Analysing and combining atmospheric general circulation model simulations forced by prescribed SST: tropical response

Vincent Moron ([1]), **Antonio Navarra** ([2]), **M. Neil Ward** ([3]), **Chris K. Folland** ([4]), **Petra Friederichs** ([5]), **Karine Maynard** ([6]) **and Jan Polcher** ([6])

([1]) *UFR des Sciences Géographiques et de l'Aménagement, Université de Provence and UMR CEREGE, CNRS, Aix en Provence, France*

([2]) *Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy*

([3]) *CIMMS, University of Oklahoma, Norman, U.S.A.*

([4]) *Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, England*

([5]) *Meteorologisches Institut des Universitaet Bonn, Germany*

([6]) *LMD-CNRS, Université Pierre et Marie Curie, Paris, France*

Abstract

The ECHAM 3.2 (T21), ECHAM 4 (T30) and LMD (version 6, grid-point resolution with 96 longitudes × 72 latitudes) atmospheric general circulation models were integrated through the period 1961 to 1993 forced with the same observed Sea Surface Temperatures (SSTs) as compiled at the Hadley Centre. Three runs were made for each model starting from different initial conditions. The large-scale tropical inter-annual variability is analysed to give a picture of the skill of each model and of some sort of combination of the three models. To analyse the similarity of model response averaged over the same key regions, several widely-used indices are calculated: Southern Oscillation Index (SOI), large-scale wind shear indices of the boreal summer monsoon in Asia and West Africa and rainfall indices for NE Brazil, Sahel and India. Even for the indices where internal noise is large, some years are consistent amongst all the runs, suggesting inter-annual variability of the strength of SST forcing. Averaging the ensemble mean of the three models (the super-ensemble mean) yields improved skill. When each run is weighted according to its skill, taking three runs from different models instead of three runs of the same model improves the mean skill. There is also some indication that one run of a given model could be better than another, suggesting that persistent anomalies could change its sensitivity to SST. The index approach lacks flexibility to assess whether a model's response to SST has been geographically displaced. We focus on the first mode in the global tropics, found through singular value decomposition analysis, which is clearly related to El Niño/Southern Oscillation (ENSO) in all seasons. The Observed-Model and Model-Model analyses lead to almost the same patterns, suggesting that the dominant pattern of model response is also the most skilful mode. Seasonal modulation of both skill and spatial patterns (both model and observed) clearly exists with highest skill (between tropical Pacific SST and tropical rainfall) and reproducibility amongst the runs in December-February, and least skill/reproducibility in March-May and June-August. The differences between each model suggest that a simple linear regression combination of each GCM's prediction indices will be improved upon by combination methods that take account of the errors in the spatial teleconnection structures generated by the GCM.

Key words *atmospheric general circulation model – inter-comparison – tropical circulation – seasonal rainfall – tropical Pacific SST*

Mailing address: Dr. Antonio Navarra, Istituto Nazionale di Geofisica e Vulcanologia, Viale Carlo Berti Pichat 8, 40127 Bologna, Italy; e-mail: navarra@ingv.it

1. Introduction

The first multi-year integration of an Atmospheric General Circulation Model (AGCM) forced with the observed time-varying Sea Surface Temperature (SST) was performed by Lau and his colleagues at GFDL (Lau, 1985; Kang

and Lau, 1986). These studies showed that the impact of SST variations is generally much stronger on tropical circulation than on flow patterns in middle latitudes (see Moron *et al.*, 2001b), and the source of much of the SST forcing was associated with El Niño/Southern Oscillation (ENSO). Such decadal to multi-decadal runs have now been performed by many modelling groups, substantially as part of the AMIP experiment (*i.e.* Latif *et al.*, 1990; Lau and Nath, 1994; Graham *et al.*, 1994; Harzallah and Sadourny, 1995; Kumar and Hoerling, 1995; Hense and Roemer, 1995; Potts *et al.*, 1996; Davies *et al.*, 1997; Renshaw *et al.*, 1998; Rowell, 1998; Moron *et al.*, 1998a – hereafter MNWR98; Livezey and Smith, 1999; Smith and Livezey, 1999). These experiments have laid the basis for experimental seasonal forecasting rooted in the response of the atmosphere at seasonal time scales to prevailing SST patterns. An important conclusion of the modelling studies is the necessity to consider multiple GCM runs for accurately identifying the SST-forced component of atmospheric variation in the model, because the land-atmosphere system in the GCM generates different seasonal variability itself inside each run, in addition to that part of the variance attributable to the SST response (*i.e.* Palmer, 1986; Barnett, 1995; Stern and Miyakoda, 1995; Bengtsson *et al.*, 1996; Zwiers, 1996; Barnett *et al.*, 1997; MNWR98; Rowell, 1998; Zwiers and Kharin, 1998).

Even when large ensembles are made, the fact remains that different GCMs respond to the same SST forcing in different ways. An interesting dilemma exists as to whether to make a seasonal forecast by using a very large ensemble from the GCM that has been shown to have most skill, or whether to make a set of ensembles from a range of skilful models, and combine the forecasts in some optimal way. In reality, the seasonal forecast community is faced with the latter situation, which will likely persist for some time given the advantages to model development of having a number of competing GCMs around the world. Krishnamurti *et al.* (2000) demonstrated clearly that an optimal weighted combination of eight different AGCM experiments forced by the same prescribed SST forcing from January 1979 to December 1988 (with only one run for each AGCM) outperforms the skill of any individual AGCM, including the best ones (Krishnamurti *et al.*, 2000). But, this comparison is a little biased since the weighted optimal solution of eight members is compared with the output of one individual run. They demonstrated also that the optimal linear combination of eight AGCMs is better than the solution where an equal weight is assigned to each model, because less weight is assigned to the poorer AGCM (Krishnamurti *et al.*, 2000). One question to be addressed in this paper is how best to combine the different output from different AGCMs, considering the estimated skill level associated with each AGCM, and with samples of the same size (*i.e.* three runs from three different AGCM *versus* three runs of the same AGCM). A range of widely-used indices like SOI and regional rainfall indices calculated for three different GCMs are studied. The three GCMs are the ECHAM3 (T21 resolution), ECHAM4 (T30 resolution) and LMD 6 (96 longitudes × 72 latitudes). Each model was integrated three times starting from different initial atmospheric conditions forced with the time-varying SST through the period 1961-1993.

Furthermore, while it is clear that different GCMs have different spatial responses to SST patterns (Sperber and Palmer, 1996; Liang *et al.*, 1997; Zwiers and Kharin, 1998), little work has actually been done to diagnose those differences. For example, one model may simulate inter-annual variability almost perfectly in phase with the Indian monsoon variability, but the model misplaces that variability by 500 km, leading to an apparent low skill-level when a geographically co-located model and observed Indian rainfall index are compared. The potential to apply coupled pattern techniques to correct such errors has been demonstrated for a set of runs with one model (Feddersen *et al.*, 1999; Moron *et al.*, 2001a) and such an approach can be expected to lead to further advances in optimal combination of output from different GCMs. We do not address that point in this paper, but make the preparatory diagnostic studies of the response of three GCMs to observed SST. One reason for a geographically misplaced AGCM teleconnection response to SST is expected to

be systematic errors in the background annual cycle. For example, if diabatic heating anomalies in the Indian monsoon region are an important forcing factor to pass on the ENSO influence through the Indian monsoon and into subtropical North Africa, then a model Indian subcontinent with much reduced climatological rainfall is unlikely to generate sufficiently large rainfall (diabatic heating) anomalies to set up the remote teleconnection from India to Africa in the atmosphere. With this in mind, the background climatologies of precipitation of each model are presented in this paper in some detail, to indicate the extent to which background vertical motion and diabatic forcings are correctly represented in each model, and to form a backdrop for interpretation of each model's teleconnection structure.

Details of the model integrations, observed data sources and methodology are given in Section 2. Section 3 considers the annual cycle of some key-fields in the Tropics for each model. The model skill for some basic indices of tropical circulation (SOI, monsoon index, regional rainfall indices) is evaluated in Section 4. The leading Observed-Model (*i.e.* which maximises the covariance between the observation and the runs) and the Model-Model (*i.e.* which maximises the covariance amongst the runs) seasonal modes of the global tropical rainfall in each season are presented in Section 5. A conclusion closes the paper (Section 6).

2. Experimental design and statistical methods

2.1. *The AGCM integrations*

All the integrations reported here have been forced by the UK Met Office's Global Sea Ice and Sea Surface Temperature data set (GISST 1.0 until 1990, GISST 2.1 after 1990 and GISST 2.2 after 1996 – refer to Parker *et al.*, 1994 and Rayner *et al.*, 1996). All other external atmospheric forcings (solar radiation input, CO_2 concentration, ecc.) are kept at standard values.

The AGCMs used for the integrations are the ECHAM 3.2 horizontal spectral resolution = T21 and 19 vertical levels), ECHAM 4 (horizontal spectral resolution = T30 and 19 vertical levels) and LMD 6 (horizontal grid-point resolution = 96 longitudes × 72 latitudes and 19 vertical levels) models – hereafter respectively EC3, EC4 and LMD. The EC3 (resp. 4) is the third (resp. fourth) generation of a climate model originating from the ECMWF spectral medium range weather forecast model. The development was undertaken at the Max Planck Institute für Meteorologie and Deutsches Klimarechenzentrum (DKRZ) in Hamburg. Details of the numerical formulation of the model and its physical parametrizations can be found in Roeckner *et al.* (1996), where comparisons are made between different versions and horizontal resolutions, and with observations. Three runs of EC3 and three runs of EC4 were made over the period 1961-1993 all with different initial atmospheric conditions. Some results of the EC4 runs have been presented elsewhere (MNWR98; Navarra *et al.*, 1999; Miyakoda *et al.*, 1999; Feddersen *et al.*, 1999; Moron *et al.*, 2001a). LMD is derived from the LMD standard GCM (Sadourny and Laval 1984; Harzallah and Sadourny, 1995; Li, 1999).

2.2. *Tropical indices*

Seasonal to inter-annual climate variability is often summarised using regional or global-scale indices. Comparing a selection of these amongst the models and observations will serve to give a first indication of the performance of the 3 models. In addition, the indices provide a good starting point to address the question of how the skill of a range of models can be combined to generate a product of maximum skill. The Southern Oscillation Index (SOI) was computed in the model and observations (provided by T. Basnett from the UKMO) according to the following procedure: monthly grid-box SLP data are averaged for a «Darwin regional index» (130°-140°E; 10°-20°S) and a «Tahiti regional index» (205°-215°E; 10°-20°S). These regional indices are then standardised to zero mean and unit variance. Note that each model run is independently standardised. Then, the SOI is defined as the difference between the «Tahiti regional index» and the «Darwin regional index».

Concerning the monsoon circulation, an important index has been defined by Webster and Yang (1992) to measure the overall intensity of the SE Asian boreal summer monsoon, the index of the difference between mean zonal components of the wind at 850 (U850 hereafter) and 200 (U200 hereafter) hPa averaged over the box 5°-20°N; 40°-110°E. A similar index was defined for West-African monsoon in Fontaine and Janicot (1992) and Fontaine *et al.* (1995). We compute here such an index as the difference between the normalised U850 speed averaged over 5°-15°N; 15°W-15°E and the same area at U200 hPa. This index is measured on June-August period according to the definition of Webster and Yang (1992) for SE Asia (SEA hereafter) and June-September for West Africa (WAF hereafter).

Sperber and Palmer (1996) and Gadgil and Sajani (1998) compared AMIP simulations of several regional rainfall indices, used in long-range forecasting. MNWR98 noticed that EC4 results are close to the mean AMIP scores for India (JJA; 12°-27°N; 70°-82°E), Sahel (JAS; 12°-18°N; 15°W-35°E) and Brazilian Nordeste (MAM; 35°-45°W, 12°-5°S). Continental grid-box rainfall are individually standardised. Normalised data are then averaged and the resultant quantity is then re-standardised. Each run is processed independently. Observed data are extracted from the Climatic Research Unit data set (Hulme, 1991) and processed in the same way as the runs.

2.3. *Statistical methods*

We use Singular Value Decomposition Analysis (SVDA) as defined in Bretherton *et al.* (1992), and as applied to analysis of AGCM integrations Ward and Navarra (1997) and MNWR98. SVDA extracts the linear combination maximising the covariance between two fields. We use it to extract:

i) The first mode maximising the covariance between the runs and the observations (*i.e.* the most skilful mode), called OM (Observation-Model). This is achieved by computing the covariance matrix between the runs (stacked beneath each others) and the observed matrix (repeatedly copied the number of available runs). For example, with three runs denoted X, Y and Z, and one set of observations denoted O (with time in rows and grid-points in columns), the left-hand (L) side and right-hand (R) side matrices (Bretherton *et al.*, 1992) are constructed as

$$L = \begin{bmatrix} O \\ O \\ O \end{bmatrix} \qquad R = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

The singular decomposition of the covariance matrix ($C = L' \times R$) yields two sets of vectors (which are the coupled modes); one set containing the observed modes (called OM_o) and one set containing the corresponding model modes (called OM_m);

ii) The first common mode amongst the runs, which is the pattern maximising reproductibility, called MM (Model-Model). This is achieved by constructing two matrices containing all possible combinations of two runs (six combinations for three runs and 72 for nine runs). For example, with three runs denoted X, Y and Z (with time in rows and grid-points in columns), the left-hand (L) and right-hand (R) side matrices are constructed

$$L = \begin{bmatrix} X \\ Y \\ Z \\ X \\ Y \\ Z \end{bmatrix} \qquad R = \begin{bmatrix} Y \\ Z \\ X \\ Z \\ X \\ Y \end{bmatrix}$$

The covariance matrix ($C = L' \times R$) is then symmetric and yields two sets of vectors which are identical (except sometimes for sign). These vectors are the sets of modes (called MM_m) ranked according to the cross-covariance amongst their associated time series, *i.e.* ranked according to a measure of the mode's reproducibility, indicating that the mode in the AGCM is at least partly being forced by the SST.

We first perform separate analyses on each model to compare the different model responses to SST. In addition, all runs from the three models are pooled together to form a super-ensemble, and the analyses are repeated to yield the

first mode that is common amongst all models, a mode that in some senses is independent of the different parametrizations, resolution and numerical calculus used by each modelling group. An extended SVDA is also performed to relate each model's base- pattern to the same observed field. For that, the LMD and EC4 ensemble anomaly fields are interpolated onto the T21 grid and an SVDA is then computed between observations and the three models pooled together. We obtain then one pattern for observations and one pattern for each model. Results show that this solution is almost equal to OM analysis performed independently on each model (not shown).

The SVDA is cross-validated following MNWR98. Each year is iteratively excluded from the SVDA and the anomalies for the excluded year (that are computed with the mean and standard deviations of the remaining period) are then projected onto the patterns computed from the remaining period. Cross-validated squared covariances are computed from the cross-validated time series for each coupled mode (Bretherton *et al.*, 1992) and these are then ranked in descending order. Some covariances computed for the lower order coupled modes may even be negative. In fact, the drop between significant covariance and near-zero (or even negative) values occurs quite fast, usually between the third and fifth values. The variance explained by the coupled mode in each of the two variables (here, observed and model) is estimated by the mean of the squared heterogeneous correlations between cross-validated time series of the leading SVDA mode of one field and the other field (Lau and Nath, 1994). All figures and values displayed here (figs. 6a-d to 10a-d) are based on ensemble mean model values; for example, heterogeneous correlations with OM_m are obtained from the correlations between the mean of the three (or nine) cross-validated OM_m time coefficients and the observed anomaly fields. The explained variance is then the weighted mean of these squared heterogeneous correlations.

3. The mean annual cycle of rainfall

The precipitation climatology in each model is compared with the precipitation climatology of Xie and Arkin (1996), which is based on blended estimates from satellite and gauge data. The model circulation fields at 850 and 200 hPa have also been compared with ECMWF reanalyses (1980-1988) and these results are mentioned in assisting interpretation of the rainfall climatologies.

For calculation of the monthly spatial correlation between observed and simulated rainfall (fig. 1), the observed grid was linearly interpolated onto the corresponding model grid. EC4 is the best model in the reproduction of the monthly patterns (fig. 1). Even the lowest scores of the three models (LMD) are fairly good relative to the larger set of AGCMs studied by Srinivasan *et al.* (1996). The performance of the models is poorest in boreal summer and best in boreal winter. This quantity masks some regional biases (figs. 2a-d and 3a-d) which are now summarised.

In December-February (fig. 2a-d), all models successfully reproduce highest rainfall rates between the Central Indian Ocean and South Pacific Convergence Zone (hereafter SPCZ). There is a general tendency to produce too much precipitation over the zone as a whole, and there are stronger 850 hPa easterlies than observed in the Central Pacific (not shown). This circulation feature is particularly the case for the LMD model, though the actual enhancement of precipitation occurs in very localised regions. EC4 remarkably captures the pattern and the rates of rainfall over Indo-Pacific longitudes, but fails in the reproduction of the Atlantic ITCZ which vanishs over the equatorial Atlantic (fig. 2c). EC4 generates also too much rainfall in the Indian Ocean and across Southern Africa. In EC3, the Atlantic ITCZ is present, albeit too weak, and the Indian Ocean is better represented (fig. 2c), yet the positive bias over Southern Africa remains. LMD is the best in the Atlantic ITCZ's reproduction, though it is too far south (fig. 2d). A strong ITCZ is generated in the LMD model in the Northern Indian Ocean, whereas in observations, convection and precipitation appear to occur over a wider region stretching further south and into Southern Africa. In the LMD model, the observed convective activity over the Southern Indian Ocean appears shifted and focused into Southern Africa, lead-

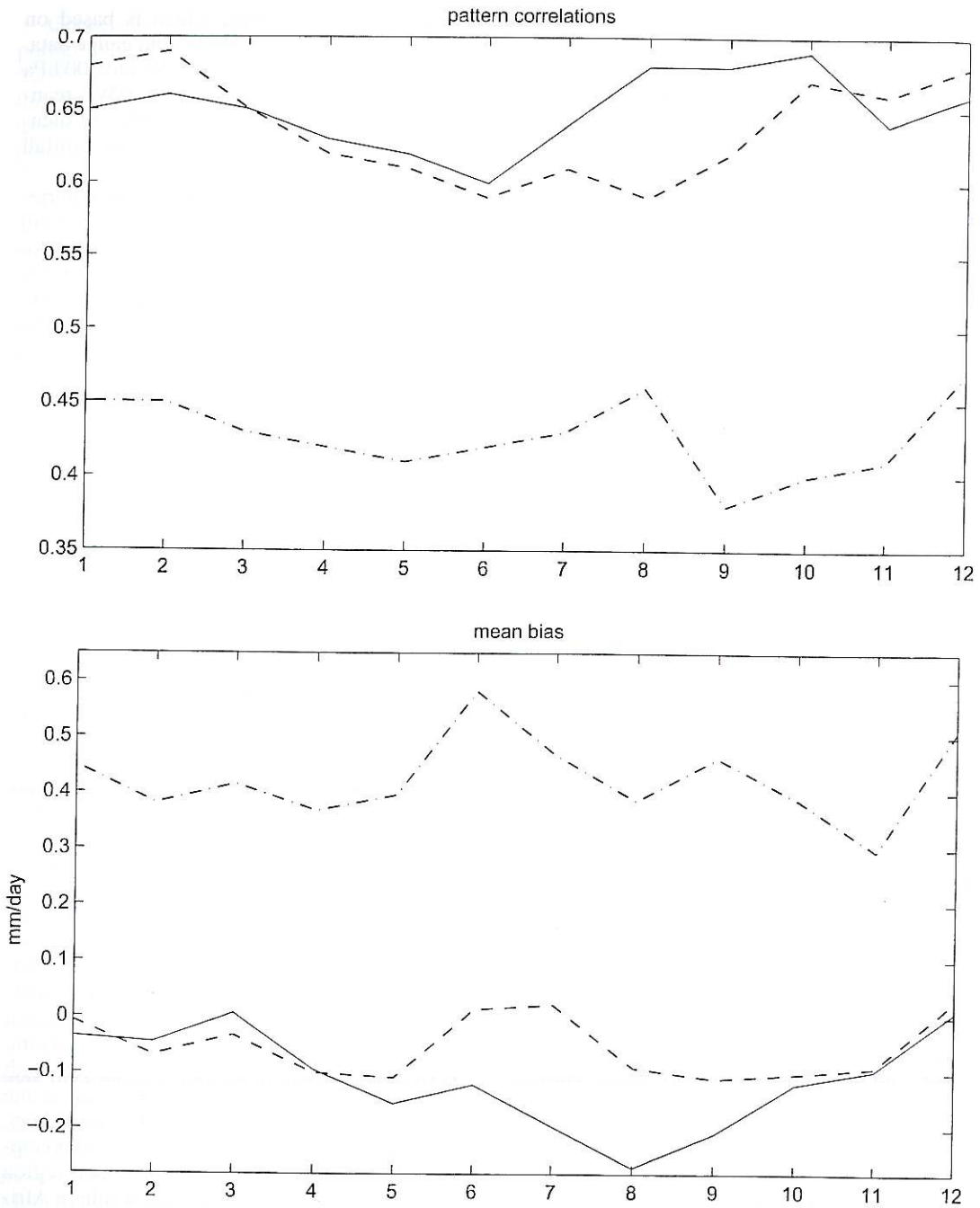

Fig. 1. Pattern correlations between monthly mean modelled rainfall and observed rainfall (extracted from the Xie-Arkin data set) on the tropical zone (30N-30S). The observations have been linearly interpolated on the corresponding model grid.

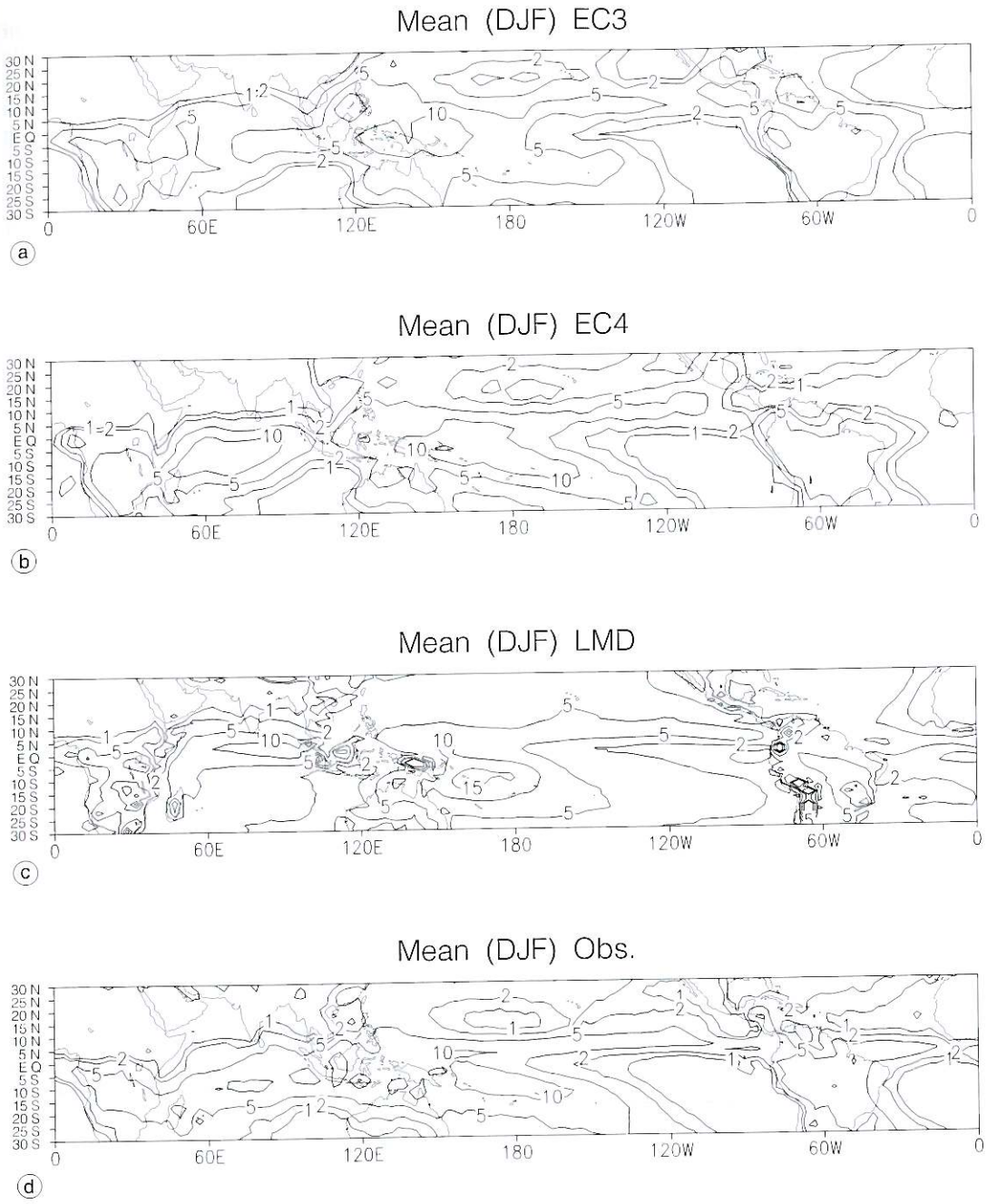

Fig. 2a-d. Mean seasonal rainfall in December-February for: a) EC3; b) EC4; c) LMD; d) observation in mm/day. The data are displayed on the original grid and isohyets are 1, 2, 5, 10 and 20 mm/day.

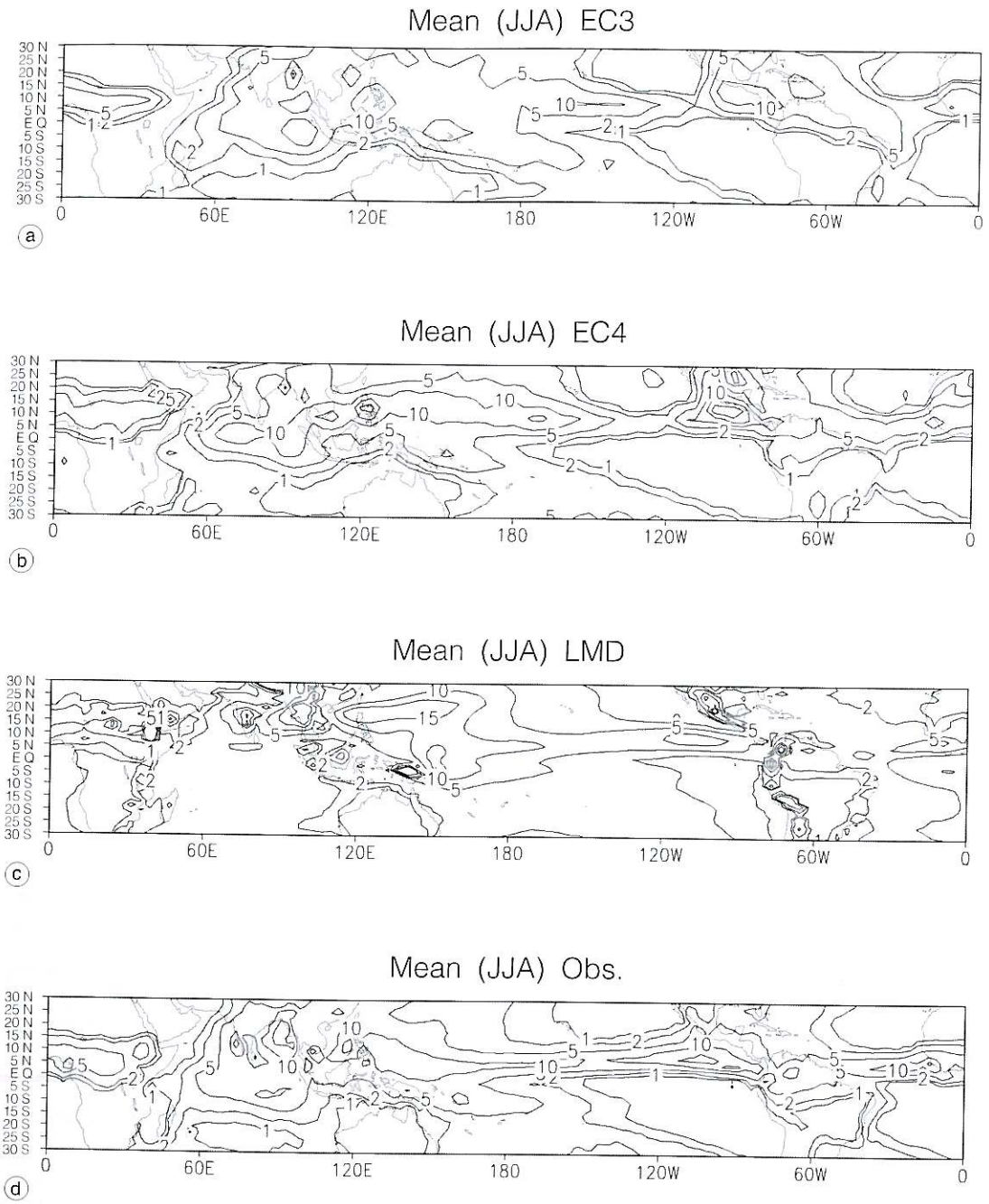

Fig. 3a-d. Same as fig. 2a-d, except for June-August.

ing to positive rainfall bias in this region. Thus, all three models have positive bias in Southern Africa, though the sources of this bias appear rather different given the model's contrasting performances in the adjacent tropical Atlantic and Indian Oceans.

In June-August, discrepancies between model and observations are stronger than in DJF (fig. 3a-d). EC4 performs well in the Pacific domain though the maximum on the eastern side of the basin is intensified and displaced northeastward into Central America and Mexico (fig. 3b). The Atlantic and West African ITCZ are well simulated by all three models, but with a slight northward displacement for EC4 relative to the observations. The main discrepancy for all models in JJA occurs in the context of the SE Asia monsoon system extending into the Western Pacific, and all models find their own solution for how to respond to this complex aspect of the annual cycle. EC4 maintains the centre of action of convection and precipitation much too far south over the Indian Ocean, and monsoon westerly flow is accordingly displaced south and penetrates too strongly into the tropical Northwestern Pacific (not shown). The situation for EC3 (fig. 3a) is somewhat improved over Indian longitudes, but the rainfall maximum in the northwestern tropical Pacific associated with extension of the monsoon flow into this region is not well simulated by EC3, which produces a general wide area of strong convection and rainfall over the Western Pacific. EC4 tends to produce too much precipitation and expands the zone of maximum precipitation north of the equator in the Western Pacific, as the monsoon flow penetrates too strongly into this region. The LMD produces extremely high rates of precipitation over this region – this strong positive bias is offset by, once again, a negative bias over the Central and Southern Indian Ocean. In between these two poles of positive and negative bias, the Indian sub-continent itself has quite accurate rainfall rates. Consistent with this, the Tropical Easterly Jet (TEJ) is too weak over the Indian Ocean for EC3 and EC4 but its intensity and location are well reproduced by LMD. The northward extension of upper-level easterlies is too weak in EC3 and 4 (not shown).

In summar:

i) The seasonal migration of the rainbelt associated with the monsoonal planetary-scale system is quite well reproduced by the three models, but the skill varies seasonally and spatially.

ii) LMD contains spurious peaks of rainfall, with sometimes excessive rainfall rates (over 80 mm/day) over small patches of mountains (Andean Cordillera, East Africa, etc.). It could be partly related to the high resolution and the grid-point (*versus* spectral) representation of the geography that could locally enhance the precipitation.

iii) The America-Africa longitudes, from the Eastern Pacific to the East Africa are quite well reproduced by the three models. In fact the main difficulty in this region is the Atlantic ITCZ itself which in DJF is almost absent for EC4 and too weak for EC3.

iv) The situation is more complicated from the Western Indian Ocean to the maritime continent. The seasonal migration of the rainbelt is well reproduced by both ECHAM models. The SPCZ is generally too weak for EC3. The highest precipitation zone is displaced over the NW tropical Pacific in boreal summer in EC4 and the rainfall remains too high over the equatorial Indian ocean in JJA. For LMD, the main rainbelt is too far north over the Indian ocean during SON through MAM, and the extent of the highest rainfall zone is too small in JJA.

4. Simulation of tropical indices

First, some general inferences are made from table I, then some specific points are made for each of the indices. We use the multiple regression least-squares fit equation to choose the weights to give to each of the three runs

$$\text{obs}' = (a \times \text{run1}) + (b \times \text{run2}) + (c \times \text{run3}) \tag{4.1}$$

where a, b, and c are the weights given to each of the runs, to try and optimally combine them to yield the best prediction obs'. The predicted obs' are generated using cross-validation, to try and get an unbiased estimate of the skill when

Table I. Cross validated correlation between observed and simulated tropical circulation and rainfall indices. Cross-validated correlation is obtained from the linear fit of the observations with runs. Negative coefficients are checked for (which would imply that a run was forecasting opposite anomalies relative to observation) and set to zero during the cross-validation. The first column gives the mean of the 27 combinations of three runs, with one run of each model. The second column gives the correlations obtained with the mean of the runs (it is equivalent to assign the same weight for the nine runs). The third column gives the correlations obtained with an optimal weighting of the three AGCM means. The fourth column gives the correlations obtained with an optimal weighting of the nine runs. For each individual model (columns 5-7), the first value gives the correlation obtained from the regression with the optimal weighting of the three runs and the second one, the correlation obtained from the regression with the mean of the three runs (it is equivalent to assign the same weight for the three runs). There is no value for WAF monsoon index of LMD because the coefficients are always set to zero. The highest value of each line is in bold.

Indices	Mean of 27 combinations	Super	Super 3	Super 9	EC3	EC4	LMD
SOI (monthly 61-93)	0.85	**0.87**	**0.87**	**0.87**	0.83 / 0.84	0.81 / 0.81	0.86 / 0.86
SEA (JJA 63-89)	0.40	0.55	0.57	0.57	0.23 / 0.36	**0.64** / 0.57	0.28 / 0.34
WAF (JAS 63-89)	0.42	0.36	0.51	0.50	0.46 / **0.53**	0.47 / 0.34	–0.23 / *
NOR (MAM 61-93)	0.67	**0.77**	0.75	0.73	0.61 / 0.62	0.58 / 0.65	0.72 / 0.73
SAH (JAS 61-93)	0.32	**0.52**	0.48	0.33	0.47 / **0.52**	0.10 /–0.02	0.48 /–0.15
IND (JJA 61-93)	0.35	0.40	0.24	**0.58**	0.44 / 0.31	0.35 / 0.16	0.41 / 0.25

combining the three runs. A 12-month period is removed each time to generate cross-validated SOI predictions (since we have monthly values for SOI), whereas one year is removed each time for the other indices, which only have one value per year. Negative coefficients are checked for (which would imply that a run was forecasting opposite anomalies relative to observation) and set to zero. It is a penalty to not inflate the explained variance by the AGCM when the sign of the model is wrong. For example, a model which forecasts quite systematically opposite anomalies relative to observation could appear as skilful if such a penalty is not applied. For a first indication in this paper, the skill score is defined as the correlation between the predicted observation (obs') and the actual observation.

First, in table I, consider the column LMD. The results for this column were generated by using, in eq. (4.1), the three LMD runs from different initial atmospheric conditions. Equation (4.1) generates the weights for an optimum combination of the three runs. In fact, we expected this to give equal weight to each of the three runs, since the only difference between the runs is the initial atmospheric conditions and there is no reason to think that a run is better than another one on the long term. We return to this issue below. The EC3 and EC4 columns were generated in a similar way, applying eq. (4.1) to the three sets of runs for each model.

Next, we generated super-ensemble forecasts by using the ensemble mean of each of the three models in eq. (4.1) (so now, in eq. (4.1), run 1 is the ensemble mean of EC3, run 2 is the ensemble mean of EC4 and run 3 is the ensemble mean of LMD). The column Super-3 shows the skill of these forecasts. We have also generated super-ensemble forecasts by using the nine runs with possible varying weights amongst them (column Super 9 in the table I) and by using the mean of the nine runs, which is equivalent to assigning the same weight to each run (column Super in table I). Results are most stable for the most skilful indices, SOI and Nordeste. For these indices the Super, Super-3 and Super-9 skills are higher than any of the individual models. However, the comparison is not really fair, because 9 runs have contributed to Super, Super-3 and Super-9, whereas only 3-runs contribute to each of the individual model scores. To make a

fairer comparison, we assessed the skill of predictions made by combining one run from each of the three models. In fact, we repeated this analysis using all 27 possible 3-run combinations always taking one run from each of the three models. The mean skill of these predictions is given in the first column of table I.

For the six indices (except IND), taking one run from three different models gives a higher average skill than the average skill when considering three runs from the same model. The mean of the three individual AGCM optimal linear combination of three runs (*i.e.* first values of columns 5-7 of table I) equals 0.83, 0.38, 0.23, 0.64, 0.27, 0.40 respectively for SOI, SEA, WAF, NOR, SAH, IND (compare to the first column of table I). The difference is particularly large for WAF where the performance of an AGCM (*i.e.* LMD here) is close to zero. A striking result is also the varying skill for each model when the mean of the three runs or an optimal combination of the three runs is considered (see the columns 5-7 of table I). Usually, both values are very close to each others, especially for the most skilful indices (as SOI, NOR), meaning that each run has the same long-term skill, even if there are possible differences for a given particular year. Sometimes, the optimal weighting of the three runs outperforms strongly the mean of the three runs (as for IND). The reverse situation could also occur (as for SEA for EC3). The robustness of these results should be also tested with different lengths of the training/verification periods during the cross-validation. This requires further investigation with a larger ensemble of models and different indices. However, even if in some situations, taking runs from one or two of the most skilful models is the best way to generate predictions, table I still endorses the advantage of having many models operational (see also Krishnamurti *et al.*, 2000). It is because each of these models turns out to have the highest skill for at least one of the indices.

Thus the results in this section have provided evidence that having more than one model available should allow us to generate an improved forecast. However, for a full interpretation of table I, we were forced to acknowledge that the results also suggest that individual runs from the same model should now be examined in detail to assess whether systematic differences develop leading to systematically different skill levels amongst the runs.

4.1. *Southern oscillation index*

The model simulations of the SOI (with a five-month running mean) are shown in fig. 4a. For table II, raw monthly mean values are used, to demonstrate the differing annual cycles of simulation skill in the models. Many authors have investigated the so-called «spring predictability barrier» in observations and reanalyses (*i.e.* Webster and Yang, 1992; Balsameda *et al.*, 1994; Lau and Yang, 1996), and its expression in weaker coupling between ocean and atmosphere at this time of year. This feature shows up in all models in both correlation skill and the percentage of external variance in the SOI, relative to the intra-model internal atmospheric variance (Rowell *et al.*, 1995; Rowell, 1998) (table II). The drop of simulation skill in EC4 is particularly marked, with SST-forced variance in the SOI almost disappearing, in contrast to EC3 and LMD, where the external SST forced variance is better maintained and the drop of correlation skill is less. All models show a secondary drop in skill and external variance in October.

4.2. *Boreal summer monsoon wind-shear indices*

There is a great discrepancy between different runs for SE Asia (fig. 4b). The internal noise due to atmospheric dynamics and/or land-atmosphere interaction is clearly high. In SE Asia, several years appear to be well reproduced (that is, almost every run is of the same sign and consistent with observations) such as 1969-1974, 1983-1989 but others are poorly reproduced such as 1965, 1968, 1979 (fig. 4b). Concerning the West Africa monsoon index (fig. 4c), skill is better for EC3 (table I). All runs reproduce fairly well the long-term negative trend (*i.e.* weakening of the zonal overturning over West-Africa with easterly anomalies at 850 hPa and/or westerly anomalies at 200 hPa) which

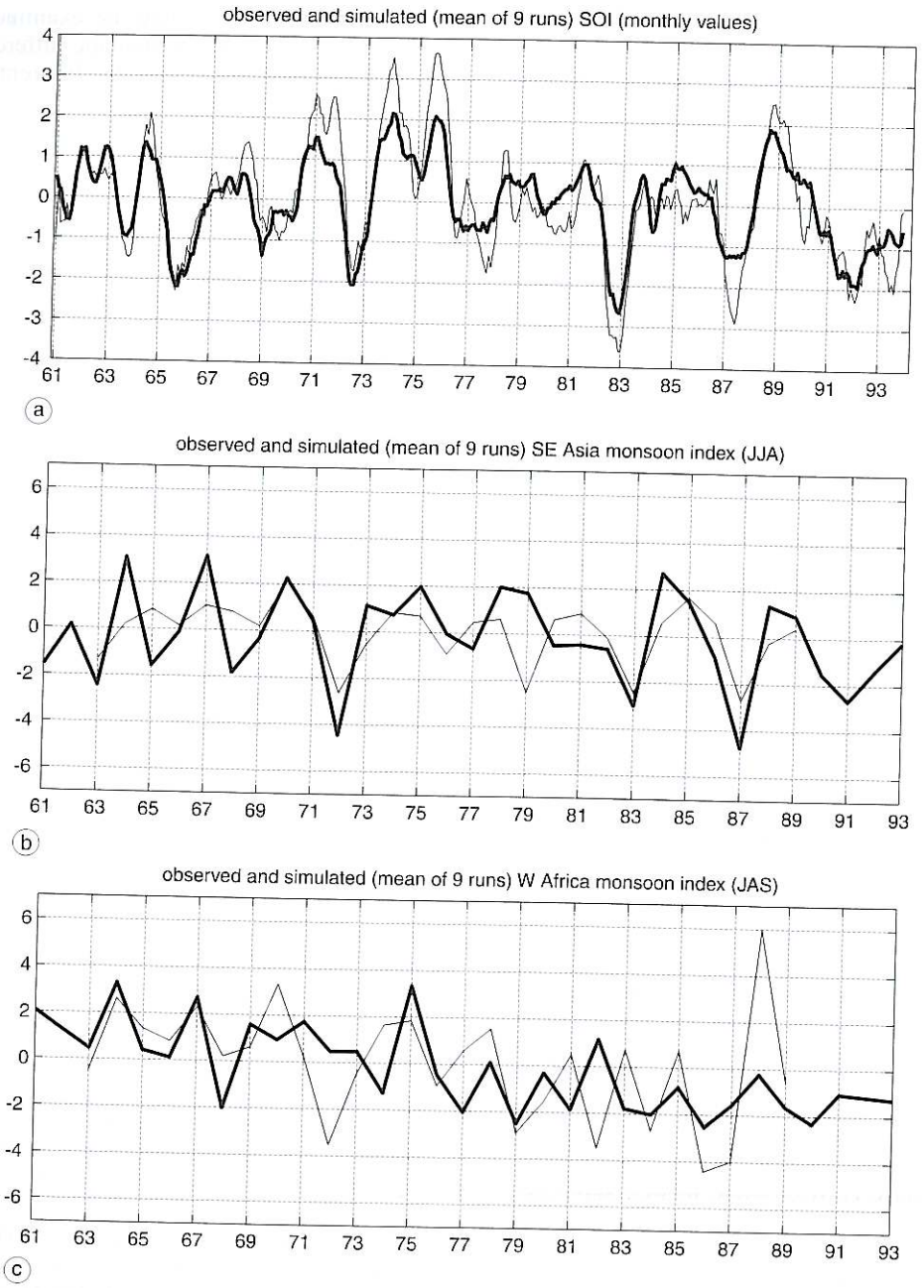

Fig. 4a-c. a) Simulated and observed Southern Oscillation Index; b) simulated south-east monsoon index; c) simulated and observed West Africa monsoon index (see text for the definition of the indices). The simulated indices are defined by the mean of the nine runs (bold line) the observed one (thin line) comes from the UKMO SLP data set and from the GDFL atmospheric data set (Oort and Liu, 1993).

Table II. Skill of the monthly simulated SOI (= correlation between observed and simulated SOI). The observed SOI are copied three times and the whole runs are used. The external variance computed from the formula of Rowell (1998) is also displayed.

	EC3 Skill	EC3 External variance (%)	EC4 Skill	EC4 External variance (%)	LMD Skill	LMD External variance (%)
January	0.60	64	0.66	62	0.52	62
February	0.56	53	0.35	58	0.42	47
March	0.53	62	0.44	36	0.27	39
April	0.52	64	0.29	19	0.56	53
May	0.40	67	0.34	8	0.47	55
June	0.49	47	0.44	41	0.63	63
July	0.61	49	0.48	57	0.65	66
August	0.50	50	0.51	64	0.56	82
September	0.70	63	0.65	62	0.70	70
October	0.53	40	0.53	35	0.62	61
November	0.58	65	0.55	52	0.66	66
December	0.65	75	0.50	59	0.59	70

is associated with the rainfall decrease observed over Soudanian and Sahelian belts during this period (*i.e.* Moron, 1994). Inter-annual skill is less consistent. EC3 reproduces well 1971-1972 and 1987-1988. This is less clear for the other AGCMs (not shown).

4.3. *Tropical rainfall indices*

As found in many other studies, skill appears to be very unstable for India and Sahel and more stable for Nordeste (Sperber and Palmer, 1996; Rowell *et al.*, 1995; Gadgil and Sajani, 1998; MNWR98) (table I and fig. 5a-c). The decreasing Sahelian rainfall trend is fairly well reproduced by EC3 (fig. 5b). Inter-annual variability is poorly reproduced especially in 1967, 1972, 1982, 1987 and 1988, related to all of these model's systematic error in the Sahel-ENSO teleconnection (see Section 5). The error is particularly strong for LMD and EC4. For India, skill improves but remains relatively low (table I and fig. 5a). For the Sahel and India rainfall indices, where average skill is low, we have analysed the simulations in more detail. We have searched for years where all runs (except at maximum, one), are $< +0.25$ standard deviations (SDs) (defining examples of consistently below normal simulations) or > -0.25 SDs (above normal simulations). For the Sahel, we obtain six above normal cases (1961, 1962, 1963, 1971, 1972, 1986) and seven below normal ones (1974, 1977, 1979, 1980, 1984, 1988, 1989). The best case is represented by 1984 where eight out of nine runs are actually below -0.25. On the 13 consistent model cases, five are completely false relatively to observations (that is the model forecasts are consistent but of the wrong sign) in 1972, 1974, 1986, 1988 and 1989. The best cases (1979 and 1984) are associated with strong equatorial Atlantic warming (Servain, 1991; Wagner and Da Silva, 1994), but overall, the spread appears not to be a good indicator of Sahel skill. As will be returned to in the later sections, this can be expected to be related to the fact that all three models tend to produce the opposite sign teleconnection between Sahel and ENSO, relative to that found in observations. For India, 15 years are consistent amongst the nine runs: eight are positive (1964, 1969,

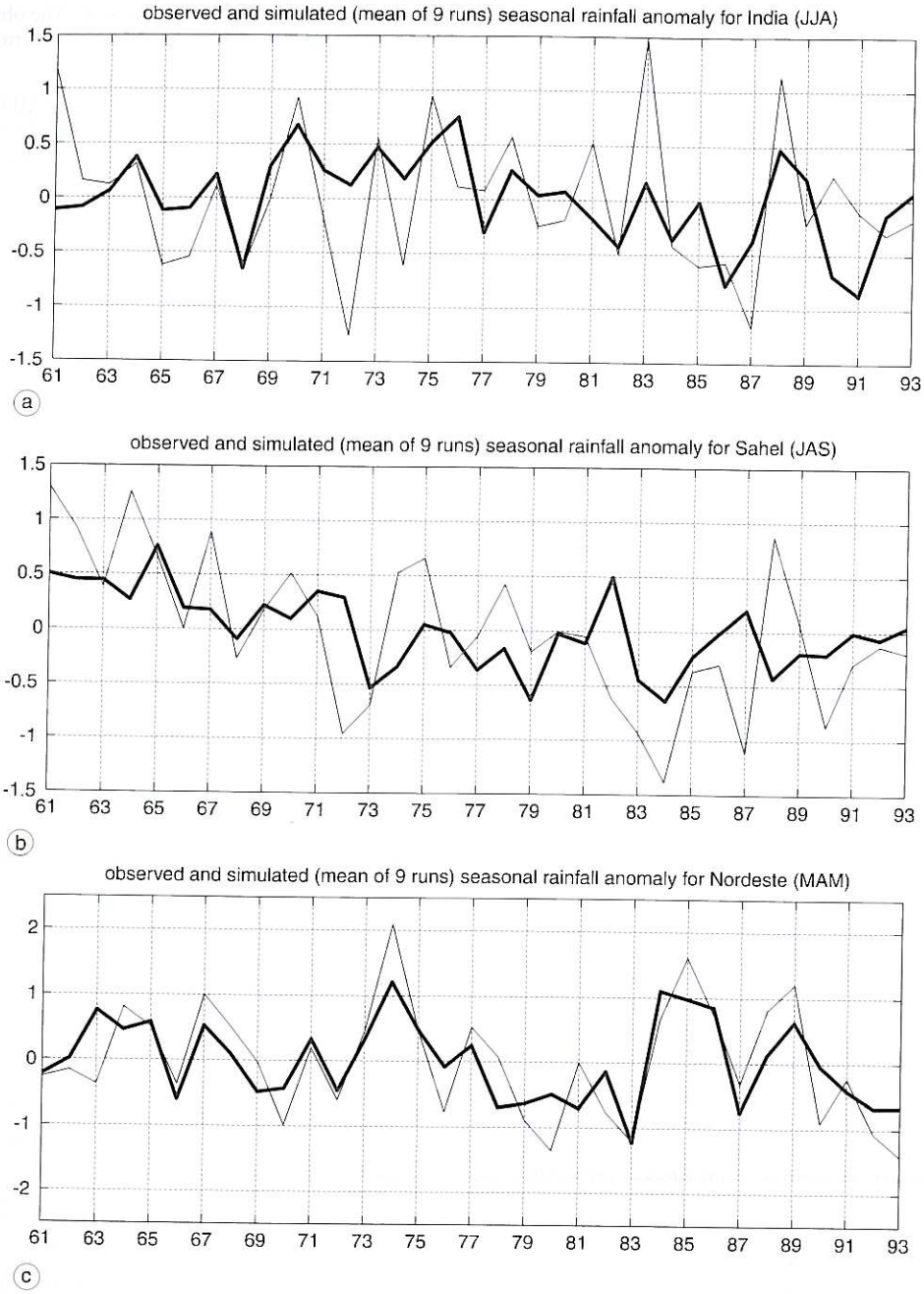

Fig. 5a-c. Same as fig. 4a-c except for: a) India (June-August); b) Sahel (July-September); c) Nordeste (March-May) seasonal rainfall anomalies. The simulated indices are defined by the mean of the nine runs (bold line) and the observed ones are defined from the Hulme data set (Hulme, 1991).

1970, 1971, 1973, 1975, 1976, 1988), seven are negative (1968, 1982, 1984, 1986, 1990, 1991, 1994). The most consistent cases are 1970 (8/9 runs are above 0.25) and 1994 (8/9 runs are below −0.25). The model's sign is wrong in 1971, 1990 and 1994, but otherwise, the years with a consistent response tend to be skilful. These results are preliminary, and require more runs to assess statistical significance (Rowell, 1998). Nonetheless, they suggest some year to year variations in the SST-forced component of the Indian and West-African monsoons, a concept that has received recent attention in the context of ENSO (Kumar and Hoerling, 1998; Montroy, 1997).

5. Seasonal OM and MM modes

We apply in this paragraph the analysis in MM and OM mode following MNWR98. The first MM mode is almost always similar to the OM_m rainfall pattern with spatial correlations for the tropical area from 0.92 to 0.96. The MM_m patterns are not shown here for that reason. The summary statistics are shown in table III.

5.1. *DJF*

The basic features of OM_o are very similar between the three models and the super-ensem-

Table III. Statistics of the first mode of OM analyses. V* (obs) indicates the explained variance (from the heterogeneous correlations between the mean of the cross-validated scores of OM_m and the observed field), V* (mod), the explained variance (from the heterogeneous correlations between the cross-validated scores of OM_o and the mean of the modelled field), Skill, the correlation between the cross-validated scores of OM_o and OM_m and Skill*, the correlation between the ensemble-mean of the cross-validated scores of OM_m and OM_o. Spatial match is the pattern correlation between OM_o and OM_m (after having interpolated linearly OM on the corresponding model grid). The spatial match can be considered as a measure of the extent to which the model's ENSO teleconnection pattern matches that of the observed ENSO teleconnection pattern. Reproducibility (*i.e.* external variance in the sense of Rowell, 1998) is estimated from MM analyses.

	V* (obs)	V* (mod)	Skill	Skill*	Spatial match	Reproducibility (%)
EC3 - DJF	8.2	10.9	0.80	0.82	0.48	92.3
EC4 - DJF	8.3	14.7	0.81	0.85	0.33	89.1
LMD - DJF	8.0	11.0	0.77	0.80	0.36	92.2
SUP - DJF	8.1	15.2	0.78	0.83	0.49	90.7
EC3 - MAM	6.2	9.5	0.80	0.84	0.43	85.1
EC4 - MAM	6.0	11.0	0.74	0.82	0.43	87.9
LMD - MAM	5.9	9.5	0.76	0.77	0.36	93.5
SUP - MAM	6.3	12.5	0.76	0.85	0.55	80.5
EC3 - JJA	6.2	8.5	0.79	0.82	0.43	90.1
EC4 - JJA	6.0	9.7	0.73	0.77	0.34	73.0
LMD - JJA	5.5	12.1	0.68	0.74	0.25	76.4
SUP - JJA	6.1	11.4	0.77	0.82	0.46	86.2
EC3 - SON	8.2	10.8	0.84	0.85	0.53	95.6
EC4 - SON	8.3	11.0	0.84	0.87	0.42	92.2
LMD - SON	7.7	11.9	0.76	0.80	0.27	85.6
SUP - SON	8.4	14.5	0.83	0.87	0.49	92.0