Vol. 51: 201–216, 2012
doi: 10.3354/cr01064

CLIMATE RESEARCH
Clim Res

Published online March 20

OPEN ACCESS

Improvements over three generations of climate model simulations for eastern India

L. Das[1,2,*], J. D. Annan[1], J. C. Hargreaves[1], S. Emori[3]

[1]RIGC / JAMSTEC, 3173-25 Showamachi, Kanazawa-ku, Yokohama City, Kanagawa 236-0001, Japan
[2]Department of Agricultural Meteorology and Physics, BCKV, Mohanpur, Nadia 741-252, India
[3]National Institutes for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan

ABSTRACT: In the present study we investigate the performance of climate models which contributed to the past 3 Intergovernmental Panel for Climate Change (IPCC) assessment reports for the Gangetic West Bengal region of east India (6° × 6°). Analysing present-day seasonal rainfall and temperature over the domain, we compare the results of the models (from the 6 modelling centres common to the second, third and fourth assessment reports—SAR, TAR and AR4, respectively) in order to judge to what extent these global models have improved on a regional scale. Metrics for model evaluation are not yet firmly established in the literature, so in this paper we compare and contrast the results from a number of different statistics used in previous studies. We also analyse the impact of topography on the results obtained for the AR4 models. We find that most models improved from SAR to AR4, although there is some variation in this result depending on seasons, variables and on which statistical methods are used in the analysis. The multi-model mean of the 6 models improves from SAR to TAR to AR4. The overall best performance in this region in the AR4 is the Japanese model, MIROC, but the best model in terms of improving skill from SAR to AR4 is the GFDL model from the United States. Correcting for errors in the model topographies produced an overall improvement of spatial patterns and error statistics, and greatly improves the performance of 1 model (CGCM) which has poor topography, but does not affect the ranking of the other models.

KEY WORDS: Seasonal cycle · Model improvement · Similarity statistics · Sub-regional scale · Topography · Model improvement index · Rank

1. INTRODUCTION

The Intergovernmental Panel for Climate Change (IPCC) published 4 consecutive assessment reports in the years 1990, 1995, 2001 and 2007 known as the first, second, third and fourth assessment reports (FAR, SAR, TAR and AR4, respectively). These reports provide huge amounts of model data for impact researchers, policy makers and other stakeholders involved in future planning and mitigation in the context of global warming. Some modelling centres have continuously provided simulations in the IPCC archive starting from SAR, whereas other modelling centres have emerged more recently to enhance and support the activities of the IPCC. Therefore, the number of models has increased from 7 in SAR to 24 in AR4, respectively. The modelling centres typically do not change the model's fundamental conceptual basis from one generation to next; rather they modify the model's code through both improved understanding of model physics and forcing, and also improved computational resources. Although climate models are sophisticated tools for predicting climate change and global warming, they still attract criticism due to their imperfections. Therefore, testing the realism of several generations of coupled climate models over a wide spatial domain is appealing. Several papers have investigated the

*Email: daslalu@yahoo.co.in

performance of climate models by measuring the degree of correspondence between observed and simulated fields using a range of conventional statistics (Willmott 1981, ASCE 1993, Boer & Lambert 2001, Taylor 2001, Moriasi et al. 2007, Gleckler et al. 2008, Reichler & Kim 2008) as well as several graphic techniques which provide a visual comparison of simulated and observed results (ASCE 1993, Legates & McCabe 1999). One popular method for model evaluation is the Taylor diagram (Taylor 2001), a 2-dimensional diagram based on the correlation coefficient, bias-corrected root mean square error (RMSE) and variance ratio, which conveys information between models and observations. Boer & Lambert (2001) also introduced a similar method, but these methods are most practical for visualising results from small numbers of models and/or small numbers of variables. Recently, model evaluation through the development of a single overall performance index such as the 'climate prediction index' (CPI; Murphy et al. 2004) is intuitively appealing, as it measures the reliability of a model based on the composite errors of a wide range of climate variables. On the other hand, performance metrics have been developed by Glecker et al. (2008), which indicate that the use of a single measure may not be appropriate, since it will conceal the multivariate, seasonal and inter-annual variations between models, and there is little agreement on the most suitable metric to use. Another method for assessing model performance is based on probabilistic approaches (e.g. Dessai et al. 2005, Alexander et al. 2006, Maxino et al. 2007), while a study by Min & Hense (2006) measured the skill of a Bayesian probabilistic approach in terms of a likelihood ratio between the model simulations. It is not clear whether conventional statistics or the probabilistic approach is superior for model evaluation.

The majority of studies evaluating model performance so far have considered a larger region (continental size) or the whole globe and have analysed only one generation of model output, although Reichler & Kim (2008) recently analysed all 3 different climate model inter-comparison projects (CMIP): CMIP-1(Meehl et al. 2000), CMIP-2 (Covey et al. 2003, Meehl et al. 2005) and CMIP-3 (PCMDI 2007, which is the IPCC-AR4 database). A few studies have also evaluated model performance on smaller scales. For example, Hulme et al. (1993) used a grid-box centred at 52.5° N and 2.5° W and compared the model simulation with observed temperature and precipitation over England and Wales, while Dessai & Hulme (2008) compared 4 generations of UK climate scenarios with observations over Central England using 5° × 5° spatial resolution in the earlier generation (CCIRG91). There is a growing challenge for the modelling communities to provide reliable simulations for impact studies and policy-oriented research. While modelling centres are working hard to adequately represent the real complexities of the earth-climate system in model simulations, it is important to continue testing against observations of present climate to show to what extent the coupled models are improving over time. Studies by the IPCC (2007) and Reichler & Kim (2008) have revealed that models have gradually been improving in accuracy, from SAR to AR4, over global and regional scales. Neither of those studies, however, has assessed model improvement over a smaller sub-regional scale, where the size of the region is comparable with the effective resolution of most of the considered models. Further, those studies validated models using gridded global observational data sets from satellites and reanalyses from numerical weather prediction models, rather than actual station-based *in situ* observational data. Model validation over such small areas and testing model improvement across generations will be increasingly important for impact researchers and policymakers. Impact assessment in such small areas requires high-resolution climate data, but output from regional climate model (RCM) simulations may not always be available, especially for those with limited computational resources. On the other hand, the IPCC provides huge climate model outputs from SAR, TAR and AR4 that are freely available for research purposes. Therefore, our aim is to validate the 3 generations of coupled climate models (those used for the 1995, 2001 and 2007 reports of the IPCC) using a wide range of conventional statistics in order to show to what extent these models are improving over time on a smaller sub-regional scale—the Gangetic West Bengal and its neighbourhood (GWBN), in eastern India. The GWBN is part of tropical monsoon region, and its mean climate is strongly influenced by the large-scale circulation patterns of the south-west and north-east monsoonal winds. The moist end of the monsoon trough (the monsoon trough is a portion of the inter-tropical convergence zone) passes through this region, resulting in heavy rainfall and cyclonic activities in the nearby sea, the Bay of Bengal, and causing heavy showers in the post-monsoon season. The region is of interest both because of the range of climatological features represented, and also because it comprises some of the most agriculturally productive land in India. We did not investigate

the reasons for changes in performance, which may include computational power (grid resolution), improved forcings, and/or better representation of physical processes, but have restricted our focus, similar to Reichler & Kim (2008), to evaluating whether improvement exists at a sub-regional scale and the quantification of this.

We analysed precipitation and temperature fields and compared model output to real observations collected over 30 consecutive years in the late 20th century. The statistics evaluated for all 3 model generations are: the spatial correlation coefficient, mean bias, the agreement index (*d*-index) and some error statistics, namely mean absolute error (MAE), total RMSE, centred RMSE and normalised error variance (NEV). Then we developed model improvement indices (MII) for each of the performance statistics to calculate the improvement of the models from SAR to TAR and from TAR to AR4. The overall improved skill of model performance across all generations and statistics is evaluated by overall model improvement indices (OMII) for the different seasons. An attempt has been made to assess the effect of model and station topography on model performance on a small scale in the present study.

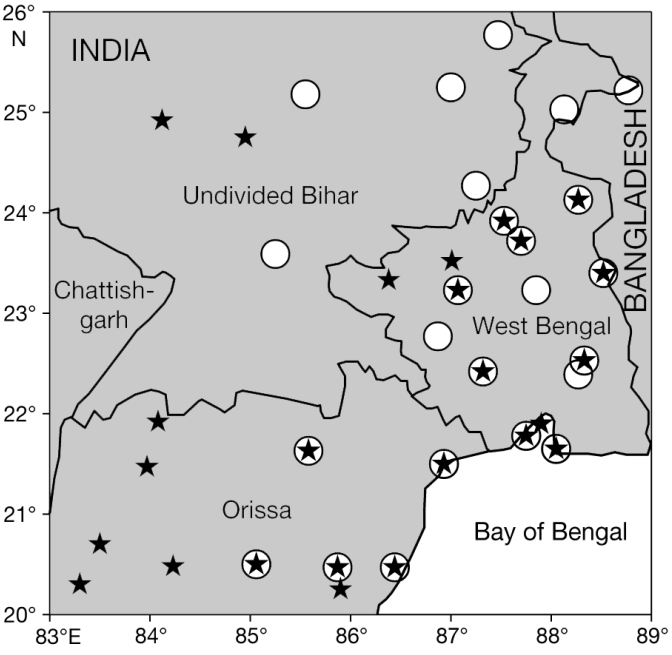

Fig. 1. Location of the study region showing the distribution of observational stations. Stations—(★) June-July-August-September (JJAS), October-November (ON) and December-January-February (DJF), (○) March-April-May (MAM); (⊛) all seasons

2. STUDY AREA, DATA AND METHODS

2.1. Study area

The GWBN region, extending from 20 to 26° N and 83 to 89°E in eastern India (Fig. 1), is one of the agriculturally most productive regions in India and is well known as the 'rice and vegetable belt of India'. The region of interest is predominantly a plain, and the Bay of Bengal in the east strongly influences the climate of this region. It also has 4 distinct seasons, namely winter (December to February; DJF), pre-monsoon (March to May; MAM), monsoon (June to September; JJAS) and post-monsoon (October and November; ON). The region also includes a number of large cities, including Kolkata, Patna, Ranchi and Bhubaneswar. It is the most densely populated area of India at 904 people km^{-2}. The present trend of expanding industrialisation and contraction of agricultural land, unused land and water bodies is a key concern in terms of ecosystem balance and environmental protection. The climate of this region has changed over the last century (Table 1), with large increases of rainfall in ON (82%; 152.7–279.3 mm) and DJF (58%; 33.7–53.4 mm) between the periods 1901–1930 and 1961–1990. It is also expected that this trend will continue in the next century due to anthropogenic influence (e.g. Das & Lohar 2005). This domain, with a small variation in topography, receives 1400–1500 mm total annual rainfall over the 4 distinct seasons. The sources and causes of rainfall in the different seasons vary. In winter we have the 'western disturbance'; in pre-monsoon early summer, squalls and thunderstorms locally known as 'kalbaisakhi' or 'Nor'westers'; in the monsoon season a large-scale south-west monsoonal wind circulation pervades; and in the post-monsoon season occasional torrential rainfall is due to cyclonic activity over the nearby area of the Bay of Bengal. The uncertainty ranges for the observed data of 25 stations during the period 1961–1990 are expressed through their standard error (SE) and the interannual variability of the same data is also calculated through the coefficient of variation (CV). The variation in the SE and CV values (%) for rainfall and temperature from a maximum to a minimum in different seasons is summarized in Table 2. Some studies indicate that thunderstorm activities are changing in the pre-monsoon season, and some local winds in the south Bengal region have altered their direction due to rapid increase of summer paddy cultivation (Lohar & Pal 1995). Evaluation of model performance over a small domain, the climate of which is dominated by so many external

Table 1. Total rainfall and mean temperature changes (with respect to 1901–1930) during the 20th century over the Gangetic West Bengal and its neighbourhood (GWBN) according to season (see Section 2.1) in 2 different time periods

Time series	ΔRainfall (%)				ΔTemperature (°C)		
	MAM	JJAS	ON	DJF	JJAS	ON	DJF
1931–1960	4	4	22	–6	0.3	0.3	0.4
1961–1990	3	5	82	58	1.2	0.6	0.7

Table 2. Seasonal ranges of standard error (SE) and coefficient of variation (CV) for observed rainfall and temperature over the GWBN

Season	Rainfall		Temperature	
	SE (mm)	CV (%)	SE (°C)	CV (%)
MAM	8.4–19.6	29.1–69.3	0.1–0.5	2.1–9.4
JJAS	32.4–79.1	15.4–36.1	0.1–0.2	0.8–4.9
ON	8.8–23.2	41.3–99.4	0.1–0.3	1.7–7.9
DJF	3.7–11.3	59.8–131.6	0.1–0.3	1.9–9.7

factors, is rather a severe test of the models. It is important, however, to understand to what extent the models can perform effectively on a local scale at which local adaptation to climate change must take place. Such studies could also lead to improved understanding of sub-grid scale processes which may aid future model development. Furthermore, this region, while influenced by a range of climate regimes, is quite large and topographically simple such that we expect that its climate is reasonably represented in all but the lowest resolution models considered here.

2.2. Data

2.2.1. Observations

Monthly total rainfall and monthly mean temperature from the India Meteorological Department (IMD) and the Department of Agriculture, Govt. of West Bengal for the period 1961–1990 were used. Some of the data are incomplete and only cover limited seasons. Therefore, after screening the data records (taking into account only the stations having ≥80% complete records), we finally selected 35 stations which are well distributed over the GWBN (Fig. 1). For the purpose of analysis we used 25 stations (shown as black stars) for the seasons JJAS, ON and DJF and 24 stations (shown as open circles) for MAM. Stations indicated by stars inside circles were used in all seasons. For more details on the locations of the observational stations, see our previous work which used the same data set (Das & Lohar 2005).

2.2.2. GCM data

The present study includes model outputs available from the IPCC, from the SAR (1995), TAR (2001) and the latest assessment report AR4 (2007).

The SAR data sets were archived as IPCC_DDC_SAR and obtained from http://cera-www.dkrz.de. We used the sulphate aerosol and greenhouse gas integration experiment (i.e. XGS01, where X is name of the model) for all 6 GCMs in this analysis. Each trial of the model involved an effective greenhouse forcing corresponding to that observed from 1850 to present. The direct effect of sulphate aerosols was also included by increasing the surface albedo. The model details are presented in Table 3.

The model simulations for the TAR were mainly based on the IPCC special report on emission scenarios (IPCC-SRES) that was published by the IPCC in 2000; this was obtained from the same website (http://cera-www.dkrz.de) as IPCC_DDC_TAR. For the present study we used a 30 yr 'normal' period for modern climate (1961–1990). The details of each model experiment are available from www.ipcc-data.org/sres/gcm_data.html.

The AR4 data sets were obtained from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) archive (www-pcmdi.llnl.gov/ipcc/about_ipcc.php). The 'climate of the 20th century simulations (20C3M)', hereafter referred to simply as the 'present-day' experiments for the period 1961–1990 were used in this analysis. The AR4 simulations were driven by a rather more realistic set of external forcing factors, including various estimated historical natural and anthropogenic forces, such as solar variation, volcanic eruptions, ozone variation, halocarbons, land use and sulphate aerosols. The modelling groups were free to use their own best estimates of these forcings (Kunkel et al. 2006). The exact formulation of these forcings was not identical among models, and the starting and ending dates of the simulations varied from model to model. For the present day experiments, multiple runs with the same forcing, but with different initial conditions were performed for the same individual model. In the present study, we only used Run 1 for all models. The detailed AR4 model information is presented in Table 3.

Table 3. Description of models for the IPCC second assessment report (SAR), third assessment report (TAR) and fourth assessment report (AR4). Grid resolution: latitude × longitude; L: number of vertical layers; CCMA: Canadian Centre for Climate Modeling and Analysis; CSIRO: Commonwealth Scientific and Industrial Research Organization; MPIfM: Max-Planck Institut für Meteorologie; GFDL: Geophysical Fluid Dynamics Laboratory; HCCPR: Hadley Centre for Climate Prediction and Research; CCSR/NIES: Centre for Climate Research Studies / National Institute for Environmental Studies

Study ID	Centre and location	Model generation	IPCC ID	Atmospheric resolution	Oceanic resolution	Source
CGCM	CCCMA, Canada	SAR	CGCM1[a]	T32 (3.8° × 3.8°) L10	1.8° × 1.8° L29	Boer et al. (2000), Flato et al. (2000)
		TAR	CGCM2[a]	T32 (3.8° × 3.8°) L10	1.875° × 1.875° L29	Flato & Boer (2001)
		AR4	CGCM3.1[a]	T63 (2.8° × 2.8°) L31	1.4° × 0.94° L29	Flato & Boer (2001), McFarlane et al. (2005)
CSIRO	CSIRO, Australia	SAR	CSIRO-Mk2b[a]	R21 (3.2° × 5.6°) L9	3.2° × 5.6° L21	Hirst et al. (1996)
		TAR	CSIRO-Mk2b[a]	R21 (3.2° × 5.6°) L9	5.6° × 3.2° L21	Hirst et al. (2000)
		AR4	CSIRO-Mk3.0	T63 L18	1.875 × 0.84 L21	Gordon et al. (2002)
ECHAM	MPIfM, Germany	SAR	ECHAM4/OPYC3[a]	T42 (2.8° × 2.8°) L19	2.8° × 2.8° L11	Bacher et al. (1998)
		TAR	ECHAM4/OPYC3[a]	T42 (2.8° × 2.8°) L19	2.8° × 2.8° L11	Stendel et al. (2002)
		AR4	ECHAM5	T63 L32	1.0° × 1.0° L41	Min et al. (2005)
GFDL	GFDL, USA	SAR	GFDL-R15[a]	R15 (4.5° × 7.5°) L9	4.5° × 3.7° L12	Haywood et al. (1997)
		TAR	GFDL-R30[a]	R30 (2.25° × 3.75°) L14	2.25° × 1.875° L18	Delworth et al. (2002)
		AR4	GFDL-CM2.0	2.5° × 2° L24	1.0° × (.33-1.0)° L50	Delworth et al. (2006)
HADLEY	HCCPR, UK	SAR	HadCM2[a]	2.75° × 3.75° L19	2.5° × 3.75° L20	Johns et al. (1997)
		TAR	HadCM3	2.75° × 3.75° L19	1.25° × 1.25° L20	Johns et al. (2003)
		AR4	HadGEM1	1.875° × 1.25° L38	1.25° × 1.25° L20	Johns et al. (2004), Johns et al. (2006)
MIROC	CCSR/NIES, Japan	SAR	CCSR/NIES[a]	T21 (5.6° × 5.6°) L20	2.8° × 2.8° L17	Emori et al. (1999)
		TAR	CCSR/NIES	T21 (5.6° × 5.6°) L20	2.8° × 2.8° L17	Nozawa et al. (2001)
		AR4	MIROC3.2(hires)	T106 (1.125° × 1.125°) L56	1.25° × 1.25° L47	K-1 Model Developers (2004)

[a]Flux adjustment (heat and water) applied

Our main purpose was to explore and analyse the extent to which model development has resulted in improvements in model performance. Therefore, we have included only the 6 modelling groups which have provided data to the IPCC archive continuously from SAR to AR4. This avoids confounding issues due to the different numbers of models in each generation and the recent introduction of new models. The actual models used by each modelling centre sometimes change significantly between generations (Table 3). For clarity we used the same abbreviated name to refer to each generation from each modelling centre. For example, we used data from HADCM2 for SAR, HADCM3 for TAR and HADGEM1 for AR4, but we used the abbreviation HADLEY to refer to all generations. In essence, therefore, we are analysing the development of the modelling expertise of each centre rather than the evolution of particular models.

As computers have increased in computational capacity, models have been built at finer resolution. This is illustrated in Fig. 2, which shows the grid

points over the study domain for all the models considered in this work. It is notable that, particularly for the SAR, very few grid points are present over the domain for some models, but that, for the more recent generation, all models have a reasonable number of grid points over the domain. Thus, we expect to find little useful spatial information for earlier models, but expect to see this improved in later generations. One model in particular stands out, the Japanese model MIROC. Not only does it have the finest resolution in the AR4, but it also has increased in resolution dramatically between the TAR and AR4.

2.3. Methods

2.3.1. Reduction and analysis of model output and observational data

First the model performance was assessed through visual comparison of the simulated and observed annual cycles of rainfall and temperature for the 6 GCMs and the mean of the multi-model ensemble (MME6). Monthly data for all months (January–December) for the period 1961–1990 were obtained for all stations. Point output from the 6 GCMs was interpolated to the station locations (Fig. 1) by bilinear interpolation, using a minimum of 4 points from the domain and nearby areas (Fig. 2), similar to studies by Palutikof et al. (1997) and Das & Lohar (2005). This interpolation technique allows the model horizontal resolution in the evaluation process to assess the model improvement due to an increase of roughly averaged resolution from ~250 km (T42) in SAR to ~180 km (T63) in TAR to 110 km (T106) in AR4. Another popular practice of interpolation to a common grid both for station data and model outputs, similar to that by Taylor (2001), is applicable for a comparatively larger region (i.e. hemispheric or continental scale), but its smoothing effect would hinder the localised analysis intended here.

Monthly means of the data were calculated for all stations in the region for both rainfall and temperature. These interpolated data were averaged together (a simple average of all stations) to analyse the overall annual cycle. During interpolation to station locations some grids from nearby ocean locations of some models have been included. We also calculate the inter-annual variability (±1 SD of the regionally averaged time series) of observed rainfall and temperature to give an indication of how the model error compares to natural variability. For comparison of the accuracy of the annual cycle (e.g. Nieto & Rodríguez-Puebla 2006) we calculated the mean bias (MB), correlation coefficient (R) and normalised mean absolute error (NMAE) between the observations and simulated monthly time series of the annual cycles. See Table 4 for the definitions and formulae for these 3 statistics.

For the purpose of assessing the spatial distribution, we analysed the model output and observational data on a seasonal scale, averaging the monthly means at each station for each of the 4 seasons (DJF, MAM, JJAS and ON). We used 7 different statistics for this purpose, the definitions of which are given in Table 4: MB, R, *d*-index, normalised total root mean square error (NTRMSE), normalised centred root mean square error (NCRMSE), normalised mean absolute error (NMAE) and NEV.

2.3.2. Calculation of model improvement across generations

In the present paper we used 2 indices for evaluating the model improvement across model generations. The improvement in model performance between generations is calculated as the difference in the similarity statistics from one generation to the next, the MII. Since there are only 3 generations to analyse, the OMII is simply calculated as the difference between the first and last generation, which is equivalent to taking the linear trend across all 3 generations.

3. ASSESSMENT OF MONTHLY MEANS AND ANNUAL CYCLE

According to Legates & McCabe (1999), graphic techniques are essential for evaluating model performance. Thus, we first examined the output from the 3 generations of models visually (and therefore subjectively) to see how well they are able to simulate the observed seasonal cycles of rainfall and temperature over the GWBN. Figs. 3 & 4 show a comparison of the seasonal cycles for observed and simulated rainfall and temperature, respectively, for the different GCMs. The dashed lines show the 1 SD range of the climatology for the region and indicate greater inter-annual variation in precipitation than in temperature, when considered in the context of the seasonal range. The model averages are presented as histograms. A large range in model results occurs for precipitation, but there is a general bias towards underprediction of rainfall in almost all models (MIROC

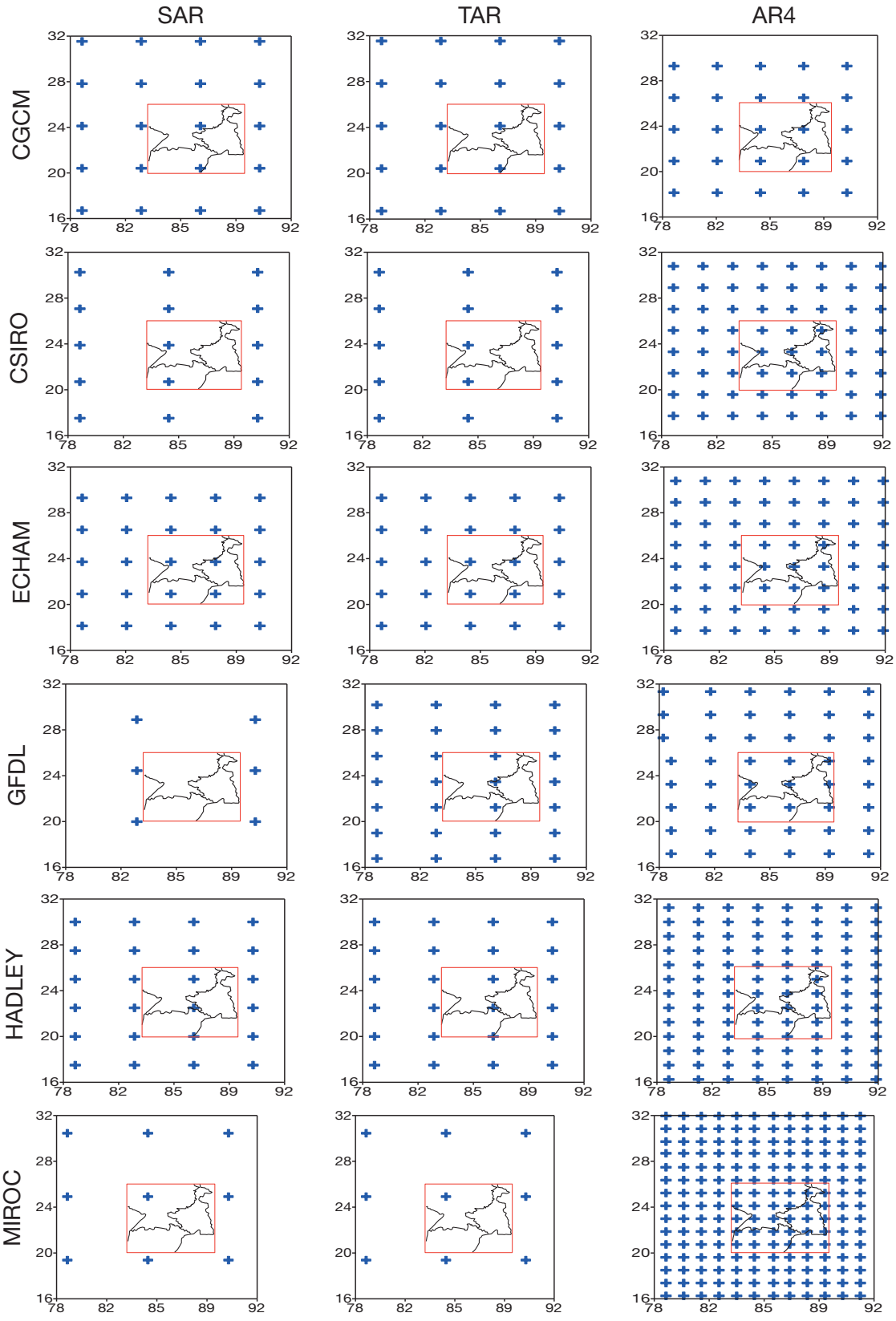

Fig. 2. Variation of effective horizontal resolution of global climate models (GCMs) in different IPCC assessment reports along with the number of grids in the study domain (red square; 20–26° N, 83–89° E) in eastern India. SAR, TAR, AR4: Assessment Reports (2nd, 3rd and 4th, respectively)

Table 4. Description of similarity statistics used in the present study. M: model output; M and σ_M: mean and standard deviation, respectively, of model output; O: observations; $\overline{O}$ and σ_O: mean and standard deviation, respectively, of observations; $\sigma_{O\text{int}}{}^2$: inter-annual variance of station data during the 1961–1990 time series; w_n: weights where n is the number of stations; RMSE: root mean square error; ABS: absolute value. For equal weight of all stations, $w_n = 1/25$

Name of similarity statistics	Equations	Reference/ studies that used this expression
Mean bias	$\mathrm{MB} = \left[\frac{1}{N}\sum_{n=1}^{N}(M_n - O_n)\right]$	Willmott (1982)
Correlation	$\mathrm{R} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(M_n - \overline{M}\right)\left(O_n - \overline{O}\right)}{\sigma_M \sigma_O}$	Taylor (2001)
Index of agreement (d-index)	$d - \mathrm{index} = 1.0 - \left(\frac{\sum_{n=1}^{N}(O_n - M_n)^2}{\sum_{n=1}^{N}\left(\lvert M_n - \overline{O}\rvert + \lvert O_n - \overline{O}\rvert\right)^2}\right)$	Willmott (1981), Legates & McCabe (1999)
Normalised total RMSE	$\mathrm{NTRMSE} = \frac{1}{\sigma_O}\left[\frac{1}{N}\sum_{n=1}^{N}(M_n - O_n)^2\right]^{1/2}$	Janssen & Heuberger (1995), Covey et al. (2002)
Normalised centred RMSE	$\mathrm{NCRMSE} = \frac{1}{\sigma_O}\left[\frac{1}{N}\sum_{n=1}^{N}\left\{\left(M_n - \overline{M}\right) - \left(O_n - \overline{O}\right)\right\}^2\right]^{1/2}$	Covey et al. (2002), Taylor (2001)
Normalised mean absolute error	$\mathrm{NMAE} = \frac{1}{\sigma_O}\left[\frac{1}{N}\left\{\sum_{n=1}^{N}\mathrm{ABS}(M_n - O_n)\right\}\right]$	Janssen & Heuberger (1995)
Normalised error variance	$\mathrm{NEV} = \sum_{n=1}^{N}\left[w_n (M_n - O_n)^2 / \sigma_{o\text{int}}{}^2\right]$	Reichler & Kim (2008)

being the clearest exception) for the months with heavier rainfall. Comparing the error to the size of the seasonal cycle gives the impression that precipitation is less well represented than temperature. For the SAR, temperature is generally underestimated, but by AR4 an approximately equal number of models under- and overestimate temperatures. Overall there seems to be an improvement in the results over the model generations. One notable exception is HADLEY where the best rainfall results were in the TAR, and both the SAR and AR4 are markedly underestimated, while ECHAM shows underestimated rainfall in the monsoon season (JJAS) and little evidence of improvement, although its temperatures remain well estimated. GFDL, MIROC and the multimodel mean (MME6) appear to produce the best rainfall results for AR4, while GFDL, MIROC, ECHAM and MME6 all seem good for AR4 temperature.

These visual impressions are confirmed for the most part by the statistics. Table 5 shows R, NMAE and the annual model MB between simulations and observations for rainfall and temperature. It is apparent that, as we found above, MIROC, GFDL and MME6 in AR4 simulate annual cycles of rainfall well, with a high correlation coefficient (0.99) and low NMAEs of 0.12, 0.19 and 0.21, respectively. In addition, all 3 statistics have generally improved across the generations, with some exceptions such as ECHAM and HADLEY. One curiosity is the decline in R for CSIRO for AR4, which may be caused by the rather late monsoon onset compared to SAR and TAR and the observations. The seasonal cycle in precipitation is very clear in the data, with low precipitation for 7 mo and high precipitation for 5 mo; this pattern is generally reproduced by the models leading to high values of R. For temperature, the changes between months are more gradual and the models follow the data pattern less closely, leading to slightly lower R values. As we identified visually, for temperature, GFDL, MIROC, ECHAM and MME6 have the lowest error (NMAE); however, CSIRO also has a low MB. The reason for this is apparent from Fig. 4; CSIRO overestimates summer and underestimates winter temperatures, giving a rather accurate annual mean.

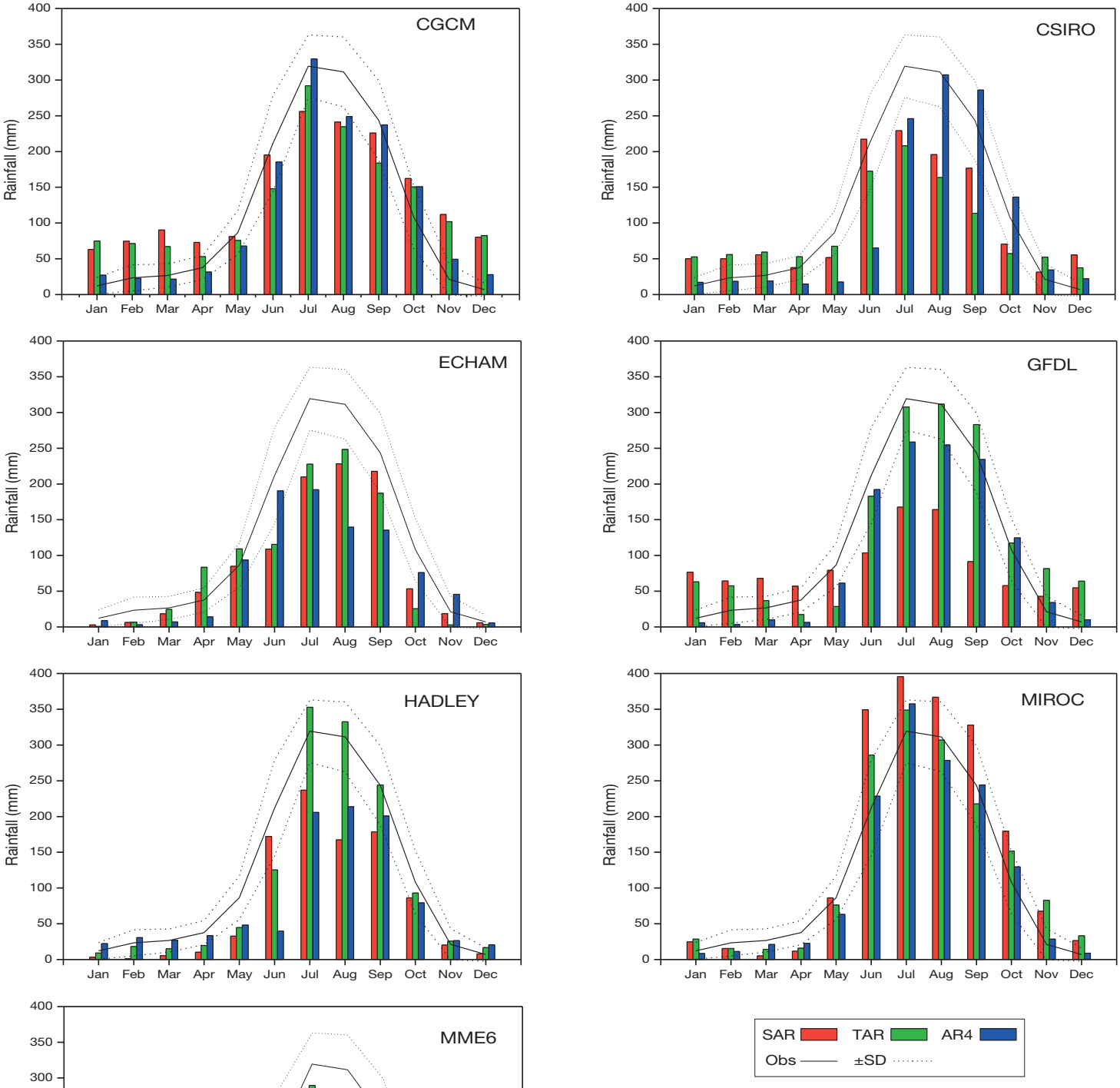

Fig. 3. Comparison of observed and simulated monthly mean rainfall from different GCMs over the GWBN (Gangetic West Bengal and its neighbourhood) for SAR, TAR and AR4 (2nd, 3rd and 4th Assessment Reports, respectively). Solid black line represents the observed cycle, and dotted lines above and below represent ±1 SD of observed climatology during the period 1961–1990

Fig. 4. Comparison of observed and simulated monthly mean temperatures from different GCMs over the GWBN (Gangetic West Bengal and its neighborhood) for SAR, TAR and AR4 (2nd, 3rd and 4th Assessment Reports, respectively). Solid black line represents the observed cycle, and dotted lines above and below represent ±1 SD of observed climatology during the period 1961–1990

4. ASSESSMENT OF SEASONAL MEANS AND SPATIAL PATTERNS

The comparison of the results of model performance are based on the 7 similarity measures discussed below. The scale used for each statistic changes across the 3 generations as the calculated values vary widely.

4.1. Mean bias

The MB in different seasons between observed and model values was calculated according to the equations described in Table 4 and illustrated in Figs. S1a–c & S2a–c, respectively (in the supplement at www.int-res.com/articles/suppl/c051p201_supp.pdf). Here we see different results for the different models. The poor JJAS results for precipitation are clear here, with large underestimates for all models except MIROC. For the other seasons, there is a mixture of positive and negative biases for rainfall, with general improvement for AR4 for ON and DJF. The JJAS results for temperature are, on the other hand, generally better than for the other seasons. On balance, over all seasons, there is an improvement in the bias for temperature across the generations.

4.2. Spatial correlation coefficient

Figs. S1d–f & S2d–f show R for rainfall and temperature for the 4 seasonal averages. The 1% significant correlation coefficient calculated using Student's *t*-test for N = 25 is 0.51, which, although not directly relevant to this analysis, provides a benchmark for the reliability of the results. The correlations are generally positive, although not all significantly, in AR4 for all seasons apart from precipitation in DJF, indicating that the spatial distribution has some similarity to the observations. The results are less consistent for the SAR and TAR. The JJAS results for rainfall and temperature shown improvement across generations for the most part and positive correlation for the AR4. Over the 4 seasons, however, the value of R is more reliably positive for temperature than for precipitation, with the DJF and ON seasons producing positive R values for all generations, while the ON season produces the best R values for precipitation in AR4.

Table 5. Comparison of the correlation coefficient (R), normalised mean absolute error (NMAE) and mean bias (MB) between observed and model-simulated seasonal cycles for 3 model generations. Other abbreviations see Table 3

Study ID	Model generation	Rainfall R	Rainfall NMAE	Rainfall MB (mm d^{-1})	Temperature R	Temperature NMAE	Temperature MB (°C)
CGCM	SAR	0.96	0.41	0.7	0.89	0.48	−1.9
	TAR	0.94	0.42	0.3	0.73	0.78	−3.0
	AR4	0.98	0.17	−0.1	0.96	0.73	−2.9
CSIRO	SAR	0.95	0.35	−0.5	0.91	0.50	−2.8
	TAR	0.93	0.47	−0.9	0.93	0.40	−2.2
	AR4	0.90	0.31	−0.6	0.96	0.22	−0.4
ECHAM	SAR	0.97	0.30	−1.1	0.97	0.23	0.7
	TAR	0.94	0.35	−1.0	0.93	0.36	1.1
	AR4	0.92	0.39	−1.4	0.98	0.23	0.9
GFDL	SAR	0.90	0.59	−1.1	0.87	1.09	−4.3
	TAR	0.95	0.26	0.4	0.96	0.31	0.7
	AR4	0.99	0.19	−0.6	0.95	0.32	−0.6
HADLEY	SAR	0.97	0.34	−1.3	0.93	0.37	1.1
	TAR	0.97	0.17	−0.3	0.97	0.31	0.6
	AR4	0.91	0.37	−1.3	0.96	0.82	−2.0
MIROC	SAR	0.97	0.39	1.2	0.96	1.03	−4.1
	TAR	0.97	0.23	0.5	0.98	1.45	−5.8
	AR4	0.99	0.12	−0.1	0.99	0.21	−0.1
MME6	SAR	0.99	0.24	−0.3	0.97	0.47	1.2
	TAR	0.99	0.21	−0.2	0.98	0.37	−1.4
	AR4	0.99	0.21	−0.6	0.99	0.32	−0.8

4.3. Agreement index

The formulation of the *d*-index is shown in Table 4, and the results are shown for rainfall in Fig. S1g–i and temperature in Fig. S2g–i (note differences in y-axis scales). Higher values of the *d*-index indicate better model performance. Random data having merely the same mean and variance as the observations achieve a *d*-index of about 0.4, so this value may be considered a minimum guideline for assessing whether the models may have any skill in reproducing the data. In general, improvement occurs across the generations from SAR to AR4, with more models having a *d*-index > 0.4 for all seasons. The precipitation *d*-index is consistently best for ON and worst for DJF, while for temperature DJF is best and JJAS worst.

4.4. Normalised error indices (NTRMSE, NCRMSE, NMAE and NEV)

We report here the results from the normalised error indices, NTRMSE, NCRMSE, NMAE and NEV, described in Table 4. The results obtained for each model for all seasons and generations are plotted in Fig. S1j–u and Fig. S2j–u for rainfall and temperature, respectively. The results vary across generations, seasons and variables. NMAE is less sensitive to outliers than is NTRMSE. In practice, however,

the 2 statistics produce only slight differences in terms of model ordering. The performance of AR4 models is presented in Fig. S1l,o,r,u for rainfall and Fig. S2l,o,r,u for temperature. In the monsoon season (JJAS), MIROC performs best for rainfall for all error estimators apart from NCRMSE, where it is third best, and for temperature it is the most consistent good performer. On the basis of all error indices for rainfall (Fig. S1j–u) and temperature (Fig. S2j–u), the magnitude of errors generally decreases from SAR to AR4, although this is not uniform for all models.

It may seem somewhat surprising that the NTRMSE is so high in some models for rainfall in the typically cool and quiet winter season (DJF) in SAR and TAR over the GWBN region, whereas for temperature it is low. The rainfall is, however, relatively low at this time, resulting in a large normalising constant. The values of NEV for all models and generations are also higher for temperature (Fig. S2s–u) than precipitation. These high values arise mainly due to dividing by the low values of observed variances (inter-annual variability) used as normalising constants.

It is clear from Figs. S1 & S2 that the ordering of models based on the relative magnitudes of errors using various error indices (NTRMSE, NCRMSE, NMAE and NEV) and other statistics (R and *d*-index) is not consistent, even for a particular season or a particular generation. This suggests that analysis may be best carried out using a range of methods to avoid the limitations of a specific statistic.

5. IMPROVEMENT SKILL AND RANKING

In the previous section all the statistical measures of model evaluation provide an overall performance skill for each model for each generation separately in terms of agreement between observation and model output. It is hard to identify the best-performing model that improved gradually from SAR to AR4 as the model performances vary from season to season, method to method, variable to variable and generation to generation. Therefore, in this section, we discuss how the models' performance improves from SAR to TAR and from TAR to AR4 on the basis of the MII defined in Section 2.3.2. We also developed an OMII from SAR to AR4, to estimate the overall numbers of improved models across generations. MII rainfall and temperature values—based on values of R, the *d*-index and other error indices (NTRMSE, NCRMSE, NMAE and NEV) in all seasons—were calculated for SAR–TAR and TAR–AR4 for all models. It is worth noting that among the statistics, higher values of R and the *d*-index indicate better agreement with data, whereas lower values for the other error statistics indicate a better model. The numbers of models improving their skill from SAR to TAR, TAR to AR4 and SAR to AR4 are summarized in Table 6 for rainfall and temperature. The total number of models analysed in each season is 7. Therefore, in total, for both for rainfall and temperature, we have 7 × 4 (MAM, JJAS, ON and DJF) × 2 (rainfall and temperature) = 56 data points or separately 28 data points each for rainfall and temperature. The numbers of models that improved performance from one generation to next were counted for each season on the basis of the MII. A model was considered improved if the estimated MII value for R and the *d*-index was >0, and if the value for the 4 error statistics was <0. The numbers of models showing improvement (not the values of the indices directly as they have different magnitudes) in each season was then summed over the 4 seasons and presented in Table 6 for each method and variable for 3 pairs of model generations, namely between SAR and TAR, TAR and AR4, and SAR and AR4. For the numbers of improved models (expressed as percentages in Table 6) averaged over 2 variables and finally averaged over 6 methods of evaluation, a modest number was improved from SAR to TAR (57%), and a higher number, from TAR to AR4 (71%). The secular trend of model improvement is most clear over the whole interval, with 75% of the analysed models indicating improvement from SAR to AR4. It is worth noting that all the percentages expressed above indicated the numbers of models with improved performance rather than the magnitude of that improvement. For rainfall, the numbers of improved models were greater than for temperature.

We have calculated the rank of each model on the basis of its overall improvement skill (OMII). This rank of each model is calculated for all seasons and statistics and for both rainfall and temperature. There is some variation in the OMII ranking of the models for the different statistical methods, and also some variation between the seasons. When, however, the seasonal ranks are summed, substantial consistency can be found in the results for the different statistics. The total accumulated rank calculated from the ranks of the 4 seasons (MAM, JJAS, ON and DJF) was summarized for all statistics, except MB, as it is similar to NMAE, for rainfall (Fig. S3a in the supplement at www.int-res.com/articles/suppl/c051p201_supp.pdf) and for temperature (Fig. S3b). Consistency is high, although there is some difference between R and the other statistics. GFDL is the most improved model for

Table 6. Number of improved models (%) on the basis of the model improvement index (MII) and overall model improvement index (OMII) from one generation to the next, across variables and methods. **Bold:** grand totals

Variable	R	*d*-index	NTRMSE	NCRMSE	NMAE	NEV	All methods
SAR to TAR (MII)							
Rainfall	46	64	54	64	61	61	58
Temperature	50	64	57	54	57	57	56
Total	48	64	55	59	59	59	**57**
TAR to AR4 (MII)							
Rainfall	75	85	68	78	64	68	73
Temperature	64	64	64	78	64	71	68
Total	70	75	66	78	64	70	**71**
SAR to AR4 (OMII)							
Rainfall	82	85	75	78	78	78	79
Temperature	64	68	68	78	71	71	70
Total	73	77	71	78	75	75	**75**

5 of the methods, with MIROC is also the second most improved model for 5 methods. Four methods agree that the least improved model is HADLEY, while all 6 methods describe ECHAM as the second least improved (Fig. S3c). The overall rank across generations calculated from all methods, variables and seasons shows that the GFDL is the most improved model, followed by MIROC, while the overall rank in the current generation (AR4) (Fig. S3d) indicates that MIROC is the best performer, followed by MME6.

6. EFFECT OF TOPOGRAPHIC CORRECTION

The observed spatial pattern of temperature is not very well simulated by any of the GCMs in any generation. In this section, we investigate to what extent this is due to topographic errors that can be expected to decrease over time simply due to increased resolution with no improvement in model physics. We restricted our attention to the AR4 models for which model topography is readily available. We modified the station temperatures by applying an average lapse rate adjustment of 0.65°C per 100 m to estimate an effective temperature at 0 m. Similarly, we also adjusted the GCM outputs at every grid location using the topography available for AR4 models. The spatial distribution of station topography, GCM topography and the results of original and topography-corrected-observation and topography-corrected-model outputs are presented in Fig. S4 in the supplement at www.int-res.com/articles/suppl/c051p201_supp.pdf. All models show quite similar patterns of topography but do not resolve the real topography well. Four of the AR4 models show a fairly similar pattern of temperature before altitude correction of temperature is applied, with CGCM and GFDL being the exceptions. After the altitude correction (Fig. S4, third column), all individual GCMs are more alike. These altitude-adjusted spatial temperature patterns of all models are also clearly more similar to the altitude-adjusted and corrected observed patterns of temperature. These results reveal that topography plays an important role in the spatially varying patterns among different models. To quantitatively investigate this effect, we recalculated all the conventional statistical measures using the topographic-corrected temperatures of models and data. In general, altitude-corrected results revealed that correction improves performance metrics (indices) (Fig. 5a–f). The improvement is most apparent for the CGCM model which had the worst reproduction of the spatial pattern of temperature (as indicated by the R statistics of model and observed temperatures, not shown here), but the other models also generally showed some improvements. The order of model performance changes somewhat with and without topographic correction (Fig. S5 in the supplement at www.int-res.com/articles/suppl/c051p201_supp.pdf), while the 2 bestmodels remain the same. For overall rank, CGCM shows considerable improvement (from sixth to third) displacing the other 3 models down 1 rank each.

7. SUMMARY AND CONCLUSIONS

The main objectives of the present study were addressed in 2 different ways: first we visually examined the ability of climate models, available for different generations viz. SAR, TAR and AR4, in simulating the seasonal cycles of rainfall and temperature over a small area in the GWBN region. A visual comparison indicates improvement for seasonal cycles across generations. These visual inspections were verified by calculating R, NMAE and MB, which are improved for the most part across generations.

We also evaluated the variation of model performance in the 4 main seasons using a wide range of conventional statistics. Due to uncertainty about the best choice of statistical method, we chose to use a variety of methods. The statistical measures such as NTRMSE, NMAE and NEV use similar types of

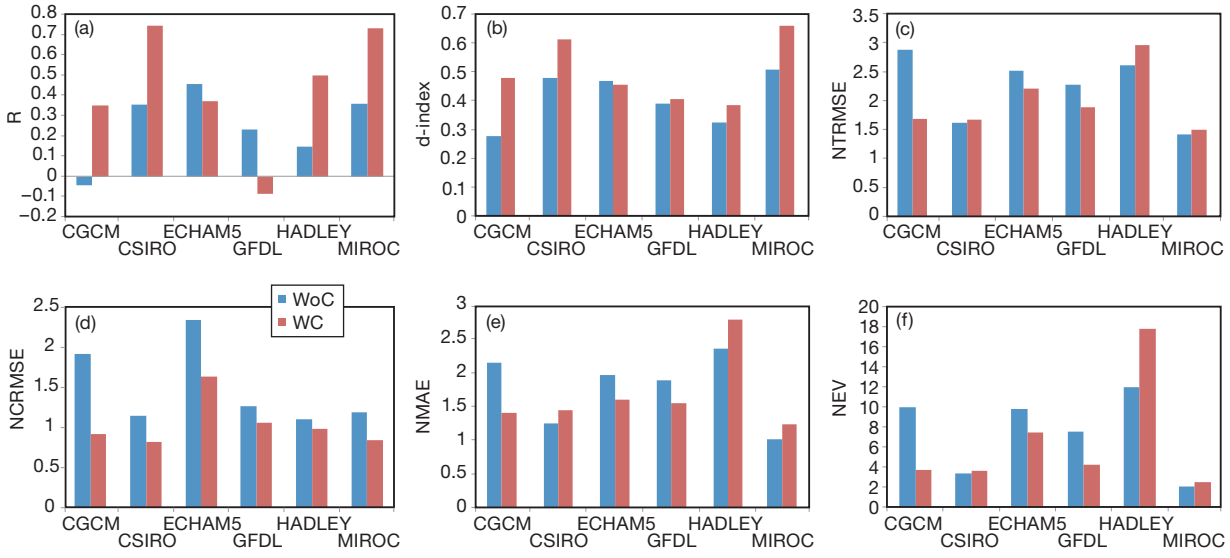

Fig. 5. Comparison for the different statistical indices (a–f) calculated for cases with and without topography correction (WC and WoC, respectively) for different AR4 models in the monsoon season (JJAS). Due to non-availability of topography data, the MME6 is absent from this analysis

mathematical expressions but are sensitive to different types of data. The NMAE is less sensitive to outliers than the NTRMSE and NEV. The results revealed that there is an improvement in the bias for temperature across generations over all seasons. The spatial correlation coefficients (R) are generally positive for the AR4, although not all are statistically significant, and mostly show improvement across generations. The majority of the models show *d*-index values >0.4 for all seasons for AR4, and results improved systematically from SAR to TAR to AR4. The magnitudes of normalised errors estimated through different statistics (NTRMSE, NCRMSE, NMAE and NNEV), in general, decrease from SAR to AR4 in almost all seasons for both temperature and rainfall.

The improvement of models from SAR to TAR, TAR to AR4 and SAR to AR4 is not uniform. On the basis of MII, the percentage of models that improved their performance in simulating current climate rose from over 57% of all statistics for SAR to TAR, to 70% for TAR to AR4. The overall percentage of improved models estimated through the OMII is about 75% from SAR to AR4.

The present study was also extended to quantify the skill of improvement of models across generations by ranking according to the OMII. Generally, the order of ranking varies slightly with season, variable and method. Averaging over variables, seasons and methods, the final ranking order of the improvement of the models (first to last) is: GFDL, MIROC, MME6, CGCM, CSIRO, ECHAM and HADCM. This ranking may be limited to the regions of the GWBN. The method of ranking by the OMII has its limitations. According to this method, any model that performs relatively poorly in the SAR but improves substantially in AR4, will secure a high rank. For example, the American GFDL model achieved the highest rank partly due to its relatively poor performance in SAR and the HADLEY model from the UK achieved the lowest rank largely because of its good performance in SAR and TAR. Because of this limitation, we also calculated the overall ranking for AR4 performance (from first to last: MIROC, MME6, ECHAM5, CSIRO, GFDL, HADLEY, CGCM). A similar performance of the MIROC model and other models is also reported by Kripalani et al. (2007) over the Indian region. Some of the weaknesses of the present study may be due, first, to observational uncertainty. We assumed that the observations are accurate; thus, the performance of the models is tested against observed data as a reference. In addition, the altitude is not taken into account when interpolating model output onto the irregularly spaced observation stations. Although the GWBN is predominantly considered a plain area, some stations within the area are situated at altitudes of a few hundred meters. In order to estimate the size of this effect, for the AR4 models, we applied a mean lapse rate correction of 0.65°C per 100 m for temperature. When the topography correction is taken into account, the changes in the improvement ranking are not dramatic, although the improvement is considerable for 1 model (CGCM). The use of flux adjustments in models may be an important factor, because almost all models (analysed in the present study) were flux

adjusted in SAR and TAR, whereas in AR4 this situation is reversed (except for CGCM). The effects of flux adjustment were discussed by Randall et al. (2007) who compared the results with non-adjustment. Our results reveal that the models improved their performance despite the loss of flux adjustments, which would (other things being equal) be expected to degrade their outputs. Finally, the perceived improvement of the models through the generations may be somewhat enhanced because the forcing of the simulations increases in realism from SAR to AR4. However, these forced changes are unlikely to be large compared to the obvious biases and errors in model output. Another important point that should be mentioned in the context of the model evaluation study is that the land use and cropping patterns in the GWBN region have changed since the 1970s in connection with the 'green revolution', especially through the increase of summer paddy crops. The summer paddy is sown in the month of December or January and harvested at the end of April or May. During this cultivation process, a huge amount of ground water was pumped out and used for irrigation because rainfall was not adequate during these months. This practice significantly affected the atmospheric moisture content and changed the local wind pattern, since moisture evaporating from the submerged summer paddy field had sufficiently high temperatures to cause disturbances in the pre-monsoon and monsoon rainfall process (Lohar & Pal 1995). This purely localized phenomenon will not be captured by the GCMs since land-use change was not incorporated into the simulations. The present study did not investigate how changes in model physics, forcing and the parameterization scheme could have led to the improvement in the models from the SAR to the AR4; rather, it includes the model resolution in the validation study by interpolating 3 generations of model grids to the station locations and assessing the impact of model resolution on model improvement. Overall, the coupled climate models have shown improvement in performance from one generation to the next over the small domain of the GWBN. The coverage of other reported studies conducted on global or comparatively larger regional scales (i.e. Randall et al. 2007, Reichler & Kim 2008) indicated improvement of model performance from one generation to the next. We do not think that our specific approach of testing model improvement across generations is unique, but we do believe that our results showing the gradual improvement across generations over a small domain adds significantly to evidence of the improvement in climate model performance over time.

Acknowledgements. We acknowledge the India Meteorological Department (IMD) and Department of Agriculture, Govt. of West Bengal, for providing observational data. We also acknowledge the modelling groups for providing their data for analysis, the PCMDI for collecting and archiving the model output, and the JSC/CLIVAR Working Groups on Coupled Modelling (WGCM) for organizing the model data analysis activity. Support of these datasets is provided by the Office of Science, U.S. Department of Energy. L.D. acknowledges the authority of Bidhan Chandra Krishi Viswavidyalaya, West Bengal, for approving a period of leave during which the work was carried out. We also thank the World Data Centre for Climate, Hamburg, for archiving the IPCC_DDC_SAR and IPCC_DDC_TAR through the link http://cera-www.dkrz.de. We also thank the members of Japanese project 'S5' funded by the Japanese Ministry of Environment (MoE), for support and discussion.

LITERATURE CITED

Alexander LV, Zhang X, Peterson TC, Caesar J and others (2006) Global observed changes in daily climate extremes of temperature and precipitation. J Geophys Res 111:D05109. doi:10.1029/2005JD006290

ASCE (American Society of Civil Engineers) (1993) Criteria for evaluation of watershed models. J Irrig Drain Eng 119:429–442

Bacher A, Oberhuber JM, Roeckner E (1998) ENSO dynamics and seasonal cycle in the tropical Pacific as simulated by the ECHAM4/OPYC3 coupled general circulation model. Clim Dyn 14:431–450

Boer GJ, Lambert SJ (2001) Second-order space–time climate difference statistics. Clim Dyn 17:213–218

Boer GJ, Flato G, Ramsden D (2000) A transient climate change simulation with greenhouse gas and aerosol forcing: projected climate to the twenty-first century. Clim Dyn 16:427–450

Covey C, AchutaRao KM, Fiorino M, Gleckler PT, Taylor KE, Wehner MF (2002) Intercomparison of climate data sets as a measure of observational uncertainty. Program for climate model diagnosis and intercomparison. UCRL-ID-147371, Lawrence Livermore National Laboratory, Livermore, CA

Covey C, AchutaRao KM, Cubasch U, Jones P and others (2003) An overview of results from the coupled model intercomparison project (CMIP). Global Planet Change 37:103–133

Das L, Lohar D (2005) Construction of climate change scenarios for a tropical monsoon region. Clim Res 30:39–52

Delworth TL, Stouffer RJ, Dixon KW, Spelman MJ and others (2002) Review of simulations of climate variability and change with the GFDL R30 coupled climate model. Clim Dyn 19:555–574

Delworth TL, Broccoli AJ, Rosati A, Stouffer RJ and others (2006) GFDL's CM2 global coupled climate models. I. Formation and simulation characteristics. J Clim 19:643–674

Dessai S, Hulme M (2008) How do UK climate scenarios compare with recent observation? Atmos Sci Lett 9: 189–195. doi:10.1002/asl.197

Dessai S, Lu X, Hulme M (2005) Limited sensitivity analysis of regional climate change probabilities for the 21st century. J Geophys Res 110:D19108. doi:10.1029/2005JD005919

Emori S, Nozawa T, Abe-Ouchi A, Numaguti A, Kimoto M, Nakajima T (1999) Coupled ocean–atmosphere model experiments of future climate change with an explicit representation of sulfate aerosol scattering. J Meteorol

Soc Jpn 77:1299–1307
Flato GM, Boer GJ (2001) Warming asymmetry in climate change simulations. Geophys Res Lett 28:195–198
Flato GM, Boer BJ, Lee WG, McFarlane NA, Ramsden D, Reader MC, Weaver AJ (2000) The Canadian Centre for Climate Modeling and Analysis global coupled model and its climate. Clim Dyn 16:451–467
Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113:D06104. doi:10.1029/2007JD008972
Gordon HB, Rotstayn LD, McGregor JL, Dix MR and others (2002) The CSIRO Mk3 climate system model [electronic publication]. Tech Pap No. 60, CSIRO Atmospheric Research, Aspendale
Haywood JM, Stouffer RJ, Wetherald RT, Manabe S, Ramaswamy V (1997) Transient response of a coupled model to estimated changes in greenhouse gas and sulfate concentrations. Geophys Res Lett 24:1335–1338
Hirst AC, Gordon HB, O'Farrell SP (1996) Response of a coupled ocean–atmosphere model including oceanic eddy-induced advection to anthropogenic CO_2 increase. Geophys Res Lett 23:3361–3364
Hirst AC, O'Farrell SP, Gordon HB (2000) Comparison of a coupled ocean–atmosphere model with and without oceanic eddy-induced advection. 1. Ocean spin-up and control integrations. J Clim 13:139–163
Hulme M, Briffa KR, Jones PD, Senior CA (1993) Validation of GCM control simulations using indices of daily airflow types over the British Isles. Clim Dyn 9:95–105
Janssen PHM, Heuberger PSC (1995) Calibration of process-oriented models. Ecol Modell 83:55–66
Johns TC, Carnell RE, Crossley JF, Gregory JM and others (1997) The second Hadley Centre coupled ocean–atmosphere GCM: model description, spin-up and validation. Clim Dyn 13:103–134
Johns TC, Gregory JM, Ingram WJ, Johnson CE and others (2003) Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. Clim Dyn 20:583–612
Johns TC, Durman CF, Banks HT, Roberts MJ and others (2004) HadGEM1—Model description and analysis of preliminary experiments for the IPCC 4th assessment report. Tech Rep 55, Met Offices, Exeter
Johns TC, Durman CF, Banks HT, Roberts MJ and others (2006) The new Hadley centre climate model HadGEM1: evaluation of coupled simulations. J Clim 19:1327–1353
K-1 Model Developers (2004) K-1 coupled model (MIROC) description. Tech Rep 1, Center for Climate System Research, University of Tokyo. www.ccsr.u-tokyo.ac.jp/kyosei/hasumi/MIROC/tech-repo.pdf
Kripalani RH, Oh JH, Kulkarni A, Sabade SS, Chaudhari HS (2007) Indian summer monsoon precipitation variability: coupled climate model simulations and projections under IPCC AR4. Theor Appl Climatol 90:133–159
Kunkel KE, Liang XL, Zhu J, Lin Y (2006) Can CGCMs simulates the twentieth-century 'warming hole' in the central United States? J Clim 19:4137–4153
Legates DR, McCabe GJ (1999) Evaluation of 'goodness of fit' measures in hydrological and hydro-climatic model validation. Water Resour Res 35:233–241
Lohar D, Pal B (1995) The effect of irrigation on premonsoon season precipitation over south West Bengal, India. J Clim 8:2567–2570
Maxino CC, McAvaney BJ, Pitman AJ, Perkins SE (2007) Ranking the AR4 climate models over the Murrray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. Int J Climatol 28:1097–1112
McFarlane NA, Scinocca JN, Lazare M, Harvey R, Verseghy D, Li J (2005) The CCCma third generation atmospheric general circulation model. Canadian Centre for Climate Modelling and Analysis internal report, Victoria
Meehl GA, Boer BJ, Covey C, Latif M, Stouffer RJ (2000) The coupled model intercomparison project (CMIP). Bull Am Meteorol Soc 81:313–318
Meehl GA, Covey C, McAvaney B, Latif M, Stouffer RJ (2005) Overview of the coupled model intercomparison project. Bull Am Meteorol Soc 86:89–93
Min SK, Hense A (2006) A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. Geophys Res Lett 33: L08708. doi:10.1029/2006GL025779
Min SK, Legutke S, Hense A, Kwon WT (2005) Internal variability in a 1000-year control simulation with the coupled climate model ECHO-G. I. Near surface temperature, precipitation, and mean sea level pressure. Tellus 57A: 605–621
Moriasi DN, Arnold JG, Van Liew MW, Bigner RL, Harmel RD, Veith TL (2007) Model evaluation guideline for systematic quantification of accuracy in watershed simulation. Transac Am Soc Agric Biological Eng 50:885–900
Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modeling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772
Nieto S, Rodríguez-Puebla C (2006) Comparison of precipitation from observed data and general circulation models over the Iberian Peninsula. J Clim 19:4254–4275
Nozawa T, Emori S, Numaguti A, Tsushima Y and others (2001) Projections of future climate change in the 21st century simulated by the CCSR/NIES CGCM under the IPCC SRES scenario. In: Matsuno T, Kida H (eds) Present and future of modeling global environmental change—toward integrated modeling. Terra Scientific Publishing Company, Tokyo, p 15–28
Palutikof JP, Winkler JA, Goodess CM, Andresen JA (1997) The simulation of daily temperature time series from GCM output. I. Comparison of model data with observations. J Clim 10:2497–2513
PCMDI (Program for Climate Model Diagnosis and Intercomparison) (2007) IPCC model output. Available at: www.pcmdi.llnl.gov/ipcc/about_ipcc.php
Randall DA, Wood RA, Bony S, Colman R and others (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M, Chen Z and others (eds) Climate change 2007: the physical science basis. Contribution of Working Group I to the 4th assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
Reichler T, Kim J (2008) How well do coupled models simulate today's climate? Bull Am Meteorol Soc 89:303–311
Stendel M, Schmith T, Roeckner E, Cubasch U (2002) The climate of the 21st century: transient simulations with a coupled atmosphere–ocean general circulation model, revised version. Climate Centre Report 02-1, Danish Meteorological Institute, Copenhagen
Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106: 7183–7192
Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194
Willmott CJ (1982) Some comments on the evaluation of model performance. Bull Am Meteorol Soc 63:1309–1313

Editorial responsibility: Peter Gleckler, Livermore, California, USA

Submitted: May 16, 2011; Accepted: October 6, 2011
Proofs received from author(s): March 14, 2012