997b: Dynamical seasonal prediction of the Indian summer monsoon. Experimental Long-lead Forecast Bulletin, Vol 6, No. 2, June 1997, 29-32.

Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. Q.J.R. Meteorol. Soc., 122, pp73-119.

Murphy, A.H., 1994: Assessing the economic value of weather forecasts: an overview of methods, results and issues. Met. Apps., 1, 69-73.

Murphy, A.H., 1997: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. Mon. Wea. Rev., 105, 803-816.

Palmer, T.N. and D.L.T. Anderson, 1994: The prospects for seasonal forecasting - A review paper. Q. J .R. Meteorol. Soc., 120, 755-793.

Rayner, N.A., E.B. Horton, D.E. Parker, C.K. Folland and R.B. Hackett, 1996: Version 2.2 of the Global sea-Ice and Sea Surface Temperature data set, 1903-1994. CTRN 74, Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, RG12 2SY.

Reynolds, R.W and T.M. Smith, 1994: Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation. J. Climate, 7(6), 929-948.

Richardson, D., 1998: Skill and economic value of the ECMWF ensemble prediction system. To be submitted.

Stanski, H.R., L.J. Wilson and W.R. Burrows, 1989: Survey of common verification methods in Meteorology. World Weather Watch Technical Report. No. 8, WMO/TD 358, 114pp.

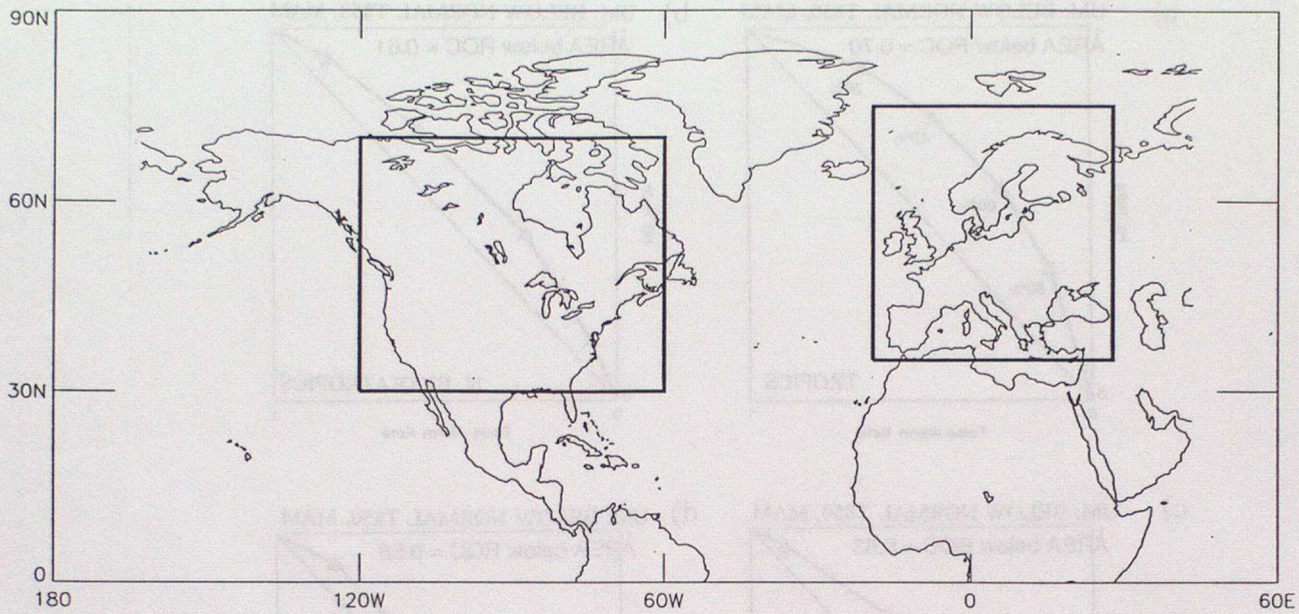


Fig. 1 The European (12.5°W to 42.5°E, 35°N to 75°N) and North American (130°W-60°W, 30°N to 70°N) assessment regions.

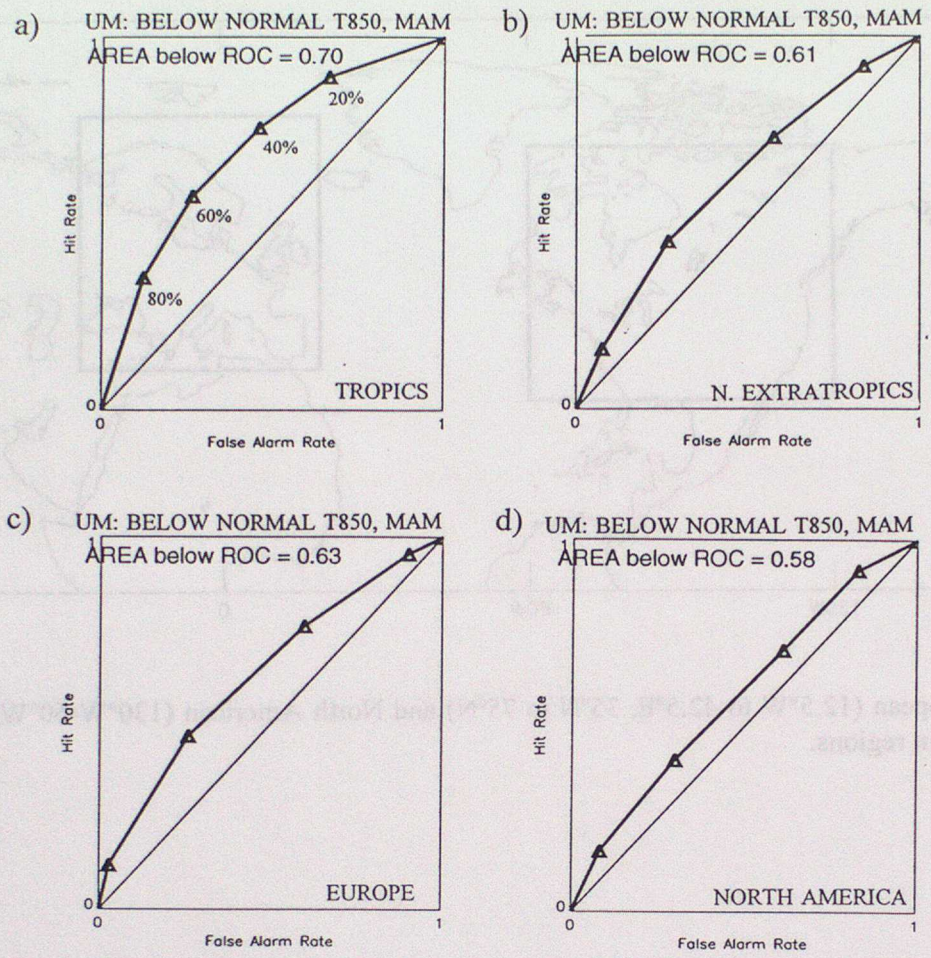


Fig. 2 Relative Operating Characteristic (ROC) curves for UM MAM simulations of the event 850 hPa temperature below normal over a) tropics, b) northern extratropics, c) Europe, d) North America. The curves are constructed from hit and false alarm rates (see Table 1b for definitions) at four thresholds on the forecast probability of the event (20%, 40%, 60% and 80%, as indicated on panel (a)). The curve is bound by the points (0,0) and (1,1) which correspond respectively to the false alarm and hit rates achieved through never and always forecasting the event. The areas under the ROC curve (the ROC score) are provided and give an overall measure of skill.

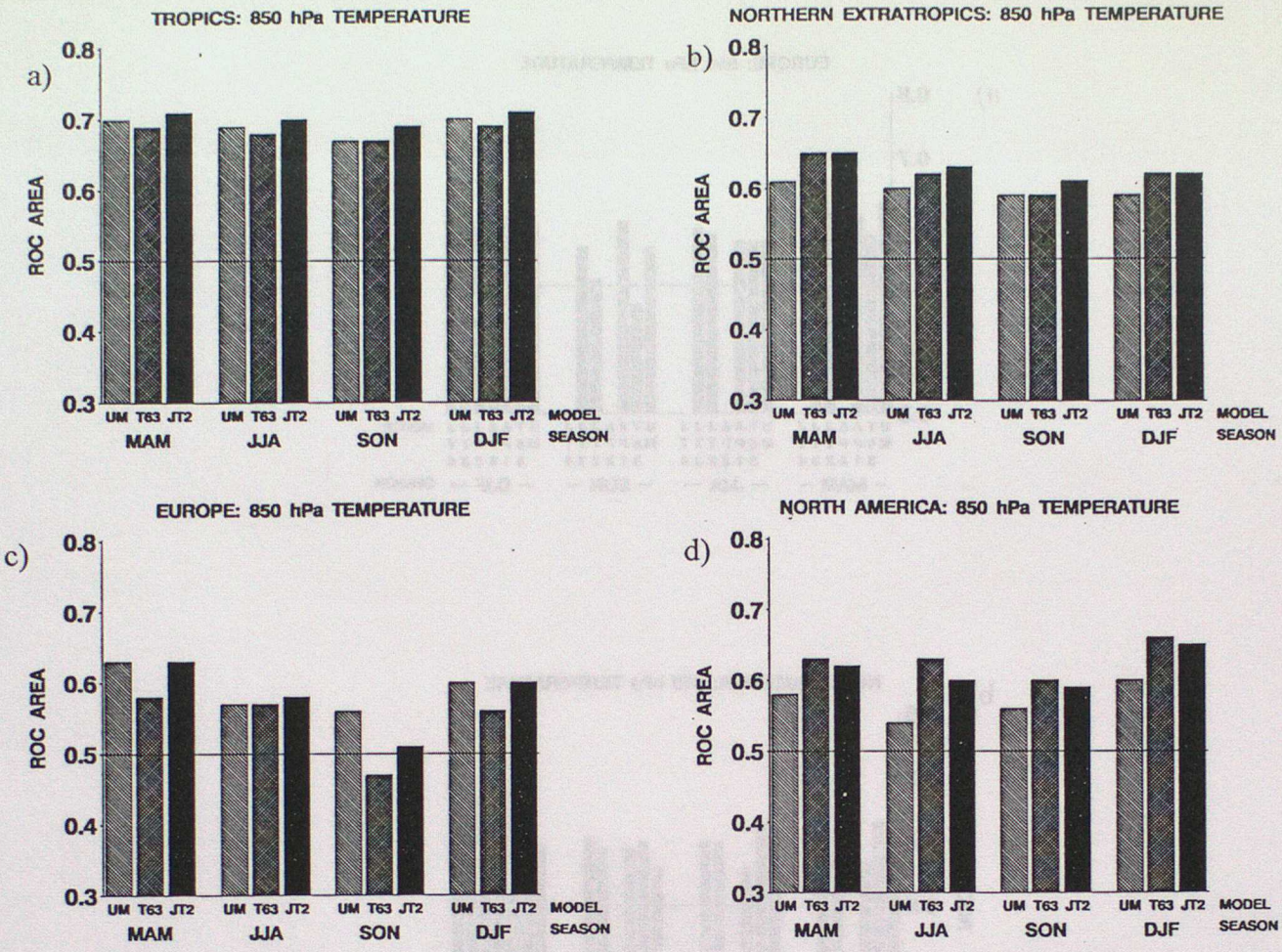


Fig. 3 ROC scores (area under the ROC curve) for the event 850 hPa temperature below normal. ROC scores for the event 850 hPa temperature above normal are identical. Results are for; UM = UKMO Unified Model ensemble, T63 = ECMWF T63 ensemble and JT2 = a combination of the UM and T63 ensembles.

a) tropics, b) northern extratropics, c) Europe, d) North America.

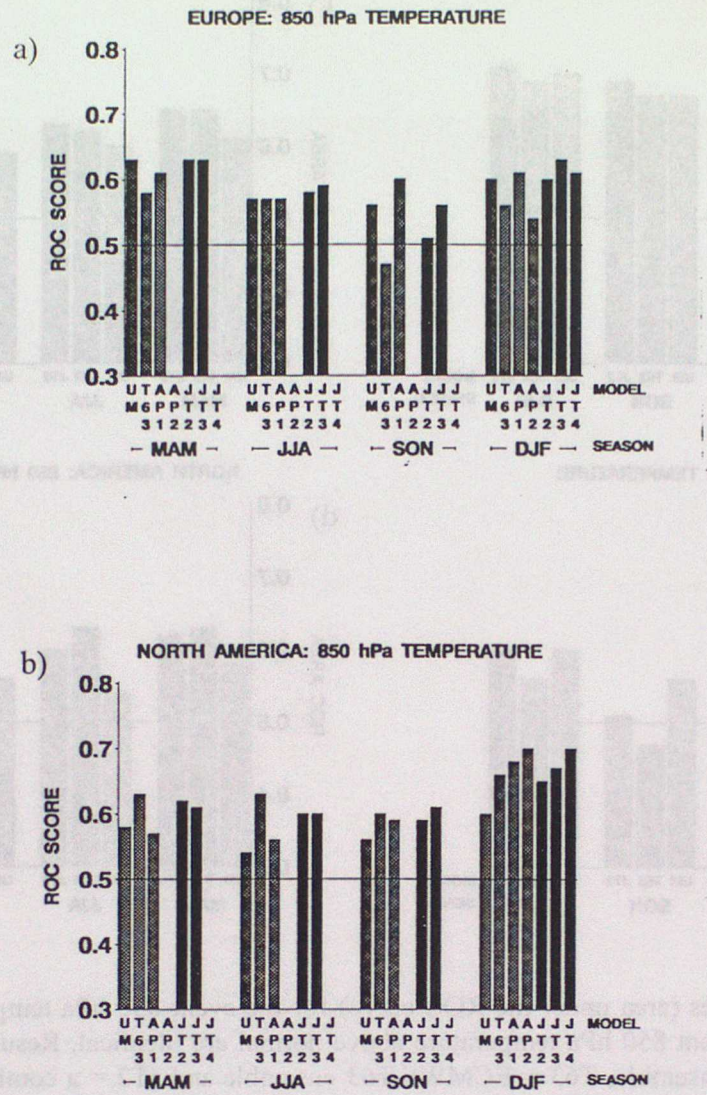


Fig. 4 As for Fig. 3 but for all four participating PROVOST models (hatched bars) and the multiple-model configurations JT2, JT3 and JT4 (solid bars).
UM = UKMO Unified Model,
T63 = ECMWF T63,
AP1 = Météo-France ARPÈGE T42 L31,
AP2 = the ARPÈGE T63 L31 (run at Electricité de France (EDF) - DJF only),
JT2 = UM+T63 (18 members),
JT3 = UM+T63+AP1 (27 members),
JT4 = UM+T63+AP1+AP2 (36 members, DJF only).

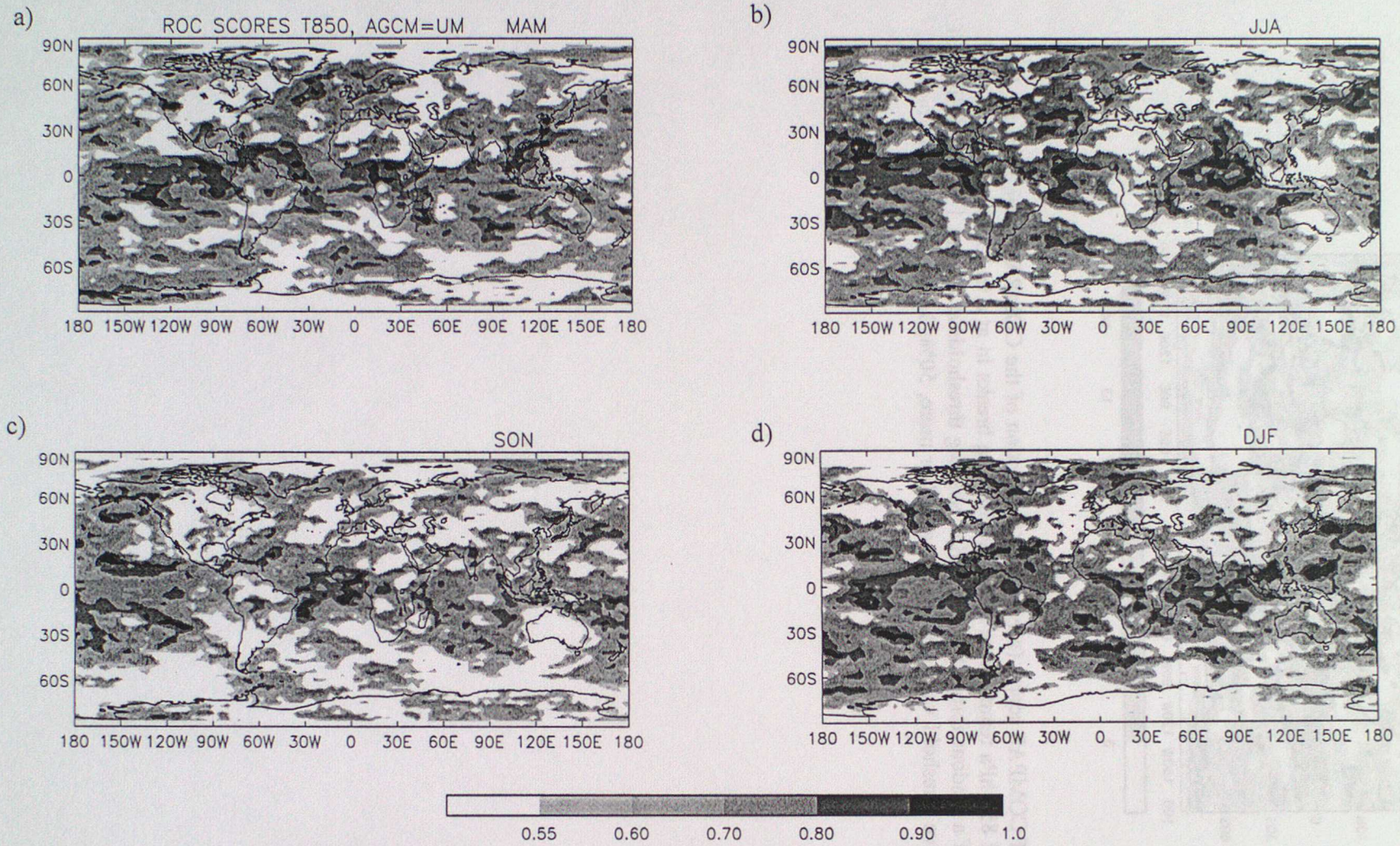


Fig. 5 Geographical distribution of the ROC score (area below Relative Operating Characteristic (ROC) curves obtained at each model grid-point) for UM simulations of the event 850 hPa temperature below (or above) normal.
a) MAM, b) JJA, c) SON and d) DJF.
Shading thresholds are 0.55 and 0.6, then at intervals of 0.1.

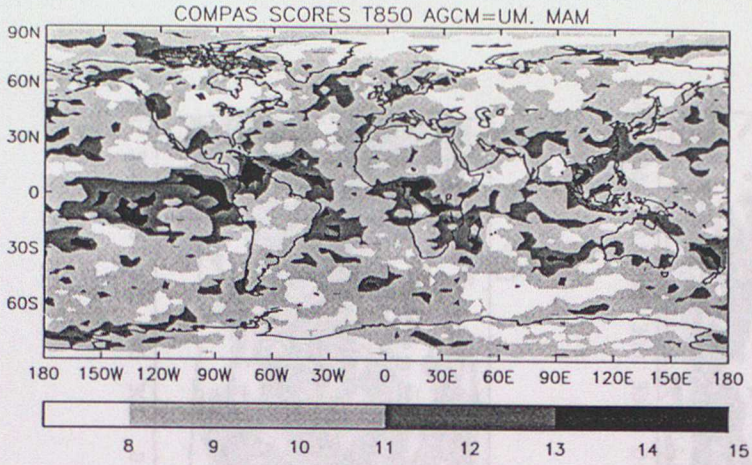


Fig. 6 Spatial plot of COMPAS scores (see text for description of the COMPAS diagnostic) for UM MAM simulations of 850 hPa temperature anomaly. Shading breaks in at a frequency of 8 (the average frequency of a random forecast), with darker shading thresholds at 11 and 13. Probabilities of achieving scores above threshold values by chance are; 8 or more, 50%; 11 or more, 6%; 13 or more, 0.4%.

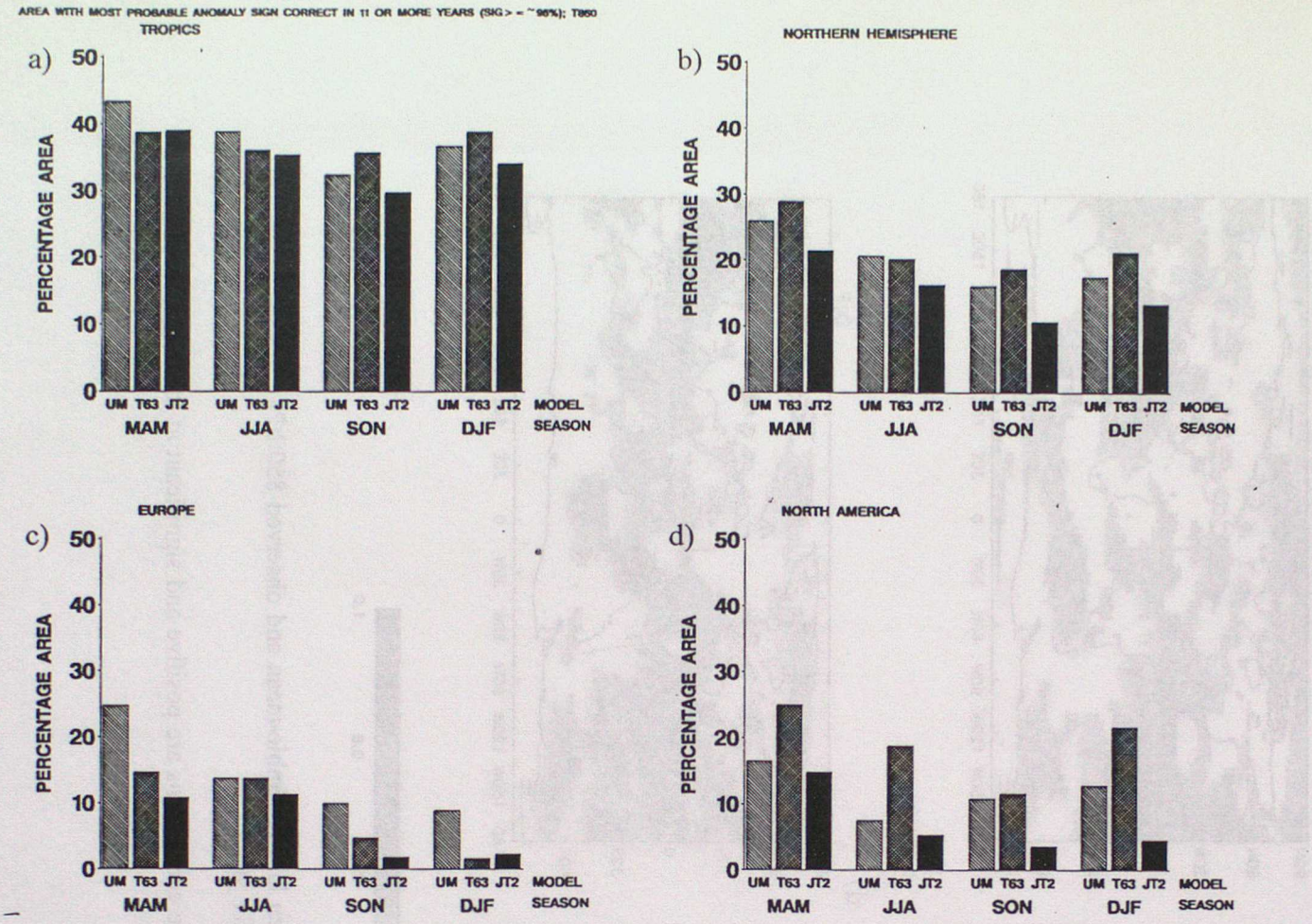


Fig. 7 Percentage area for which the COMPAS score equals or exceeds a value of 11 (indicating significance at 95% level or higher - see text) for UM, T63 and JT2 simulations of 850 hPa temperature. DJF values for T63 and JT2 have been scaled to adjust for calculation over 14 years rather than 15 years (as for the UM). The scaling factor used was the ratio of the UM percentage area obtained using 15 years to that obtained using 14 years.
a) Tropics, b) northern extratropics, c) Europe, d) North America.

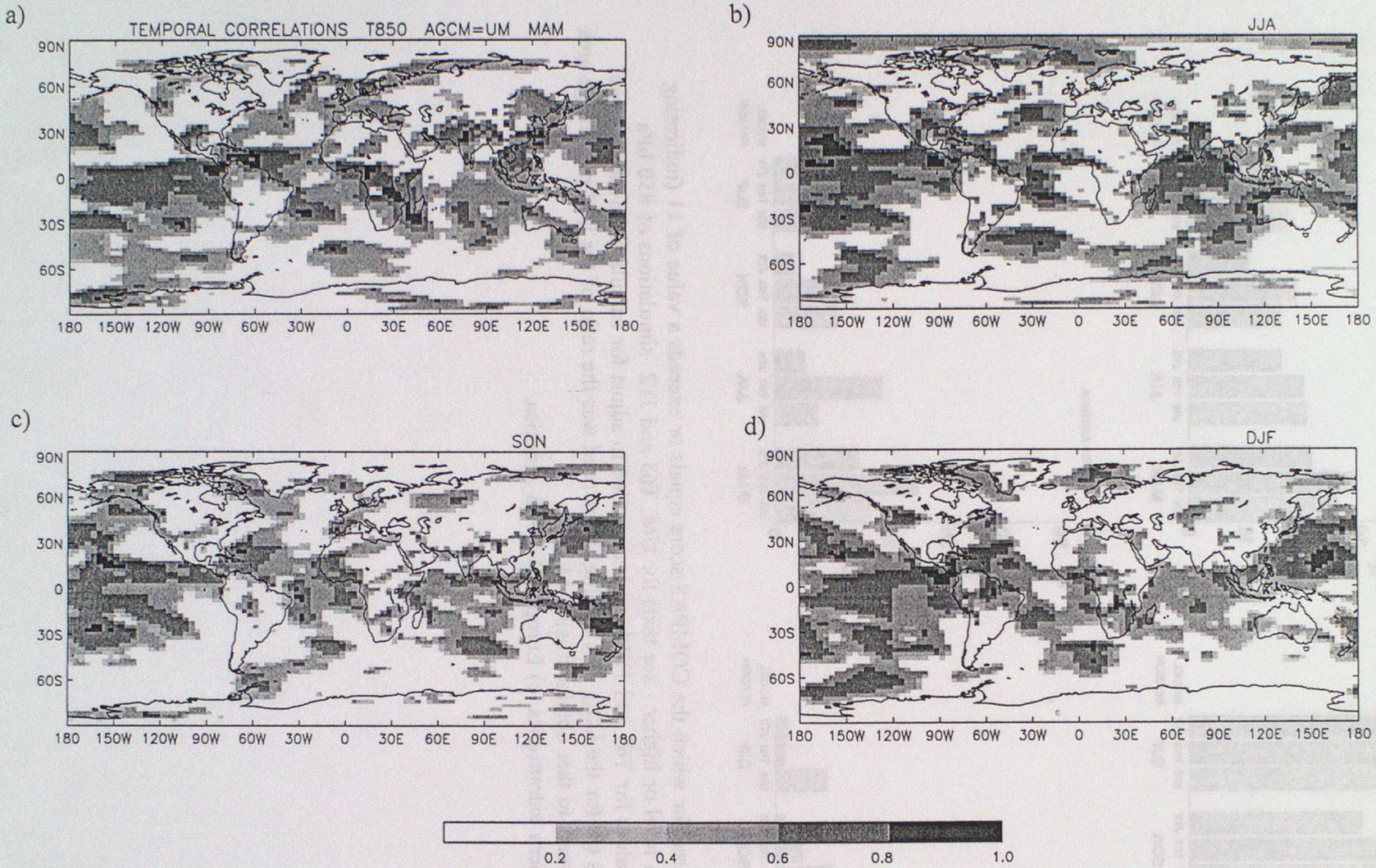


Fig. 8 Spatial map of correlation coefficients between UM ensemble-mean and observed 850 hPa temperature at each grid point over the 15 PROVOST years.
a) MAM, b) JJA, c) SON, d) DJF.
Contour interval is 0.2; values are shown only where correlations are positive and significant at the 90% level or higher.

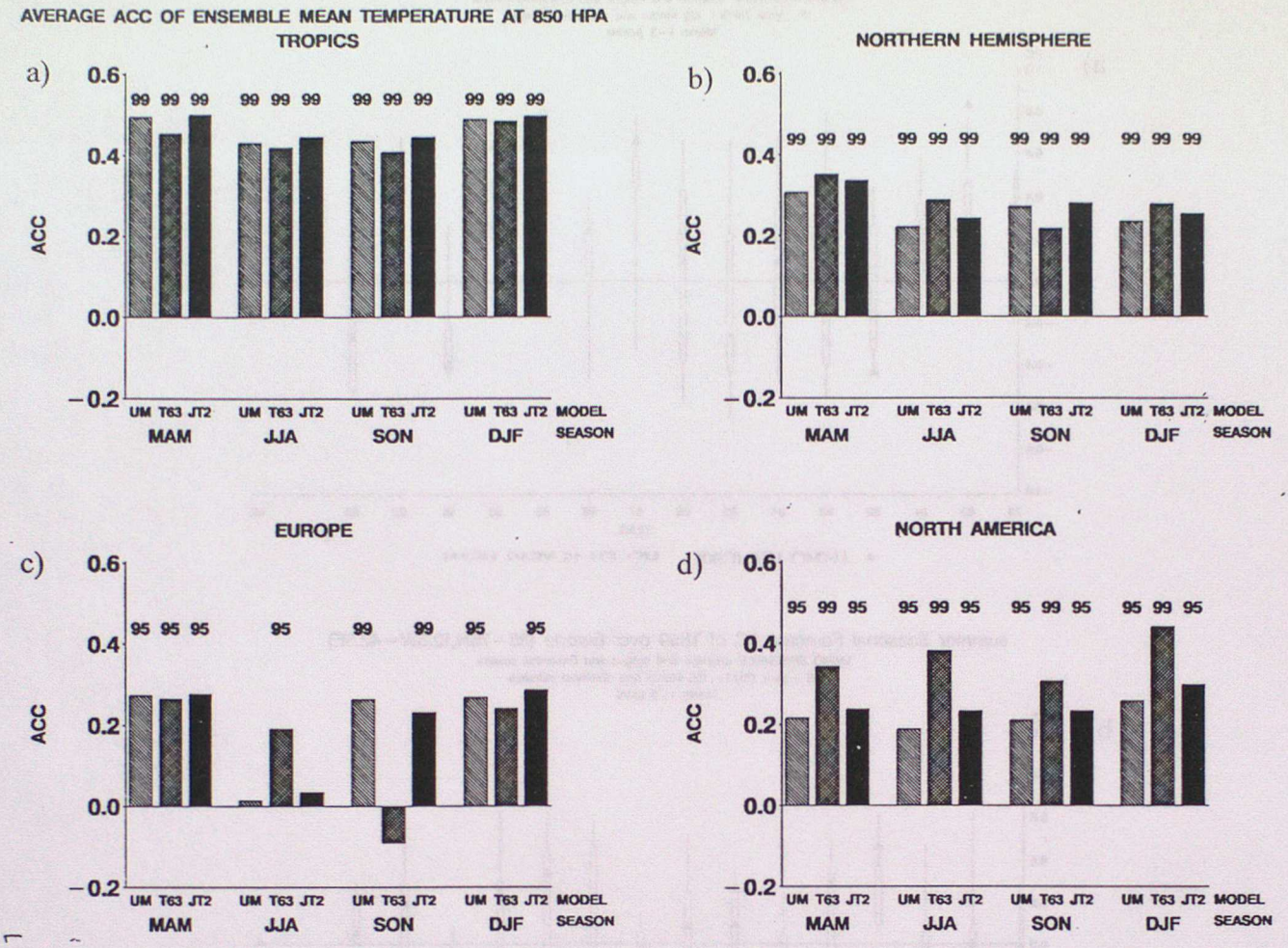
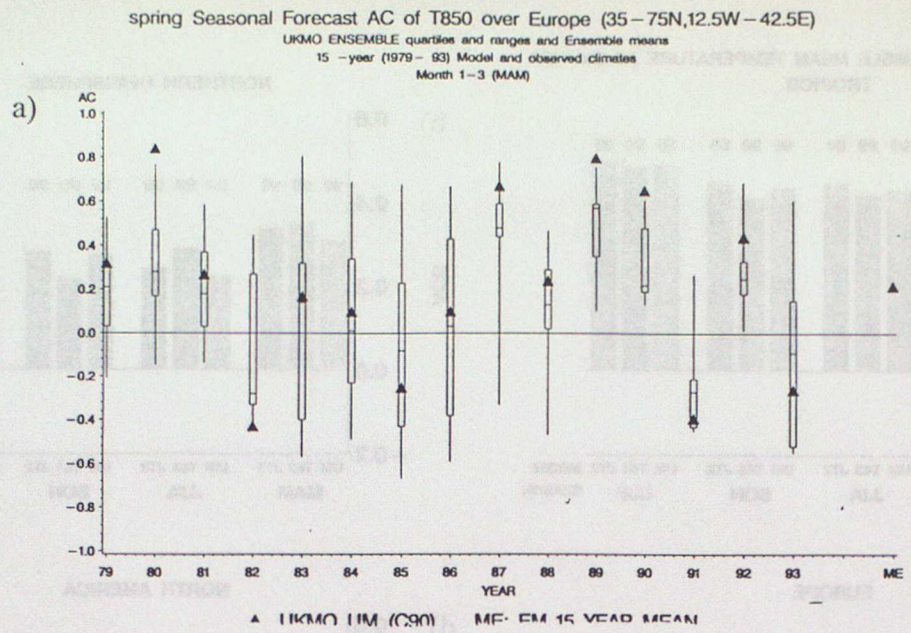


Fig. 9 Average anomaly correlation of ensemble-mean and observed 850 hPa temperature for the UM, T63 and JT2 ensembles. Averages are over the 15 year period (14 years for T63 and JT2 DJF simulations) and are calculated using the Fisher z-transform method. When significance with which the average is different from zero exceeds a threshold of 95% or 99%, the threshold value is plotted above the bars.

a) tropics, b) northern extratropics, c) Europe, d) North America.



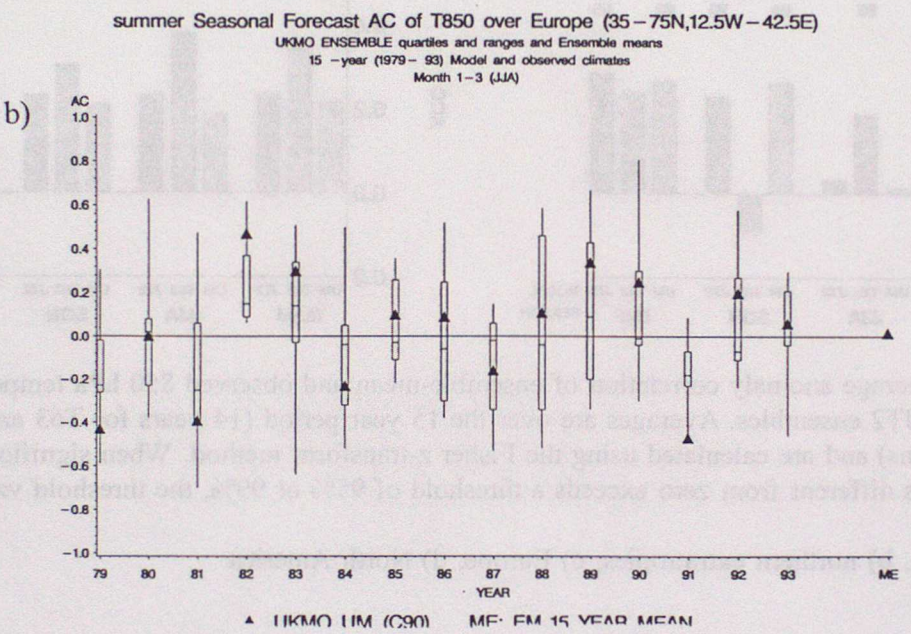
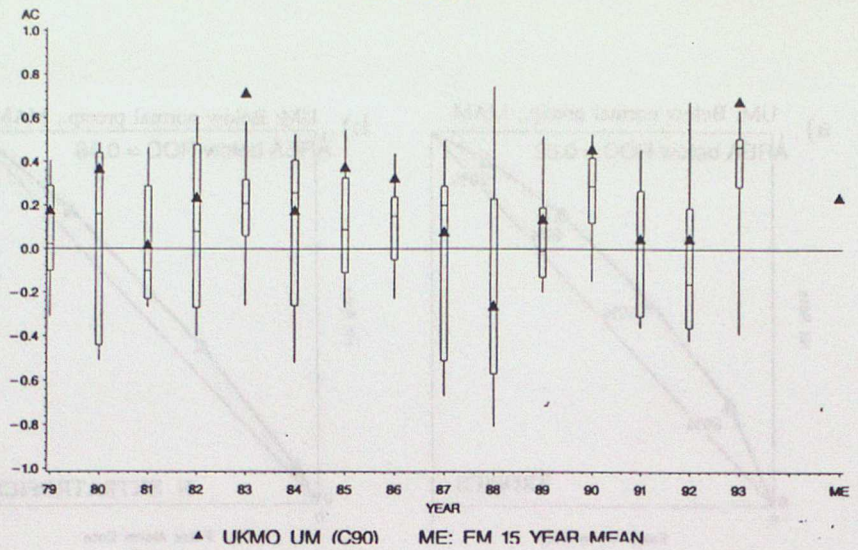


Fig. 10 Annual anomaly correlations coefficients for the 9 UM ensemble members for simulations of 850 hPa temperature over Europe (1979-93). Each stem and whisker plot indicates median, quartile and extreme values. Ensemble mean values are shown as solid triangles.
a) MAM, b) JJA, and (on following page) c) SON, d) DJF.
Note for DJF (d), the year refers to the December of the period, i.e. 79 = DJF79/80.

c)



d)

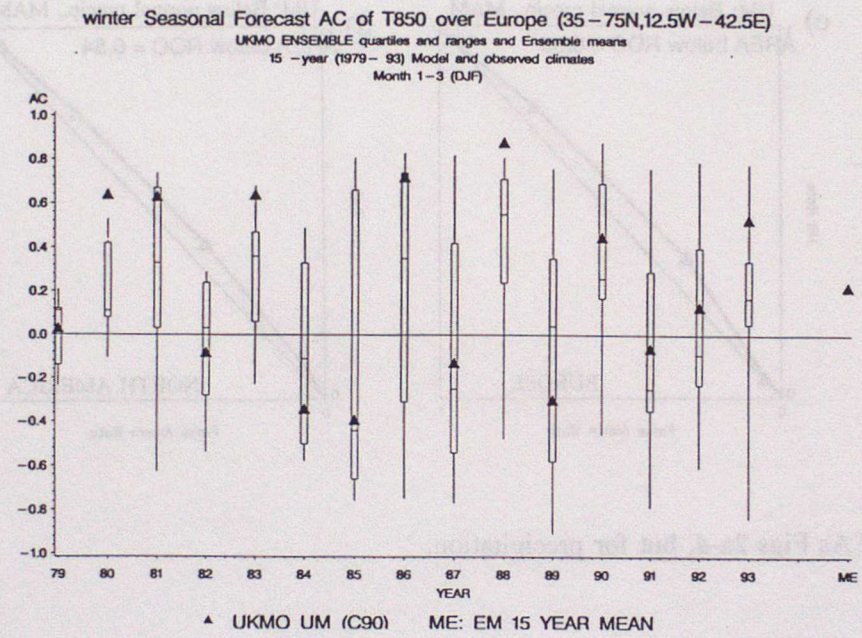
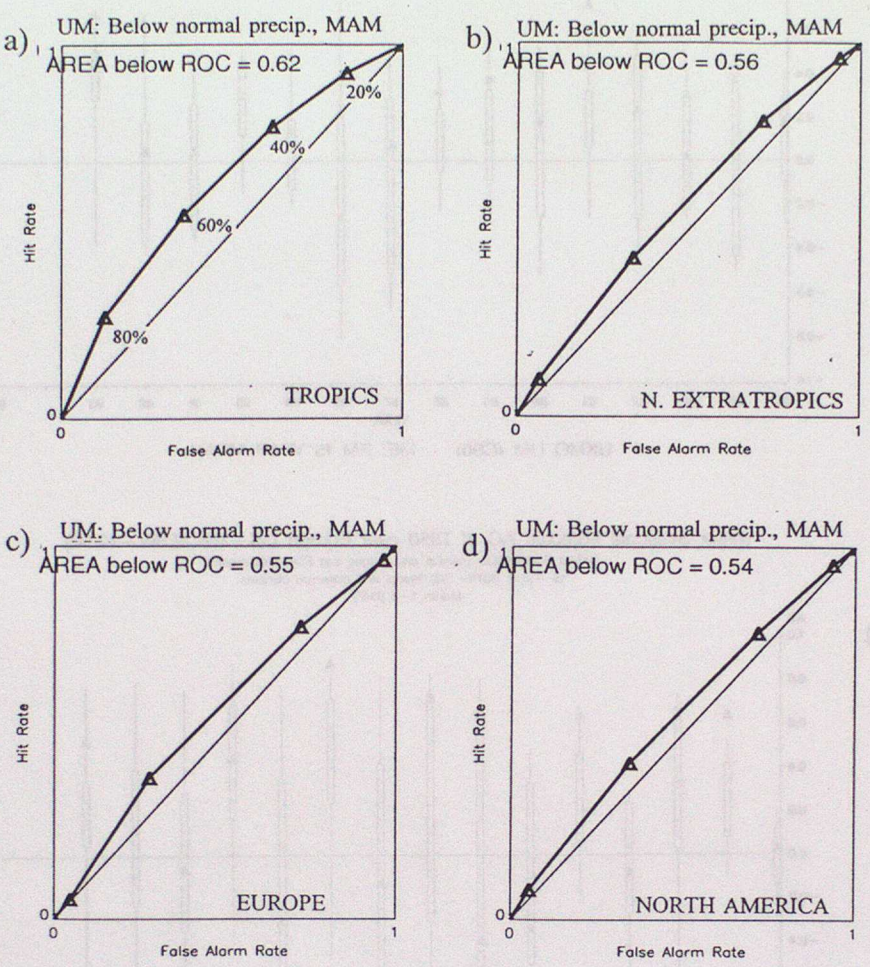
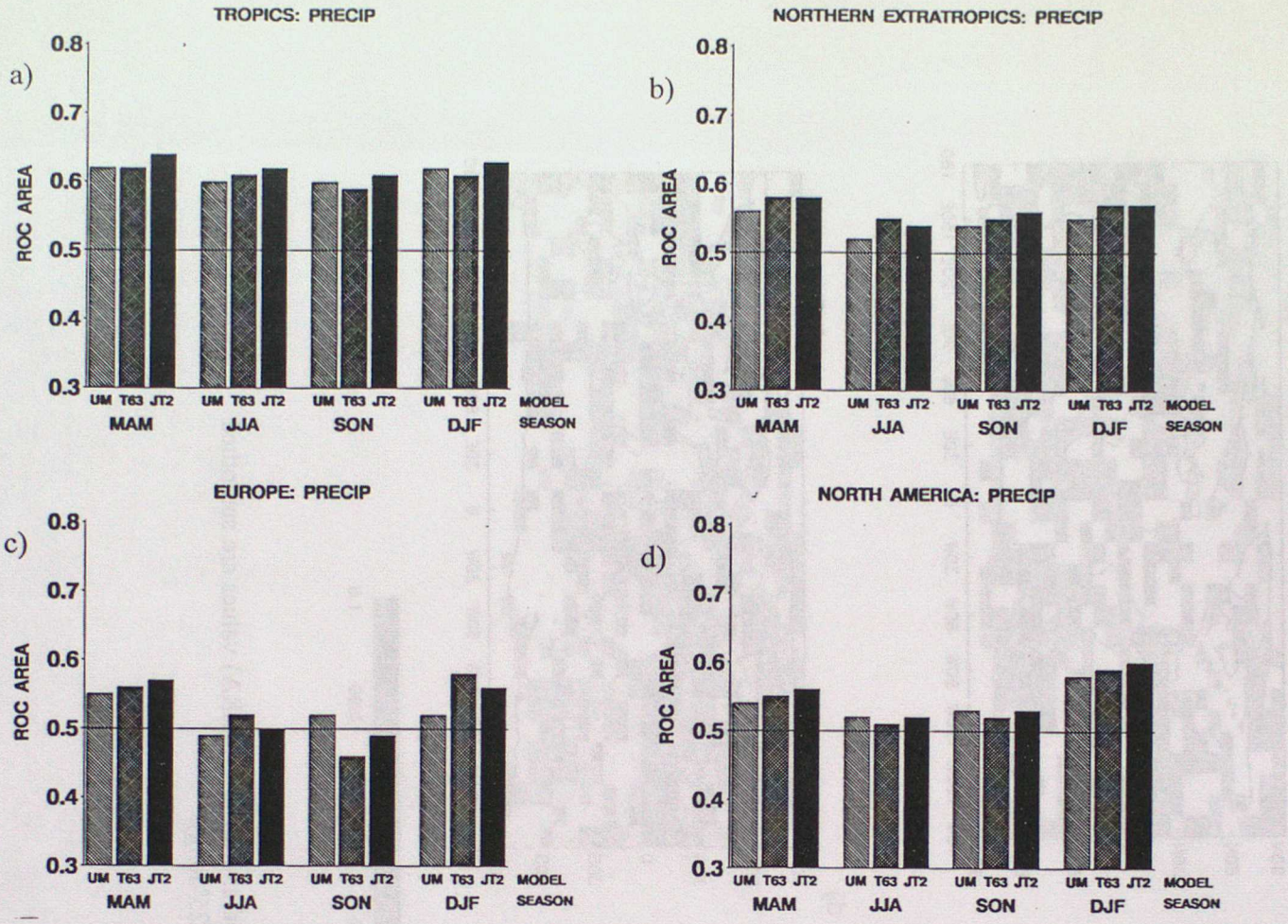


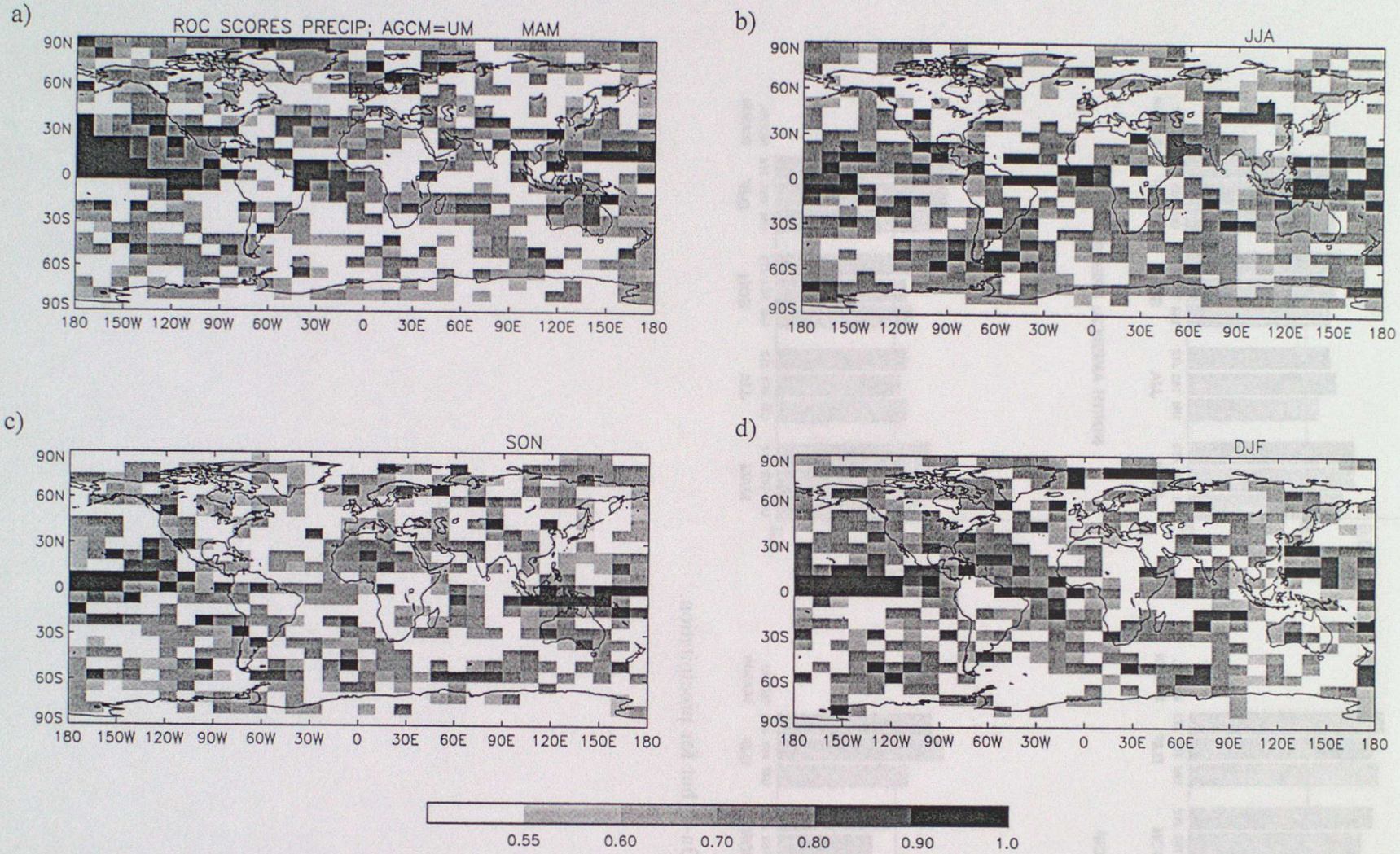
Fig. 10 (cont)



Figs 11a-d As Figs 2a-d, but for precipitation.



Figs 12a-d As Figs 3a-d, but for precipitation.



Figs 13a-d As Figs 5a-d, but for precipitation. Simulated and observed (ERA) values are smoothed over a 3x3 grid-boxes (corresponding to 7.5° lat. by 11.25° long).

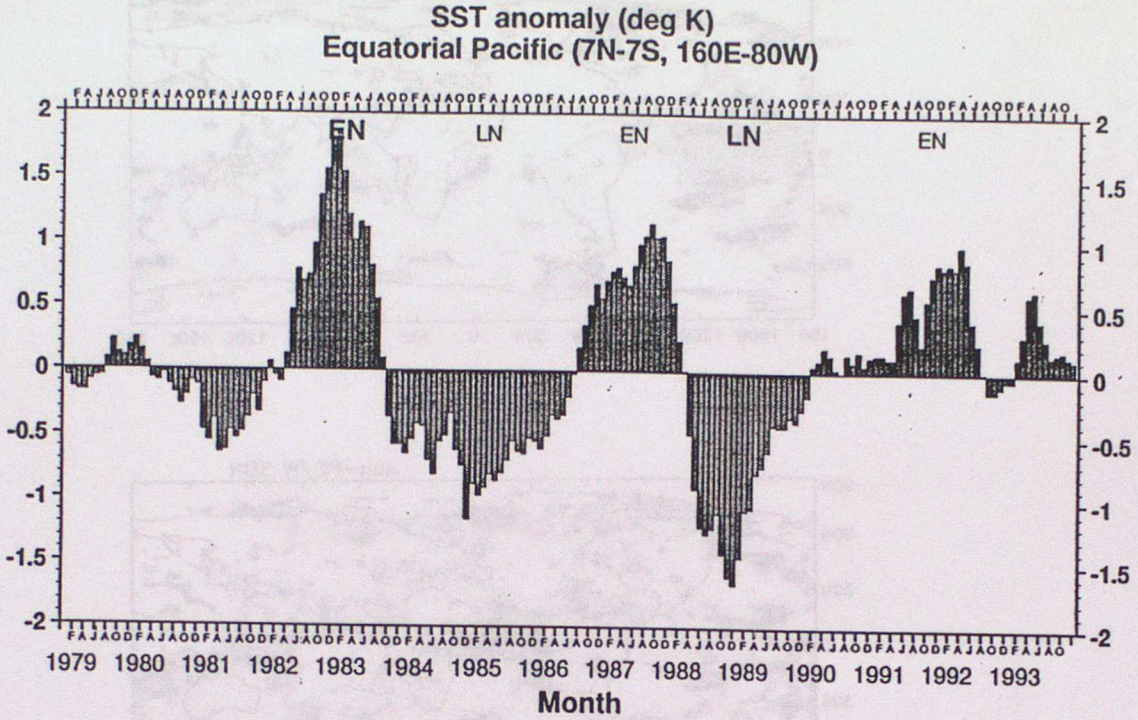


Fig. 14 Monthly mean SST anomalies (deg K) from 1979 to 1993 averaged over the region 7°N-7°S, 160°E-80°W. Strong PC/W events are denoted by large heavy EN (El Niño, PW) and LN (La Niña, PC); moderately strong events are denoted by small EN and LN. (From Branković and Palmer 1999)

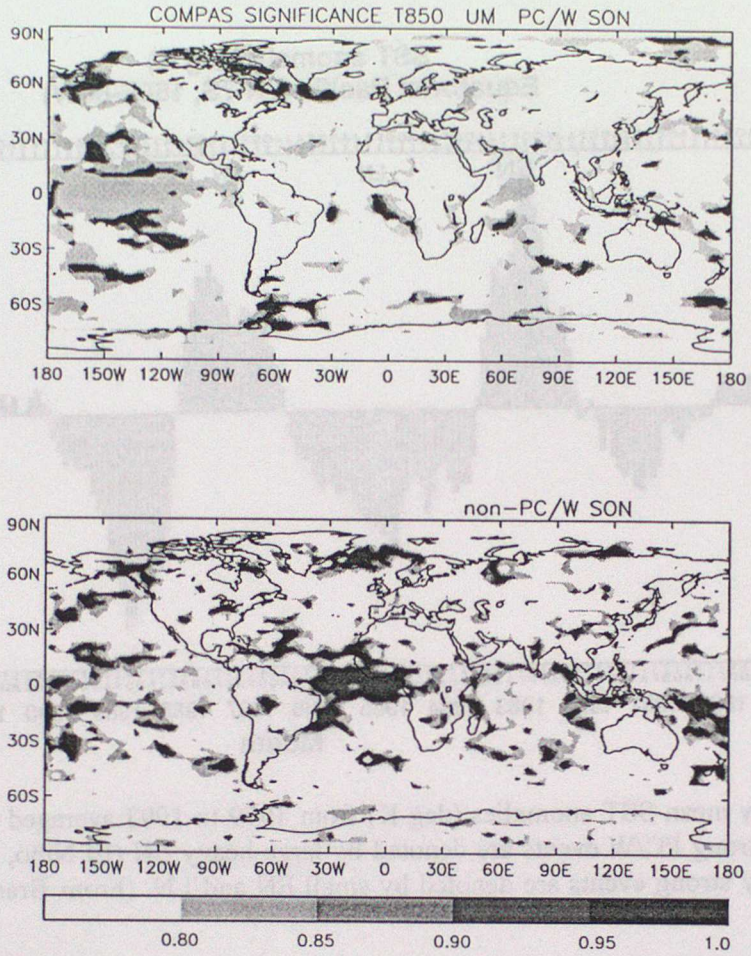


Fig. 15 Spatial plots of COMPAS score significance for UM simulations of 850 hPa temperature anomaly for:
upper panel; the 5 PROVOST SON periods with PC/W events (SON periods selected precede the DJF SST peak, see text for details).
lower panel; the 10 PROVOST SON periods with no PC/W event.
The COMPAS score significance is derived by assigning a significance value, defined as the complement of the probability of achieving an equivalent score with a random forecast. For PC/W years, darkest shading corresponds to regions where the most probable anomaly sign is correct in all 5 years (probability by chance = 0.03, thus significance = 0.97). For non-PC/W years darkest shading corresponds to 8 correct simulations out of 10 years.

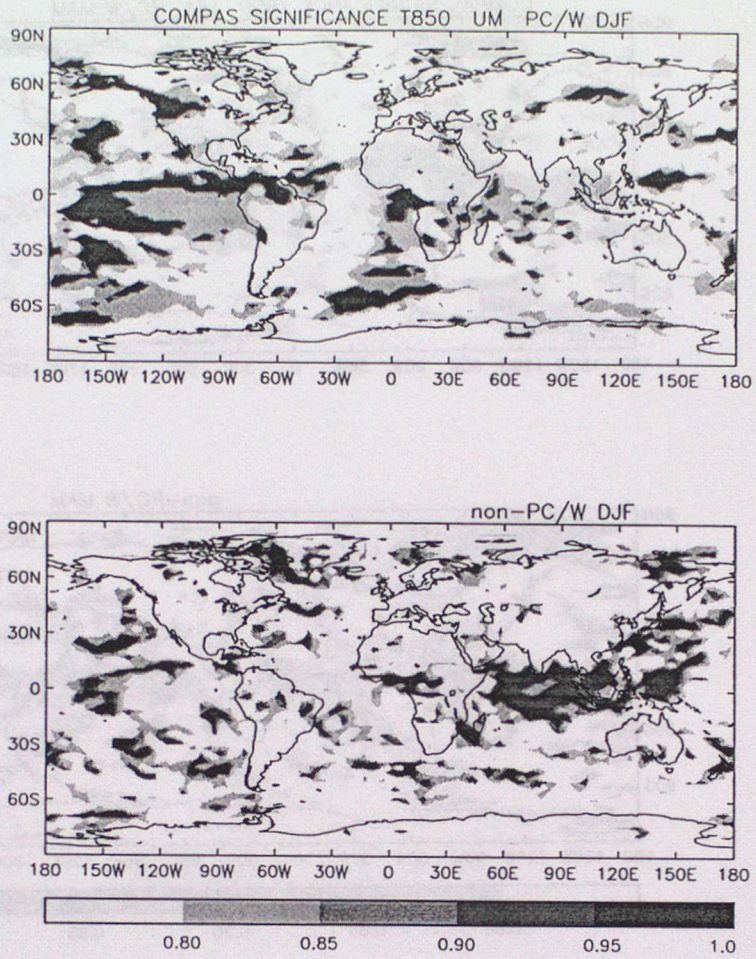


Fig. 16 As Fig 14, but for,
upper panel; the 5 DJF periods with PC/W events
lower panel; the 10 DJF periods with no PC/W event.

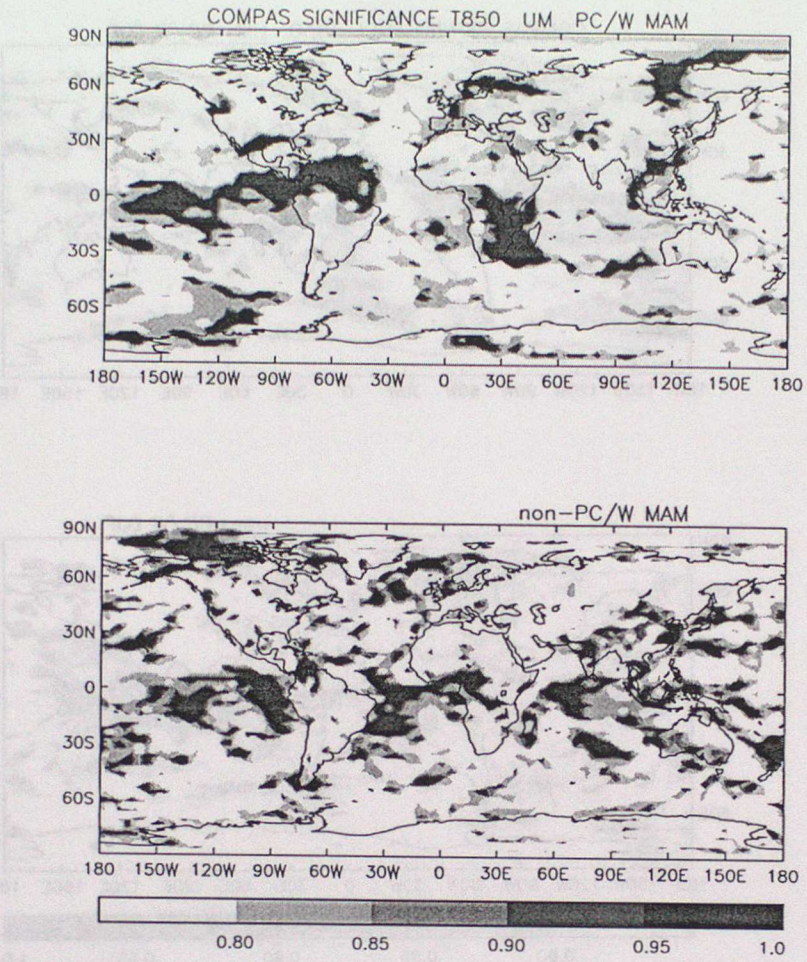


Fig. 17 As Fig. 14, but for,
upper panel: the 5 MAM periods with PC/W events (MAM periods selected follow the DJF SST peak, see text for details).
lower panel; the 10 MAM periods with no PC/W event.

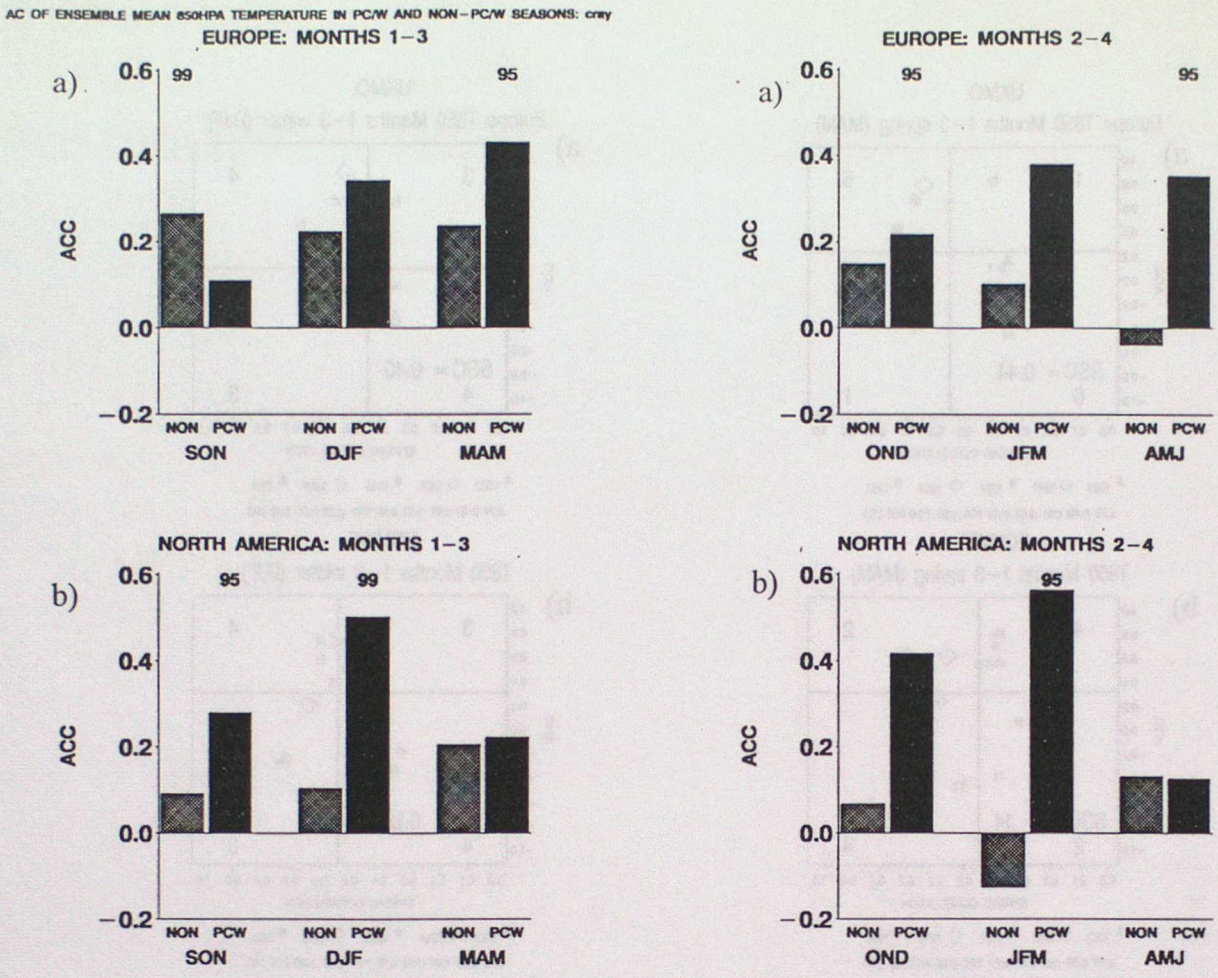


Fig. 18

Fig. 19

Fig. 18 Average anomaly correlation of UM month 1-3 simulations of 850 hPa temperature for PC/W (bars labelled "PCW") and non-PC/W (bars labelled "NON") SON, DJF and MAM periods (see text for definition of PC/W and non-PC/W seasons). When significance (from a t-test) with which the average is different from zero exceeds a threshold of 95% or 99%, the threshold value is plotted above the bars.
a) Northern Hemisphere; b) North America.

Figs 19a&b As Figs 17a&b, but for UM month 2-4 simulations.

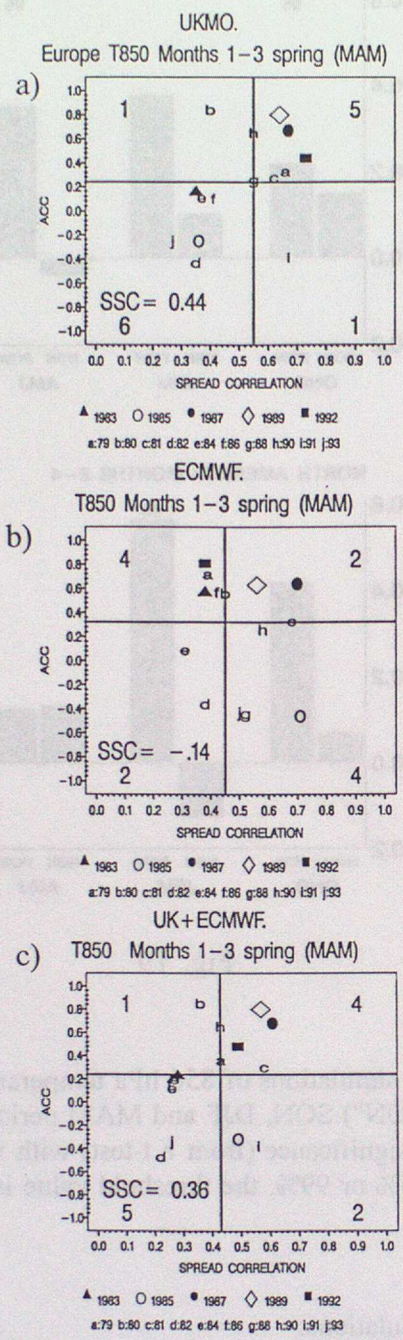


Fig. 20

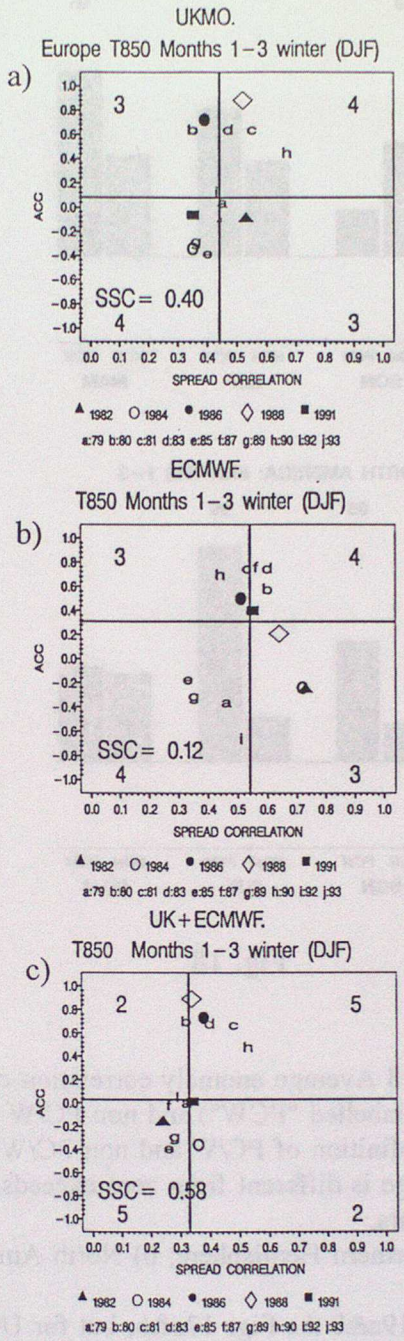


Fig. 21

Fig. 20 Scatterplots of ensemble-mean skill (defined using anomaly correlation) and ensemble spread (defined as the average, Fisher z-transformed, anomaly correlation of the ensemble members with the ensemble mean) for month 1-3 MAM simulations of 850 hPa temperature over Europe,
a) UM, b) T63, c) JT2.
The total entries in the diagonal quadrants (constructed using the ensemble median values of AC skill and spread) are given. Years corresponding with PC/W events are indicated with symbols, and the identity of other individual years is denoted with letters. The linear skill/spread correlation (SSC) is also provided.

Figs 21a-c As Figs 20a-c, but for DJF.

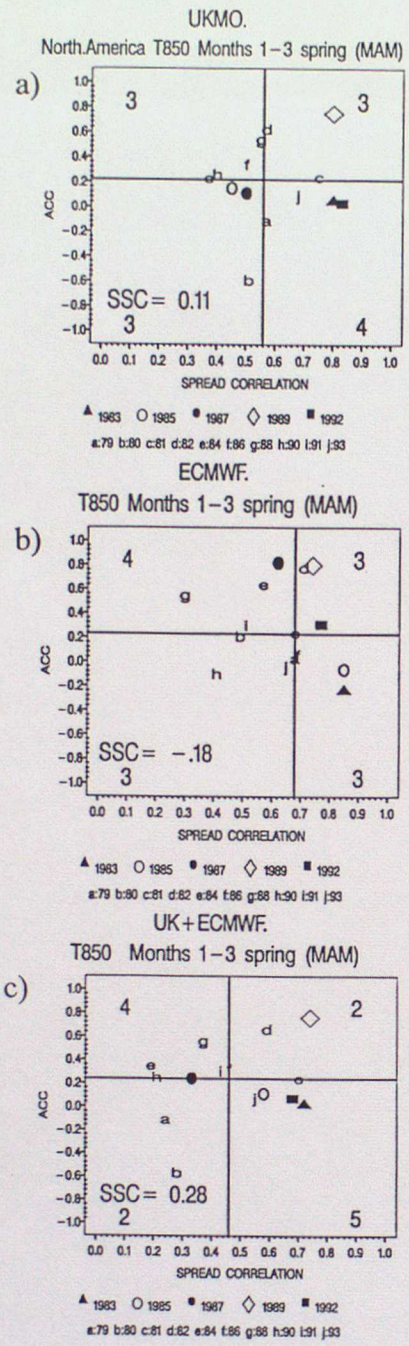


Fig. 22

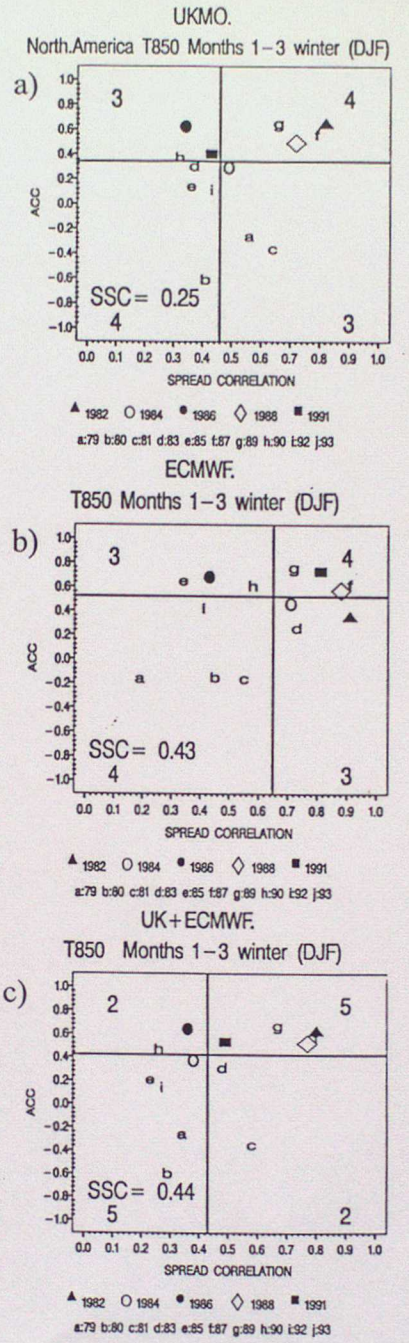


Fig. 23

Figs 22a-c As Figs 20a-c, but for MAM over North America.

Figs 23a-c As Figs 20a-c, but for DJF over North America.

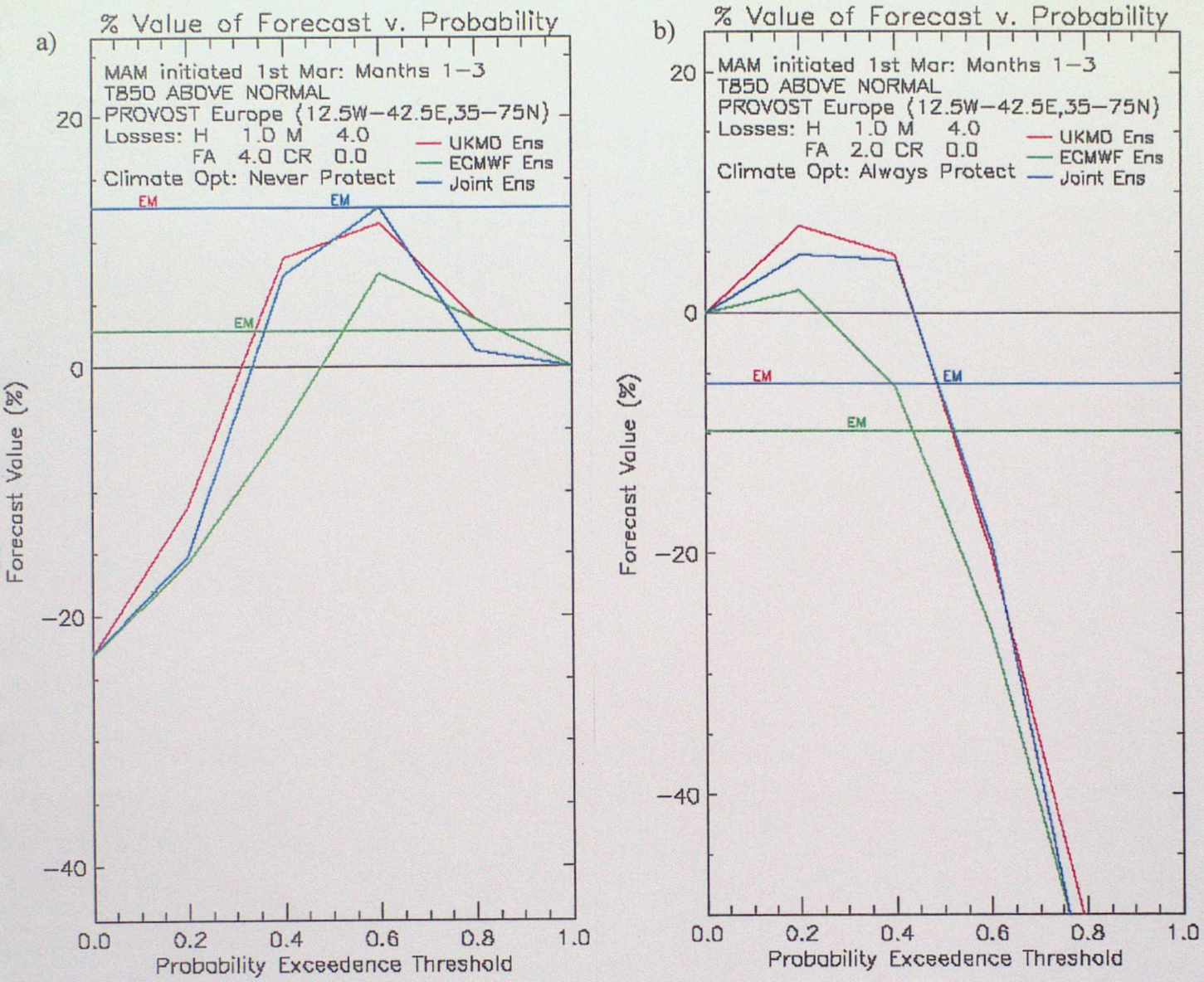


Fig. 24 Potential value (V) of month 1-3 MAM simulations of 850 hPa temperature above normal over Europe with user losses of,

a) $L_h = 1$, $L_m = 4$, $L_f = 4$

b) $L_h = 1$, $L_m = 4$, $L_f = 2$.

Plotted curves show the potential value of probabilistic predictions at thresholds of 0%, 20%, 40% ... 100%, while horizontal lines labelled EM show corresponding potential value of deterministic predictions based on the ensemble mean, for three ensembles UM (red), T63 (green) and JT2 (blue).