UCRL-JC-129367

PCMDI

Program for Climate Model Diagnosis and Intercomparison

PCMDI Report No. 48

ARE REVISED MODELS BETTER MODELS? A SKILL SCORE ASSESSMENT OF INTERANNUAL VARIABILITY

Kenneth R. Sperber[1] and Participating AMIP Modelling Groups [2-11]

[1]Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
Livermore, CA 94551 USA

[2]Bureau of Meteorology Research Centre, Melbourne, Australia
[3]Centre National de Recherches Meteorologiques, Toulouse, France [4]Department of Numerical Mathematics, Moscow, Russia
[5]Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[6]Laboratoire de Meteorologie Dynamique, Paris, France
[7]Max Planck Institut für Meteorologie, Hamburg, Germany
[8]Meteorological Research Institute, Tsukuba, Japan
[9]Naval Research Laboratory, Monterey, CA, USA
[10]State University of New York, Albany, NY USA
[11]Yonsei University, Seoul, Korea

November 1998

Program for Climate Model Diagnosis and Intercomparison
University of California, Lawrence Livermore National Laboratory
Livermore, CA 94551 USA

Are Revised Models Better Models?
A Skill Score Assessment of Interannual Variability

Kenneth R. Sperber[1] and Participating AMIP Modelling Groups[2-11]

[1]**Program for Climate Model Diagnosis and Intercomparison**
Lawrence Livermore National Laboratory
Livermore, CA 94551 USA

[2]Bureau of Meteorology Research Centre, Melbourne, Australia
[3]Centre National de Recherches Meteorologiques, Toulouse, France
[4]Department of Numerical Mathematics, Moscow, Russia
[5]Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[6]Laboratoire de Meteorologie Dynamique, Paris, France
[7]Max Planck Institut für Meteorologie, Hamburg, Germany
[8]Meteorological Research Institute, Tsukuba, Japan
[9]Naval Research Laboratory, Monterey, CA, USA
[10]State University of New York, Albany, NY USA
[11]Yonsei University, Seoul, Korea

Corresponding Author:
Dr. Kenneth R. Sperber
Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory
P.O. Box 808, L-264
Livermore, CA 94551
E-mail: sperber@space.llnl.gov

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (615) 576-8401, FTS 626-8401

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161

Abstract

Various skill scores are used to assess the performance of revised models relative to their original configurations. The interannual variability of all-India, Sahel and Nordeste rainfall and summer monsoon wind shear is examined in integrations performed under the experimental design of the Atmospheric Model Intercomparison Project. For the indexes considered, the revised models exhibit greater fidelity at simulating the observed interannual variability. Additionally, the revised models have an improved signal-to-noise ratio (with the exception of Sahel rainfall). The results suggest that changing a models convection scheme is more beneficial for the simulation of tropical interannual variability than modification of the land surface package. Improvement in the simulation of interannual variability is directly related to the reduction of systematic error in the mean state.

1. Introduction Standardized experimentation and sensitivity testing are prerequisites for ascertaining the fidelity of new physical and dynamical elements of model formulations. To this end, the climate modelling community has adopted the Atmospheric Model Intercomparison Project (AMIP, Gates 1992), a World Climate Research Program/Working Group on Numerical Experimentation initiative, as one of many vehicles for model validation. The benefit of the AMIP experimental design is that models are integrated in a standardized fashion, all using the same SST boundary conditions, solar constant and CO_2 concentration for the 10-year period 1979-88. Thus, differences among the simulations are directly attributable to the wide variety of model formulations. In an effort to evaluate systematic error and perform process studies with this suite of integrations, AMIP diagnostic subprojects, the modelling groups, and the staff of the Program for Climate Model Diagnosis and Intercomparison have performed a wide variety of analyses associated with diurnal through interannual time scales for tropical, extratropical, tropospheric and stratospheric phenomena (see Gates et al. 1998 for a summary of AMIP).

Given that the period of the integrations incorporated two strong El Niño events (1982/83, 1986/87) and one La Niña event (1987/88), it is of interest to investigate the ability of the models to simulate El Niño/ Southern Oscillation (ENSO) teleconnections. The Tropical Ocean Global Atmosphere Monsoon Numerical Experimentation Group (TOGA MONEG, WCRP 1992, 1993) investigated the ability of many atmospheric general circulation models (GCMs) to simulate boreal summer Indian monsoon and Sahel rainfall variability during the ENSO extremes of 1987 and 1988 in seasonal length integrations. Based upon an AMIP diagnostic subproject proposed by MONEG, Sperber and Palmer (1996) evaluated the ability of 32 AMIP GCMs to simulate interannual variability over the afore-mentioned regions, as well as the interannual variations of boreal spring rainfall over northeast Brazil (Nordeste). The majority of these models analyzed were vintage 1990-93. While this work served to compare and contrast the ability of the models to represent aspects of tropical rainfall and circulation, the real benefit of these integrations is that they serve as a benchmark against which model development can be assessed.

Table 1: Revised Model Attributes (modified convection scheme; modified land-surface scheme)

Modelling Group	Acronym	Resolution	Changes relative to original submission
Bureau of Meteorology Research Centre, Australia	BMRC_A	R31L9	Tiedtke replaces Kuo
Bureau of Meteorology Research Centre, Australia	BMRC_B	R31L9	BASE land surface scheme implemented
Centre National de Recherches Météorologiques, France	CNRM_A	T42L30	Arpège rather than Emeraude: Land surface scheme, clouds, GWD, etc.
Department of Numerical Mathematics, Russia	DNM_A	4^ox5^oL7	Moist convective adjustment rather than Kuo
Geophysical Fluid Dynamics Laboratory/Dynamic Extended-Range Forecasting, USA	DERF_A	T42L18	bucket model replaced by 3-layer soil model
Laboratoire de Météorologie Dynamique, France	LMD_A	50 sin(lat) x 64 lon L11	Sechiba surface scheme, diurnal cycle. surface drag formulation
Laboratoire de Météorologie Dynamique, France	LMD_B	50 sin(lat) x 64 lon L11	7 layer soil thermodynamics using the same bucket model as the original model, diurnal cycle. surface drag formulation
Max Planck Institut für Meteorologie, Germany	MPI_A	T42L19	ECHAM4 rather than ECHAM3: Nordeng convection replaces Tiedtke, etc.
Meteorological Research Institute, Japan	MRI_A	4^ox5^oL15	modified the gravity-wave drag generation factor
Naval Research Laboratory, USA	NRL_A	T47L18	modification of shallow convection
State University of New York at Albany/National Center for Atmospheric Research/Genesis, USA	SNG_A	T31L18	change in radiation scheme, cloud emissivity and cloud formation
Yonsei University, South Korea	YONU_A	4^ox5^oL5	Vertical resolution increased to 7 layers from 5

The models and their associated (major) revisions are described in section 2, and the standardization and skill scores are described in section 3. The comparison of original and revised models is presented in section 4, and the sensitivity to physical parametrization is assessed in section 5. The role of systematic error and its relation to interannual variability is presented in section 6, and the conclusions are given in section 7.

2. The models Subsequent to their initial AMIP submissions, ten modelling groups contributed twelve additional integrations with revised models that are typically representative of mid-1990's development. Select attributes of the revised models are given in Table 1. In this table are listed the modifications that are likely to be associated with the most substantial changes to the integrations. BMRC_A, DNM_A, MPI_A and NRL_A modified their convection schemes, a revision that demonstrably affects rainfall and its variability (Slingo et al. 1994). BMRC_B, CNRM_A, DERF_A and LMD_A changed their land surface schemes, which through feedback with the circulation may affect variability. For the majority of models, changes over and above those listed in Table 1 were also incorporated in the revised models. Comprehensive documentation of the original AMIP models is given in Phillips (1994). This information, and the complete suite of changes made to the revised models can be found at: "http://www-pcmdi.llnl.gov/modeldoc/amip/01toc.html".

3. Standardization and Skill Scores In order to account for the different amplitudes of interannual variability among the models, as well as their individual biases in simulating the time-mean, each area-weighted index has been standardized by removal of the time mean and division by the standard deviation of the interannual variability. All results are based on these standardized indexes.

Given the large number of models analyzed, succinct measures of skill are desirable to provide a quantitative measure of performance. The Brier score has been used as a measure for assessing numerical weather prediction performance (Murphy and Katz 1985, Perrone and Miller 1985). In this application it measures the skill of an ensemble of models to simulate the correct sign of a seasonal anomaly with respect to observations over the 10-year period of the integrations. It is required that the observed standardized departure for a given year i exceed +/- 0.25 to be included in the calculation of the Brier score, in which case the number of years n_y over which the average Brier score is calculated may be less than ten. The average Brier score is calculated as follows:

$$\overline{Bs} = \frac{1}{n_y}\sum_{i=1}^{n_y}[(1.0 - Y_i)^2 + (0.0 - N_i)^2] \quad (1)$$

which reduces to

$$\overline{Bs} = \frac{1}{n_y}\sum_{i=1}^{n_y} 2N_i^2 \quad (2)$$

where:

Y_i = fraction of models simulating an anomaly of the correct sign during year i

N_i = fraction of models simulating an anomaly of the incorrect sign during year i,

and $N_i = (1 - Y_i)$

Brier scores may range from 0.0 (a perfect score) to 2.0 (total disagreement with observations). The Brier score of a climatological forecast (Y_i = 0.5, N_i =0.5) is 0.5. The significance levels of the Brier scores are estimated using the Monte Carlo technique from which the probability distribution function of the Brier scores is calculated by randomly sampling the simulated indexes 100,000 times.

While the Brier score is a model-observed data verification tool, reproducibility, which is essentially a signal-to-noise ratio, is a model-model intercomparison. The realizations intercompared may consist of different models, or multiple realizations of the same model that differ only in the specification of initial conditions. The reproducibility is a measure of the models' ability to respond robustly (not necessarily correctly) to the imposed boundary forcing. Using the standardized indices of each model, y_i, we calculate the ensemble mean time series for years i = 1, 2,...10 as follows:

$$\langle y_i \rangle = \frac{1}{m} \sum_{\alpha = 1}^{m} y_{i\alpha} \qquad (3)$$

where m = the number of models considered. Noting that the time mean of $\langle y_i \rangle$ is zero since the standardized indices were used in Eq. 3., we next calculate the variance of the ensemble mean time series:

$$\sigma_{tsig}^2 = \frac{1}{9} \sum_{i = 1}^{10} \langle y_i \rangle^2 \qquad (4)$$

σ_{tsig}^2 is a measure of the interannual variability of the signal extracted from the suite of simulations considered, and it is the numerator of the reproducibility. The spread of the ensembles during a given year is associated with the unpredictability of the signal. The denominator of the reproducibility is the average of the variance based upon the simulated anomalies from each year i from the m prognostications. It is a measure of the unpredictable portion of the signal $\langle y_i \rangle$, and is calculated as follows:

$$\overline{\sigma^2_{noise}} = \frac{1}{10}\sum_{i=1}^{10}\left[\frac{1}{(m-1)}\sum_{\alpha=1}^{m}(y_{i\alpha}-\langle y_i \rangle)^2\right] \quad (5)$$

The reproducibility is:

$$R = \frac{\sigma^2_{tsig}}{\overline{\sigma^2_{noise}}} \quad (6)$$

When denominator equals zero, each realization is an exact replica and the reproducibility equals ∞.

The statistical significance of reproducibility can be easily tested since this quantity is related to the F-distribution for testing variances of two normally distributed populations. When:

$$mR > F[n-1, n(m-1), a] \quad (7)$$

where n is the number of years of data, m is the number of simulations, and a is the level of significance, we may reject the null hypothesis that sea surface temperature has no effect on the interannual variations of the indexes.

4. Revised vs. Original Simulations The ability of the models to simulate rainfall variability over India, the Sahel and Nordeste, and the wind shear over the Indian Ocean during the Asian summer monsoon is assessed as in Sperber and Palmer (1996). In order to make a direct comparison, the analysis is performed on the subset of original AMIP models for which there is a revised model. Early work by Walker (1928) and Walker and Bliss (1932) established that Indian monsoon rainfall is linked to the Southern Oscillation. Simulation of interannual variability of Indian monsoon rainfall has proven to be a challenge (e.g., WCRP 1992, 1993, Sperber and Palmer 1996, Gadgil and Sajani 1998). Variations in monsoon wind shear (e.g., Webster and Yang 1992), Sahel rainfall (e.g., Lamb 1978, Folland et al. 1986, Lough 1986, Palmer

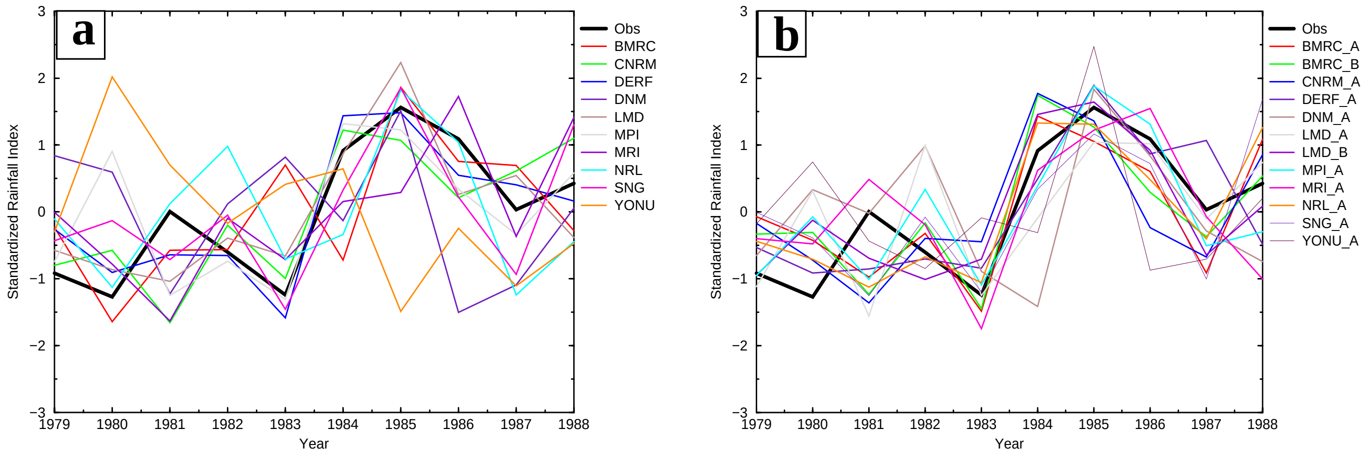

Figure 1: Simulated (MAM) and observed (MA) averaged Nordeste rainfall indexes for (a) the original AMIP simulations; (b) the revised AMIP simulations.

1986, Nicholson and Entekhabi 1986, Hastenrath 1988, and Folland et al. 1991), and Nordeste rainfall (e.g., (Hastenrath and Heller 1977, Moura and Shukla 1981, Ropelewski and Halpert 1987, Aceituno 1988, Ward et al. 1988, Ward and Folland 1991, and Sperber and Hameed 1993) are linked to ENSO and other regional SST variations (Sperber and Palmer 1996). Given the importance of rainfall to these mainly agrarian societies and the tendency for the rainfall amounts to typically be extreme, constituting either flood or drought conditions, the skill of the revised models is assessed relative to their original counterparts.

Figure 1 shows the March-May (MAM) averaged area-weighted land-only Nordeste rainfall indexes from the original and revised AMIP models. The ensemble dispersion of the simulated rainfall is large, particularly for the original models. Even so, the models capture the envelope of the observed rainfall variability. It is clear that the revised integrations (Fig. 1b) exhibit greater consistency than the original models (Fig. 1a) in their ability to simulate the observed (March-April) variations (Hastenrath 1992, personal communication), the thick solid black line in each panel of Fig. 1. To quantify this, the Brier and reproducibility skill scores are shown in Tables 2 and 3. The reduced Brier scores and the improved level of significance of the revised models in Table 2 indicate that they are better able to capture the observed interannual variations of June-September (JJAS) all-India rainfall, the JJAS averaged monsoon wind shear (40-100^{o}E, 0-25^{o}N, similar to Webster and Yang 1992), the July-Sep-

Table 2: Brier skill scores for the indexes. Based on Monte Carlo experiments the significance levels of the skill scores are given in parentheses. A Brier score of 0.0 indicates perfect agreement with observations. Years in brackets are those that were excluded from the calculation of the Brier score since the observed anomaly was near zero (see text).

Model Set	India Rainfall [1984]	Monsoon Wind shear [1982]	Sahel Rainfall [1985, 1986]	Nordeste Rainfall [1981, 1987]
Original	0.33 (.017)	0.27 (.001)	0.64 (.781)	0.13 (.000)
Revised	0.31 (.008)	0.20 (.000)*	0.58 (.664)	0.08 (.000)

*wind shear not available from DNM_A

Table 3: Reproducibility ratios for the simulated indexes. Ratios in excess of 1 indicate that the signal of the ensemble mean time series exceeds that of the internal chaotic variability. Reproducibility is related to the F-distribution for testing variances of two normally distributed populations. The significance levels at which we may reject the null hypothesis that the sea surface temperature has no effect on the interannual variations of the indexes are given in parentheses.

Model Set	India Rainfall	Monsoon Wind Shear	Sahel Rainfall	Nordeste Rainfall
Original	0.22 (.05)	0.86 (.005)	0.23 (.025)	0.50 (.005)
Revised	0.49 (.005)	1.37 (.005)*	0.15 (>.05)	1.53 (.005)

*wind shear not available from DNM_A

tember (JAS) averaged Sahel rainfall and MAM averaged Nordeste rainfall. For example, in the case of all-India rainfall for the revised models, the Brier score of 0.31 is smaller (better, recall a Brier score of 0.0 indicates perfect agreement with observations) than 99.2% of the Brier scores estimated from the Monte Carlo generated probability distribution, while for the original models the Brier score (0.33) is poorer, as is its significance, since it is smaller than 98.3% of the Brier scores estimated from the Monte Carlo generated probability distribution. The Brier score of the Sahel rainfall improves for the revised models, but since it is greater than 0.5 this indicates that the models are still not as skillful as climatology.

Table 3 indicates that the reproducibility of the revised models improves for the all-India rainfall, the monsoon wind shear, and Nordeste rainfall. The greater consistency of the revised models at simulating Nordeste rainfall (Fig. 1) is reflected by the their dramatic improvement in reproducibility noticed in Table 3. The increased reproducibility of the revised models is reflected in their higher level of statistical sig-

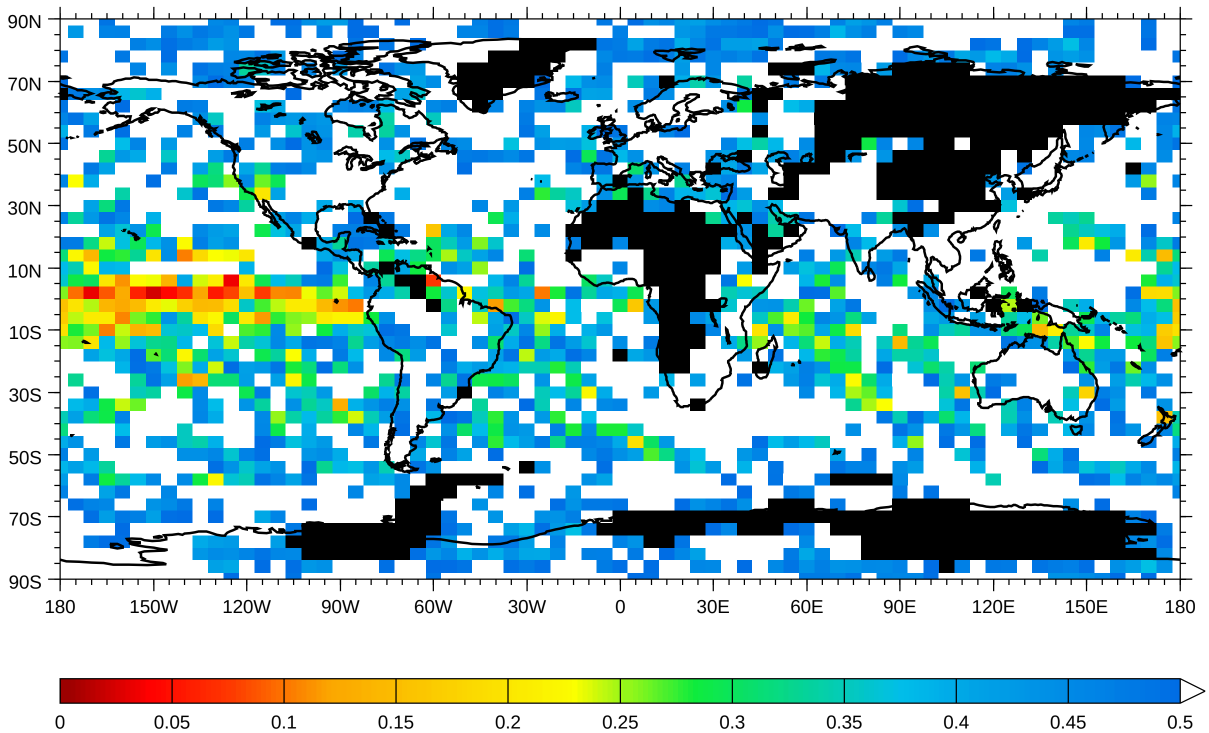

Figure 2: Brier skill scores of JJAS averaged precipitation for the revised models relative to MSU/rain gauge data. Brier scores worse than climatology (>0.5) are unshaded, and areas shaded black correspond to locations where Brier scores could not be calculated due to missing MSU/rain gauge data.

nificance which indicates a more robust relation to SST boundary forcing. For example, in the case of all-India rainfall for the revised models, we are 99.5% confident that we may reject the null hypothesis that the boundary forcing does not affect the interannual variability of all-India rainfall, while for the original models we are only 95% confident that we may reject the null hypothesis. Thus, as in Sperber and Palmer (1996), the interannual variations of Nordeste rainfall are most readily simulated by the models. The Nordeste is one of the few regions for which long-lead forecasts of precipitation are made routinely (Graham 1993). The reproducibility of the Sahel rainfall has decreased, indicating a weaker signal-to-noise ratio, even though the agreement with observations has improved as per the Brier score. Simulation of interannual variations of Sahel rainfall still remains problematic, possibly due to land surface moisture feedback mechanisms (Rowell et al. 1995).

Figure 2 shows the Brier scores of JJAS precipitation for the revised models relative to MSU/raingauge data (Schemm et al. 1992, Spencer 1993). Relative to the skill scores in Tables 2-7, which are based upon indexes that are averaged over large domains, this is a more stringent test of the models ability to simulate interannual variability at gridpoint spatial scales. For India, the skill tends to be located along the west coast of the subcontinent. The dominant skill, which is associated with ENSO forcing, occurs over the central/eastern Pacific Ocean. Other areas of notable skill include the South Atlantic Convergence Zone, the southwestern Indian Ocean, and eastern and northern Australia. Moderate extratropical skill is found over central North America and western Europe. Thus, while the largest skill is located in the tropics, where the link between SST and rainfall is strongest, the models also exhibit pockets of skill in the extratropics.

5. Parameterization Sensitivity As noted in Table 1, four groups modified their convection schemes and four groups modified their land surface process parameterizations. While these are admittedly small sample sizes, this enables one to test the relative impact that modifying two different physical parameterizations may have on the simulation of interannual variability. However, this is not an ideal experiment since as noted earlier additional modifications were typically made to the revised models. Additionally, testing the effect of convective vs. land surface changes were not performed with the same subset of models. Thus, attribution of the improved simulation of interannual variability to either of these changed parameterizations is not conclusive. Therefore, the results may only be taken to be suggestive of the relative impact of modifying either of these schemes. However, Sperber and Palmer (1996) found that the verisimilitude of the simulation of all-India and Sahel rainfall was sensitive to the type of convective closure used in the complete suite of original AMIP integrations.

As seen in Table 4, relative to their original formulations, the revised models with modified convection schemes are in better agreement with observed interannual variability for all indexes as indicated by their lower Brier scores and improved statistical significance. The most pronounced improvement occurs to the rainfall indexes, although the monsoon wind shear index also exhibits a moderate improvement. For the models with revised land surface schemes Table 5 indicates, perhaps surpris-

Table 4: Brier skill scores for the models that revised their convection schemes (BMRC, DNM, MPI, and NRL). Based on Monte Carlo experiments the significance levels of the skill scores are given in parentheses.

Model Set	India Rainfall [1984]	Monsoon Wind Shear [1982]	Sahel Rainfall [1985, 1986]	Nordeste Rainfall [1981, 1987]
Original	0.43 (.14)	0.44 (.119)	0.78 (.810)	0.42 (.146)
Revised	0.24 (.005)	0.38 (.059)*	0.64 (.566)	0.16 (.001)

*wind shear not available from DNM_A

Table 5: Brier skill scores for the models that modified their land surface parameterizations (BMRC, CNRM, DERF, and LMD). Based on Monte Carlo experiments the significance levels of the skill scores are given in parentheses.

Model Set	India Rainfall [1984]	Monsoon Wind Shear [1982]	Sahel Rainfall [1985, 1986]	Nordeste Rainfall [1981, 1987]
Original*	0.38 (.080)	0.41 (.071)	0.56 (.404)	0.02 (.000)
Revised*	0.40 (.114)	0.20 (.001)	0.58 (.428)	0.08 (.000)

*BMRC_B was developed subsequent to BMRC_A (here grouped in the original category)

ingly, that the monsoon wind shear exhibits the most improvement, while the simulated interannual rainfall variability degrades slightly. The improved ability to simulate the large scale circulation may have occurred through improving the land-sea temperature contrast, which is important for establishing the monsoon flow. However, a more in-depth analysis is required to confirm this hypothesis.

As seen in Table 6, relative to their original formulations, the revised models with modified convection schemes show improved reproducibility of all-India and Nordeste rainfall, the two indexes that showed the most substantial improvement with respect to observed variability (Table 4 Brier scores). Table 7 indicates that models with revised land surface parameterizations have enhanced reproducibility of the monsoon wind shear and all-India rainfall. Thus, where the improvement in the Brier scores is largest, there tends to be a commensurate improvement in the signal-to-noise ratio as indicated by the reproducibility.

Modification of the convection is associated with a more realistic representation of observed variability over the regions considered as indicated by the Brier scores in

Table 6: Reproducibility skill scores for the models that revised their convection schemes. The significance levels at which we may reject the null hypothesis that the sea surface temperature has no effect on the interannual variations of the indexes are given in parentheses.

Model Set	India Rainfall	Monsoon Wind Shear	Sahel Rainfall	Nordeste Rainfall
Original	0.43 (>.05)	1.99 (.005)	0.49 (>.05)	0.53 (>.05)
Revised	0.57 (.05)	1.13 (.025)*	0.24 (>.05)	1.49 (.005)

*wind shear not available from DNM_A

Table 7: Reproducibility skill scores for the models that modified their land surface parameterizations. The significance levels at which we may reject the null hypothesis that the sea surface temperature has no effect on the interannual variations of the indexes are given in parentheses.

Model Set	India Rainfall	Monsoon Wind Shear	Sahel Rainfall	Nordeste Rainfall
Original*	0.26 (>.05)	1.09 (.005)	0.39 (>.05)	4.09 (.005)
Revised*	0.85 (.01)	1.28 (.005)	0.24 (>.05)	1.96 (.005)

*BMRC_B was developed subsequent to BMRC_A (here grouped in the origianl category)

Table 4, while the land surface modification was only associated with a more realistic large-scale flow over the summer monsoon region (Table 5). This is not to say that modifying the land surface scheme is not beneficial for simulating the rainfall variability. Sperber and Palmer (1996) employed a teleconnection criterion which enabled the further stratification of model performance. Models that (qualitatively) simulated the observed teleconnection patterns of the indexes with SST had improved Brier and reproducibility skill scores relative to those models that did not simulate the observed teleconnection patterns. The original version of CNRM did not meet the teleconnection criterion for all-India rainfall. However, the revised CNRM model, which included a substantial modification to the land surface scheme (Table 1) simulated the observed teleconnection pattern with SST.

6. Mean State vs. Interannual Variability Sperber and Palmer (1996) found that the quality of the mean state was related to the ability of the models to simulate interannual variability. Thus, they were able to relate systematic error of the mean

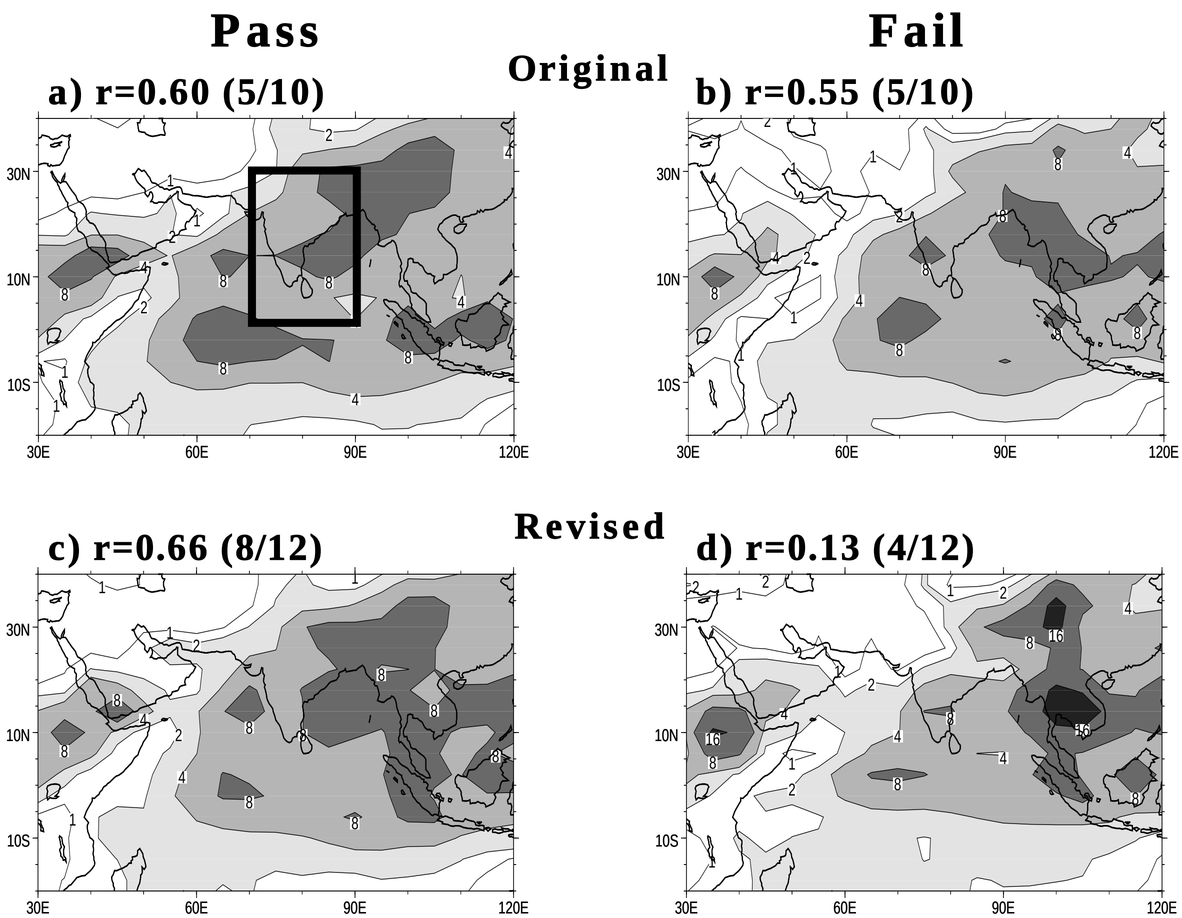

Figure 3: Simulated JJAS rainfall climatologies for the original models that (a) did, (b) did not, and the revised models that (c) did and (d) did not simulate the observed all-India rainfall SST teleconnection pattern. Above each panel are given the pattern correlation coefficients with MSU/rain gauge data for the region 2-30^{o}N, 70-90^{o}E [box in panel (a)] and the fraction of models that simulated or did not simulate the observed teleconnection pattern. The unit of precipitation rate is mm/day.

state to biases in the simulation of interannual variability. As described above, a teleconnection criterion was employed to assess the ability of each model to simulate the observed teleconnection patterns. For all-India rainfall, correlation with SST results in an ENSO pattern with a region of anticorrelation in the tropical central/eastern Pacific SST and positive correlations in the western Pacific that extend into the extratropical Pacific of each hemisphere (e.g., Sperber and Palmer 1996). The models were then grouped into two sets, those that were able, and those that were unable to simulate the observed teleconnection (qualitatively). Then, the ensemble mean rainfall climatology of each group of models was compared with observations using the pattern correlation technique. For the 32 original AMIP models, the ensemble of models that were able to simulate the observed all-India rainfall/SST teleconnection had a substantially higher pattern correlation with observed rainfall in the vicinity of India relative to the ensemble of models that did not simulate the observed rainfall/SST teleconnection.

The same analysis is performed for the revised models in Table 1 in comparison to their original versions. For the original models that simulated the observed teleconnection (Fig. 3a), the pattern correlation with MSU/raingauge data (Schemm et al. 1992, Spencer 1993) in the vicinity of India is slightly higher (0.60 vs. 0.55) relative to the original models that did not to simulate the observed teleconnection (Fig. 3b). Thus, this subset of original models weakly supports the Sperber and Palmer (1996) indication of a direct link between the quality of the mean state and successful simulation of observed interannual variability (which was based on the analysis of 32 AMIP models). However, the revised models firmly support this hypothesis, since as seen in Fig. 3c the revised models that simulated the observed rainfall/SST teleconnection have a pattern correlation of 0.66 with observations, while those that did not simulate the observed teleconnection have a pattern correlation of only 0.13 (Fig. 3d). Additionally, 8/12 (67%) of the revised models passed the teleconnection criterion as compared to 5/10 (50%) of their original counterparts. This indicates that a larger majority of revised models show realistic interannual variability, and that this facet is strongly associated with the quality of the mean state.

7. Conclusions In this paper, revised AMIP models are compared against their original simulations in order to assess the impact of modifying model formulation. The revised models exhibit better agreement with observations as indicated by the Brier scores and their improved statistical significance in Table 2. Table 3 indicates that the signal-to-noise ratio also improves for the revised models (with the exception Sahel rainfall). Revision of convective parametrizations had a more beneficial impact on rainfall variability than did modifications to land surface schemes. The results regarding the model sensitivity to convection and land surface processes (Tables 4-7) are only suggestive given that other changes were also incorporated in the revised models. Systematic intercomparison of different physical parameterizations would be facilitated by plug-compatibility of physics modules among models. The revised models firmly support the Sperber and Palmer (1996) finding that systematic error in the mean state is linked to the quality of the interannual variability. The results further reinforce the beneficial nature of standardized experimentation, with the original AMIP integrations providing an excellent benchmark against which model development and improvement can be assessed.

These results are however tempered by the fact that only one realization from each model was available for analysis under the AMIP I experimental design. Many authors have shown that multiple realizations are necessary to firmly assess the robustness of a given models response to SST forcing (e.g. Brankovic et al. 1994, Sperber and Palmer 1996). However, the approach here has not been to assess the skill of individual models, but rather it is over the ensemble of different realizations that skill is assessed. To the extent that the models are basically realistic in their representation of the physics and dynamics of the atmosphere, they serve a proxy for multiple realizations of nature. The next phase of AMIP experimentation calls for the submission of multiple realizations (differing only in the specification of initial conditions) from the participating modelling groups in order to address the robustness of individual models (Gleckler 1996). With ensembles a more direct assessment of the significance of the skill scores will be possible.

Acknowledgments I gratefully acknowledge the AMIP modelling groups for making available the data from their integrations. Special thanks to Dr. Tim Palmer for inspiring many of the ideas discussed in this paper, to Dr. Julia M. Slingo for helpful comments on an earlier draft of this manuscript, to Prof. Stefan Hastenrath for providing the observed Nordeste rainfall index, and to Drs. David Rowell, Neil Ward and Mike Hulme for providing the observed Sahel rainfall index. This work was performed under the auspices of the U.S. Dept. of Energy Environmental Sciences Division at the Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

References

Aceituno, P., 1988: On the functioning of the Southern Oscillation in the South American sector: Part 1: Surface climate. *Mon. Wea. Rev.*, **116**, 505-524.

Brankovic, C., T. N. Palmer and L. Ferranti, 1994: Predictability of seasonal variations. *J. Clim.*, **7**, 217-237.

Folland, C. K., J. Owen, M. N. Ward and A. Coleman, 1991: Prediction of seasonal rainfall in the Sahel region using empirical and dynamical methods. *J. Forecast.*, **10**, 21-56.

Folland, C. K., T. N. Palmer and D. E. Parker, 1986: Sahel rainfall and worldwide sea temperatures, 1901-1985. *Nature*, **320**, 602-607.

Gates, W. L., 1992: The atmospheric model intercomparison project. *Bull. Amer. Meteorol. Soc.*, **73**, 1962-1970.

Gadgil, S., and S. Sajani, 1998: Monsoon precipitation in the AMIP runs. *Clim. Dynam.*, **14**, 659-689.

Gates, W.L., J. Boyle, C. Covey, C. Doutriaux, M. Fiorino, P. Gleckler, J. Hnilo, S. Marlais, T. Phillips, G. Potter, B. Santer, K. R. Sperber, K. Taylor, and D. Williams, 1998: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP). Bull. Amer. Meteorol. Soc. (submitted).

Gleckler, P. J., (ed.) 1996: AMIP Newsletter No. 8. 20pp. PCMDI/LLNL P.O. Box 808, L-264, Livermore, CA 94551 USA.

Graham, N. E., 1993: Experimental predictions of wet season precipitation in northeastern Brazil. in *Proceedings of the Eighteenth Annual Climate Diagnostics Workshop*, U.S. Department of Commerce, NOAA/NWS/CAC/NMC, Springfield, VA, 378-381.

Hastenrath, S., 1988: Climate and Circulation in the Tropics. 2nd ed., D. Reidel, Boston, Massachusetts. 455pp.

Hastenrath, S., and L. Heller, 1977: Dynamics of climate hazards in Northeast Brazil. *Quart. J. Roy. Meteorol. Soc.*, **103**, 77-92.

Lamb, P. J., 1978: Large-scale tropical Atlantic circulation patterns associated with sub-Saharan Africa weather anomalies. *Tellus*, **30**, 240-251.

Lough, J. M., 1986: Tropical Atlantic sea-surface temperatures and rainfall variations in Subsaharan Africa. *Mon. Wea. Rev.*, **114**, 561-570.

Moura, A. D., and J. Shukla, 1981: On the dynamics of droughts in Northeast Brazil: Observations, theory and numerical experiments with a general circulation model. *J. Atmos. Sci.*, **38**, 2653-2675.

Murphy, A. H., and R. W. Katz, 1985: Probability, statistics and decision making in the atmospheric sciences. Westview Press, 379-437.

Nicholson, S. E., and D. Entekhabi, 1986: The quasi-periodic behavior of rainfall variability in Africa and its relationship to Southern Oscillation. *Arch. Met. Geoph. Biocl.*, **A34**, 311-348.

Palmer, T. N., 1986: Influence of the Atlantic, Pacific and Indian Oceans on Sahel Rainfall. *Nature*, **322**, 251-253.

Parthasarathy, B., Kumar, K. R., and Kothawale, D. R., 1992: Indian summer monsoon rainfall indices: 1871-1990. *Meteorol. Mag.*, **121**, 174-186.

Perrone, T. J., and R. G. Miller, 1985: Generalized exponential Markov and model output statistics: a comparative verification. *Mon. Wea. Rev.*, **113**, 1524-1541.

Phillips, T. J., 1994: A summary documentation of the AMIP models. PCMDI Report No. 18, UCRL-ID-116384. Lawrence Livermore National Laboratory, P.O. Box 808, L-264, CA 94551.

Ropelewski, C. F., and Halpert, M. S., 1987: Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Wea. Rev.*, **115**, 1606-1626.

Rowell, D. P., C. K. Folland, K. Maskell, and M. N. Ward, 1995: Variability of summer rainfall over tropical North Africa (1906-92): Observations and modelling. *Q. J. R. Meteorol. Soc.*, **121**, 669-704.

Schemm, J.-K., S. Schubert, J. Terry, S. Bloom, 1992: Estimates of monthly mean soil moisture for 1979-1989. NASA Technical Memo 104571, Goddard Space Flight Center, Greenbelt, MD, 260 pp.

Slingo, J. M., M. Blackburn, A. Betts, R. Brugge, B. J. Hoskins, M. J. Miller, L. Steenman-Clark, and J. Thuburn, 1994: Mean climate and transience in the tropics of the UGAMP GCM: sensitivity to convective parametrization. *Q. J. R. Meteorol. Soc.*, **120**, 881-922.

Spencer, R. W., 1993: Global oceanic precipitation from the MSU during 1979-91 and comparison to other climatologies. *J. Clim.*, **6**, 1301-1326.

Sperber, K. R., and S. Hameed, 1993: Phase locking of Nordeste precipitation with sea surface temperatures. *Geophys. Res. Lett.*, **20**, 113-116.

Sperber, K. R., and T. N. Palmer 1996: Interannual tropical rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project. *J. Clim.*, 9, 2727-2750.

Walker, G. T., 1928: World weather, III. *Mem. Roy. Meteorol. Soc.*, **2**, 97-134.

Walker, G. T., and E. W. Bliss, 1932: World weather, V. *Mem. Roy. Meteorol. Soc.*, **4**, 53-84.

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

Ward, M. N., S. Brooks and C. K. Folland, 1988: of seasonal rainfall in the northern Nordeste region of Brazil. Chap. 21 in Recent Climatic Change: A regional approach. S. Gregory (ed.), Belhaven Press, NY, 237-251.

WCRP, 1992: Simulation of interannual and intraseasonal monsoon variability. WMO/TD-No. 470, 185pp.

WCRP, 1993: Simulation and prediction of monsoons: Recent results. WMO/TD-No. 546, Geneva, Switzerland, 73pp.

Webster, P. J. and Yang, S., 1992: Monsoon and ENSO: selectively interactive systems. *Quart. J. Roy. Meteorol. Soc.*, **118**, 877-926.