Reducing uncertainty in selecting climate models for hydrological impact assessments

A. J. PITMAN[1] & S. E. PERKINS[2]

1 *Climate Change Centre, University of New South Wales, Sydney, New South Wales, Australia*
a.pitman@unsw.edu.au

2 *Department of Physical Geography, Macquarie University, New South Wales 2109, Australia*

Abstract Deciding which climate models to use to assess the impact of climate change on water resources is particularly difficult in environments where precipitation dominates resource vulnerability. We show that assessing climate models based on their simulation of mean precipitation provides little guide to a model's ability to simulate the more extreme events that affect hydrological systems. In contrast, a probability density function based assessment using daily climate model data provides a good basis for confidence in a model's ability to simulate the 95th rainfall percentile. We demonstrate that climate models have useful skill in simulating observed probability density functions over two regions of Australia, although the well-known bias of excess rainfall at low rates remains common. We conclude by identifying those climate models that produce the best basis for hydrological impacts assessment over two regions of Australia.

Key words climate models; probability density function; skill-score

INTRODUCTION

Climate models are our principal tools for projecting future climate (Houghton *et al.*, 2001). The Intergovernmental Panel on Climate Change (IPCC) Third Assessment Report concluded that they provide "credible simulations of climate, at least down to sub-continental scales and over temporal scales from seasonal to decadal" (McAvaney *et al.*, 2001). This evaluation was based on the ability of climate models to simulate a range of diagnostics including means and variances, past climates, the El Nino-Southern Oscillation, monsoons and other specific modes of variability.

One application of climate models is the projection of future climate. Some variables that are directly resolved by the models using primitive equations (see McGuffie & Henderson-Sellers, 1997) are likely reliable at large spatial scales (e.g. temperature, wind and pressure). Variables that result from the interactions of many physical processes are more challenging to simulate (e.g. precipitation). Variables that are calculated using quantities like precipitation, coupled with spatial variability in soil properties, vegetation, slope, etc. are likely increasingly uncertain as the complexity of interactions of nonlinear processes at different time and space scales increases. Unfortunately, impact modellers commonly need these more difficult-to model-quantities. There are therefore on-going attempts to improve model estimates of soil moisture, runoff, etc. (Wood *et al.*, 1998; Schlosser *et al.*, 2000; Nijssen *et al.*, 2003).

As the starting point in simulating water resources well in climate models is a good simulation of precipitation, an analysis of the skill of climate models in simulating precipitation is useful. Arora (2001) analysed one climate model's skill in simulating runoff and precipitation. He noted limits in the model's skill in simulating precipitation at regional scales and subsequent limits to basin-scale runoff. While mean annual precipitation was within 20% of the observed estimates for 13 of 23 basins considered, this reduced to only 4 of 23 basins in the simulation of runoff.

Climate models simulate large-scale (say continental) rainfall well on the annual and seasonal timescales (McAvaney *et al.*, 2001). However, Sun *et al.* (2006) analysed daily precipitation data from 18 coupled global climate models that underpin the Fourth Assessment Report conducted by the IPCC (AR4) in terms of precipitation frequency, intensity, and the number of rainy days contributing to most (i.e. 67%) of the annual precipitation total. They showed that, for light precipitation (1–10 mm d^{-1}), most models overestimate the frequency but produce patterns of the intensity that are in broad agreement with observations. In contrast, for heavy precipitation (>10 mm d^{-1}), most models considerably underestimate the intensity but simulate the frequency relatively well. This has significant potential implications for hydrological modelling – the AR4 climate models simulate total rainfall well but fail to capture the magnitude–frequency relationships well. An implication of this might be, for example, to encourage the use of idealized climate change scenarios where rainfall is changed by a given amount to explore the runoff responses (e.g. Chiew & McMahon, 2002). This approach is not particularly popular within the climate community because it decouples the hydrological response from the climate and vegetation responses. This means that changes in many significant feedbacks, such as snow accumulation/melt, evaporative demand, or vegetation (stomatal function, root depth etc.), are ignored or prescribed. Among climate modellers, a preferred approach has been to minimize biases via multi-model ensembles. This has been believed to reduce overall biases (Cubasch *et al.*, 2001; Nohara *et al.*, 2006). Clearly, if most models contain systematic biases and they are averaged, this risks providing a false sense of agreement. Our key problem is therefore the recognition that climate model projections of changes in precipitation and runoff are very important for water resource planning (Seckler *et al.*, 1999; Vörösmarty *et al.*, 2000; Arnell, 2004) but we are aware of systematic biases in climate models' simulation of the key driving variable, precipitation (Sun *et al.*, 2006).

To partially resolve this issue and to try to provide guidance to users of climate model data we have explored alternative ways to assess the skill of climate models. As part of this assessment we have used daily simulated precipitation because this averaging over many days to monthly, seasonal or annual averages can hide systematic biases. We have chosen to assess models *not* in terms of mean precipitation but in terms of the probability of the simulation of an amount of precipitation on a given rain day. This has enabled us to assess models for frequency of rainfall, the amount of precipitation that occurs with a particular time frequency, and ultimately to assess the models across the full range of observed probabilities. We utilize the model results submitted to the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory in the USA (http://www-pcmdi.llnl.-gov/about_ipcc.php) as part of the AR4. We base our analysis on probability density functions (PDFs) to study the distribution of simulated values for a given variable using daily data archived at PCMDI to allow an evaluation of variables at timescales that are lost in monthly or seasonal means. Using PDFs as the basic unit of analysis has other advantages over using means as spatially inhomogeneous data and observed time series of stations that contain fractional temporal coverage can be used reliably.

The data and analysis methods used in this paper are described in detail in the Appendix and in Perkins *et al.* (2007). Briefly, daily climate model data over Australia for precipitation taken from the *Climate of the Twentieth Century* simulations were used from a large sample of climate models (Table 1). Due to the problem of missing data from some models, a total of 16 models and 39 runs were available for precipitation. We compared these data to daily observed precipitation from the Australian Bureau of Meteorology (BOM) for the period 1961–2000. In calculating PDFs we combined all data within each 9.75° × 10.75° rectangle (Fig. 1). PDFs were calculated for two ~10° × 10° regions centred on New South Wales and Queensland. Observed and model data were centred on bins of 1 mm d^{-1}. Modelled precipitation <0.2 mm d^{-1} was omitted as these are not recorded in the observations. We then derived a PDF-based metric that calculates the cumulative minimum value of two distributions of each binned value, thereby measuring the common area between two PDFs. This measure provides a robust and comparable measure of the relative similarity between model and observed PDFs, and is likely preferable to ad hoc weightings based on statistical tests. We now provide an examination of the skill-based performance, a discussion of these results and conclusions.

ANALYSIS OF MODELLED PRECIPITATION

Mean precipitation

A traditional evaluation of a climate model would focus on seasonal or annual mean precipitation. Table 2 shows the mean simulated and observed precipitation for each

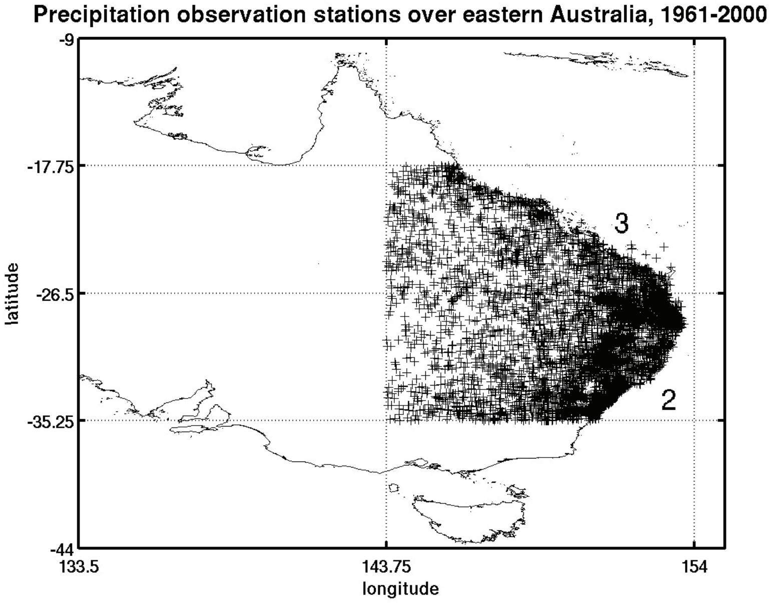
Precipitation observation stations over eastern Australia, 1961-2000
-9
-17.75
-26.5
-35.25
-44
latitude
133.5
143.75
154
longitude
3
2

Fig. 1 Locations of observed precipitation over Regions 2 and 3 with data available between 1961 and 2000. The large rectangles are numbered for reference in the text.

Table 1 All climate models with daily data for precipitation available from PCMDI. Column 1 is the acronym used in the text. Column 2 is the name of the model used in the PCMDI archive, column 3 is the source of the model (see http://www-pcmdi.llnl.gov/ipvv/about ipcc.php).

Acronym	Model	Affiliation
BCCR	bccr_bccm2_0	Bjerknes Centre for Climate Research, Univ. Bergen, Norway
CGCM-h	cccma_cgcm3_1_t63	Canadian Centre for Climate Modeling and Analysis
CGCM-l	cccma_cgcm3_1_t47	Canadian Centre for Climate Modeling and Analysis
CSIRO	csiro_mk3_0	Australian Commonwealth Scientific and Res. Organization
GFDL2.0	gfdl_cm2_0	Geophysical Fluid Dynamics Laboratory
GFDL2.1	gfdl_cm2_1	Geophysical Fluid Dynamics Laboratory
GISSAOM	giss_aom	Goddard Institute of Space Studies (NASA)
GISSER	giss_model_e_r	Goddard Institute of Space Studies (NASA)
FGOALS	iap_fgoals1_o_g	Institute of Atmospheric Physics, Chinese Academy of Sci.
IPSL	ipsl_cm4	Insitut Pierre Simon Laplace, France
MIROC-h	miroc3_2_hires	Centre for Climate System Research, University of Tokyo; National Institute for Environmental Studies; Frontier Research Centre for Global Change
MIROC-m	miroc3_2_medres	As MIROC-h
ECHO-G	miub_echo_g	Max Planck Institut für Meteorologie
ECHAM	mpi_echam5	Max Planck Institut für Meteorologie
MRI	mri_cgcm2_3_2a	Japan Meteorological Agency
CCSM	ncar_ccsm3	National Centre for Atmospheric Research
PCM	ncar_pcm1	National Centre for Atmospheric Research

model for Regions 2 and 3. This type of assessment is not very useful as a model could simulate the mean well but fail to capture the observed magnitude–frequency relationships. However, if the models are ranked by annual mean precipitation, for Region 2 only GISS ER, MIROC-m, and FGOALS fall within 20% of the observed. For Region 3, ECHAM, FGOALS, MIROC-m fall within 20% of the observed. Some models simulate mean errors exceeding 50% (Region 2, IPSL, MRI, GISS AOM, GFDL2.0; Region 3, IPSL, BCCR, GISS AOM, MRI, GFDL2.0, ECHAM). One might therefore decide to choose the only models that fall within 20% for both regions (MIROC-m and FGOALS) for impact assessment. The remainder of this paper explores whether this would be a sound decision.

PDF-based model evaluation

Figure 2 shows that in both regions most models underestimate the probability of rainfall in low amounts (note the *x*-axis is the square root of the simulated and observed values). At precipitation amounts of less than ~1 mm d^{-1} most models over predict the probability of precipitation by up to a factor of four. Figure 3 shows the skill-scores for precipitation for the same two regions. These range from ~0.4 (GISS AOM, Region 3) to >0.8 (BCCR, ECHO-G and ECHAM for both Regions 2 and 3). Thus, in the case of BCCR, ECHO-G and ECHAM, these models overlap the observed PDF by >80%. Given these are fully-coupled climate models run globally and then assessed over two 10° × 10° regions is an impressive achievement. The r^2 value of each model's skill score in Region 2 compared with Region 3 is 0.74, indicating that the good models are consistently good and the poor models are consistently poor on a

Table 2 Mean precipitation amount for each model for Regions 2 and 3. The difference of the mean from the observed is shown as an amount and as a percentage. The ranking of the models based on mean performance is shown in column 5. The PDF skill score and the rank based on this score are shown in the final two columns.

Region 2	Mean (mm d^{-1})	Difference from observed	% difference from the observed	Mean Rank	PDF-skill score	PDF rank
BCCR	3.26	–0.92	39.13	8	0.82	3
CGCM-l	1.35	0.99	–42.39	9	0.57	14
CSIRO	1.57	0.77	–32.93	6	0.70	8
GFDL2.0	1.16	1.19	–50.67	11	0.74	4
GFDL2.1	1.46	0.88	–37.61	7	0.73	5
GISS AOM	0.94	1.40	–59.73	12	0.67	10
GISS ER	2.24	0.11	–4.63	1	0.72	6
FGOALS	2.02	0.33	–14.02	3	0.68	9
IPSL	0.54	1.80	–76.86	14	0.64	12
MIROC-m	2.22	0.12	–5.30	2	0.72	6
ECHO-G	1.75	0.59	–25.36	5	0.86	1
ECHAM	1.23	1.11	–47.45	10	0.84	2
MRI	0.83	1.52	–64.70	13	0.61	13
CCSM	1.82	0.53	–22.43	4	0.66	11
Observed	2.34					

Region 3	Mean (mm d^{-1})	Difference from observed	% difference from the observed	Mean Rank	PDF-skill score	PDF rank
BCCR	3.96	–1.70	74.95	13	0.80	3
CGCM-l	1.33	0.93	–41.11	8	0.49	13
CSIRO	1.57	0.69	–30.58	6	0.68	5
GFDL2.0	1.06	1.21	–53.32	10	0.65	7
GFDL2.1	1.36	0.90	–39.93	7	0.61	9
GISS AOM	0.65	1.61	–71.15	12	0.43	14
GISS ER	2.81	–0.54	24.07	4	0.71	4
FGOALS	2.14	0.13	–5.60	2	0.63	8
IPSL	0.42	1.84	–81.39	14	0.51	12
MIROC-m	2.56	–0.29	12.96	3	0.68	5
ECHO-G	2.15	0.11	–4.88	1	0.81	2
ECHAM	1.12	1.15	–50.61	9	0.84	1
MRI	0.79	1.48	–65.20	11	0.55	11
CCSM	1.60	0.66	–29.24	5	0.58	10
Observed	2.26					

PDF-based assessment of two regions. FGOALS, which performed well in the mean assessment, ranked 8th/9th for the two regions (of 14 models), while MIROC-m ranked 5th/6th for the two regions.

The skill scores shown in Fig. 3 assess the overall PDFs. This contrasts with an evaluation based on the mean precipitation that assesses one part of the PDF. It is a harder test of a model to simulate the full PDF than (say) the mean. However, because the shape of the observed PDF shows most rainfall falls at rates that are hydrologically relatively unimportant (less than 1 mm d^{-1}), a model could appear to be very good

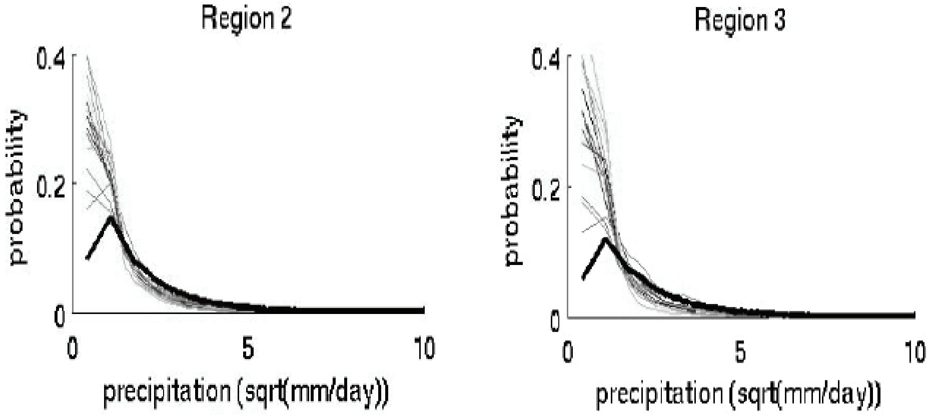

Fig. 2 Probability density functions for precipitation for each climate model. The observed PDF is shown in the solid line. The *x*-axis has been square-rooted to aid interpretation.

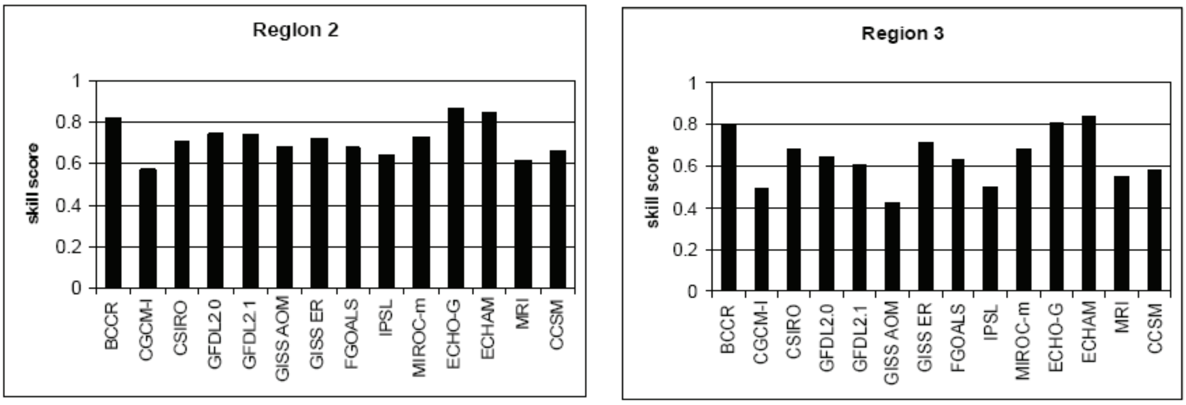

Fig. 3 PDF-based skill scores for precipitation for Regions 2 and 3 for each model. A perfect score is 1.0.

based on the skill score while being very bad at the extremes of the distribution (no rain, or precipitation exceeding amounts that are low probabilities). It is therefore useful to explore whether models that simulate the PDFs well also capture the tails of the observed distribution well.

Frequency/magnitude relationships

Figure 4 shows the percent of total precipitation that occurs at rates between 0.01–0.5 mm d^{-1}, 0.5–1.0 mm d^{-1} and over 1 mm d^{-1} compared to the observed. Note that the bar representing rates >1 mm d^{-1} has been square rooted to allow the other bars to be displayed clearly. This can be compared to Fig. 5 which shows the percentage of days with precipitation <0.01 mm d^{-1}, 0.01–0.5 mm d^{-1}, 0.5–1.0 mm d^{-1} and >1 mm d^{-1}. The observed shows that 99% of rainfall occurs at rates >1 mm d^{-1}. The models range from 83% (GISS AOM, Region 3) and less than 90% (CGCM-l, Region 2 and 3, IPSL Region 3) to several models that simulate more than 98% of precipitation at rates

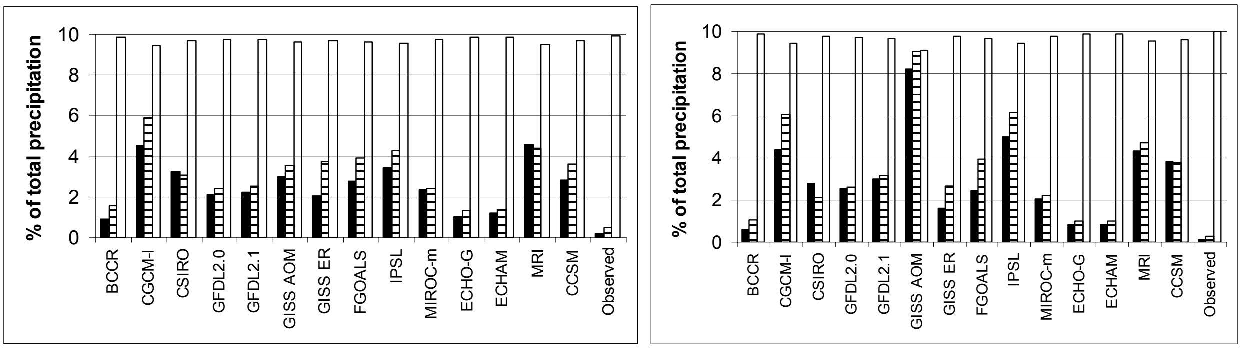

Fig. 4 The percentage of total precipitation that occurs at rates of 0.01–0.5 mm d^{-1} (solid bar), 0.5–1.0 mm d^{-1} (middle bar) and >1 mm d^{-1} (open bar) compared to the observed (right hand of each panel). Note that the open bar representing rates >1 mm d^{-1} has been square-rooted to allow the other bars to be displayed clearly.

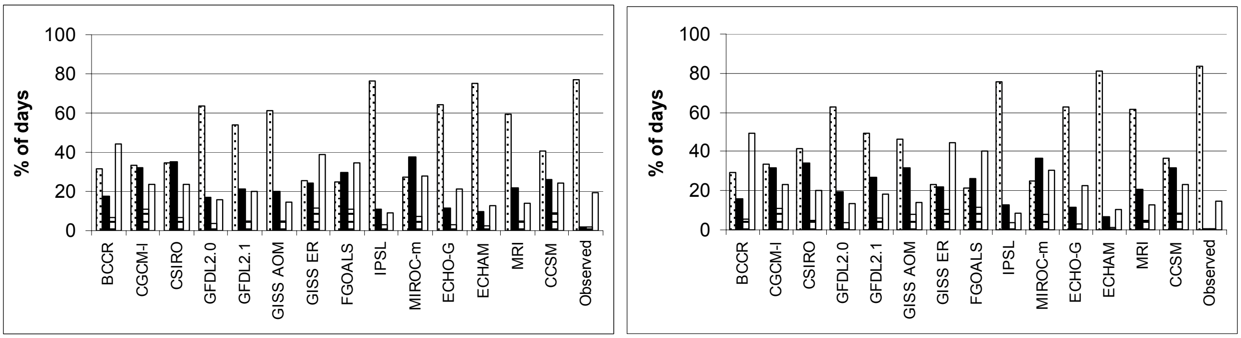

Fig. 5 Percentage of days with total precipitation occurring at rates <0.01 mm d^{-1}, (dotted bar), 0.01–0.5 mm d^{-1} (solid bar), 0.5–1.0 mm d^{-1} (middle bar) and >1 mm d^{-1} (open bar) compared to the observed (right hand of each panel). Note that the open bar representing rates >1 mm d^{-1} has been square-rooted to allow the other bars to be displayed clearly.

>1 mm d^{-1} (ECHO-G, ECHAM, BCCR) in at least one of the regions. A 1–2% error in this measure in a model is unlikely to be of concern, but 10% of total rainfall occurring at rates below observed may affect the terrestrial hydrometeorology. Specifically, excessive drizzle is likely to lead to excessive interception and evaporation loss, relative to soil moisture recharge and runoff (see Pitman *et al.*, 1990), and lead to poor partitioning of available energy between sensible and latent heat and poor partitioning of available water between runoff and evaporation.

The percentage of days with rainfall exceeding the various thresholds (Fig. 5) shows a similar result. The observed suggests approximately 80% of days should be dry in both Regions 2 and 3. Model simulations by IPSL and ECHAM are close to 75% of days which is likely to be good enough. However, BCCR, CGCM-l, GISS ER, FGOALS and MIROC-m all simulate less than 35% of days being dry. In MIROC-m and CGCM-l this is the result of excessive rainfall in the range 0.01–0.5 mm d^{-1} but in FGOALS, GISS ER and BCCR the excessive rainfall days are at rates >0.5 mm d^{-1}. These models are basically raining at significant rates on about double the number of days observed in both regions. This has the potential to affect the regional hydrometeorology. Recall that FGOALS ranked 3rd and 2nd and GISS ER ranked 1st and 4th for Regions 2 and 3 respectively in terms of mean rainfall (Table 2). These models simulate the *amount* of precipitation well, but do not capture the magnitude–frequency relationships.

Simulation of 80th, 90th and 95th percentiles

One advantage of using PDFs as the basis for the analysis of the models is that they can be assessed against higher percentile values. We used the 80th, 90th, and 95th percentiles as measures of how well the models could simulate these rarer values that are not easily interpreted from the PDFs. Note that all days where precipitation was <0.2 mm d^{-1} were omitted from this analysis.

Figure 6(a) and (b) shows the model results corresponding to each percentile for the two regions. Precipitation, at these higher percentiles, is poorly captured by all models. Specifically, the highest values simulated by any model for the 95th percentile is 25.8 mm d^{-1} (ECHAM, Fig. 6(a)) compared to the observed 37.6 mm d^{-1} (Region 2). For Region 3, ECHAM is again best but the simulated 95th percentile is 27.0 mm d^{-1} compared to the observed 50.4 mm d^{-1} (Fig. 6(b)). Recall that ECHAM ranked 10th and 9th for Regions 2 and 3 respectively on the means-based assessment (Table 2). A general result is that for both regions, most models' 95th percentile matches the observed 80th percentile most closely.

Figure 6(c) shows the relationship between the simulated 95th percentile expressed as a difference from the observed and the model's skill score integrated across the whole PDF. A clear relationship is visible for both Region 2 (squares) and Region 3 (circles). As a model's skill score increases, the difference between the simulated and observed precipitation at the 95th percentile declines. The improvement in the simulated 95th percentile as the skill score improves is not trivial – the best models are substantially better although those with skill scores exceeding 0.8 still underestimate the observed 95th percentile by 30–40% in Region 2 and by 50% in Region 3. However, the clear relationship between PDF skill score and skill in the 95th percentile shown in Fig. 6(c) demonstrates that the skill score is a reasonable measure of a model's ability at the tail of the simulated distribution. We find no examples of where a model captures the 95th (or other percentiles) well when the overall skill score is weak. In contrast, the skill of a model to capture the mean provides no guide to how well the 95th percentile will be simulated (Fig. 6(d)).

DISCUSSION

A climate model with skill across a range of observed PDFs shows a capacity to simulate the full range of climates within a region. If a climate model can accurately simulate the probabilities two standard deviations from the current mean, this suggests that they should be able to simulate the greater proportion of future climates, at least until rainfall changes such that the PDFs are substantially different. An evaluation of the regional PDFs of precipitation for each of the AR4 models show, as expected, a range of performances. These performances were quantified via a skill-score that measured the degree of overlaps of the PDFs.

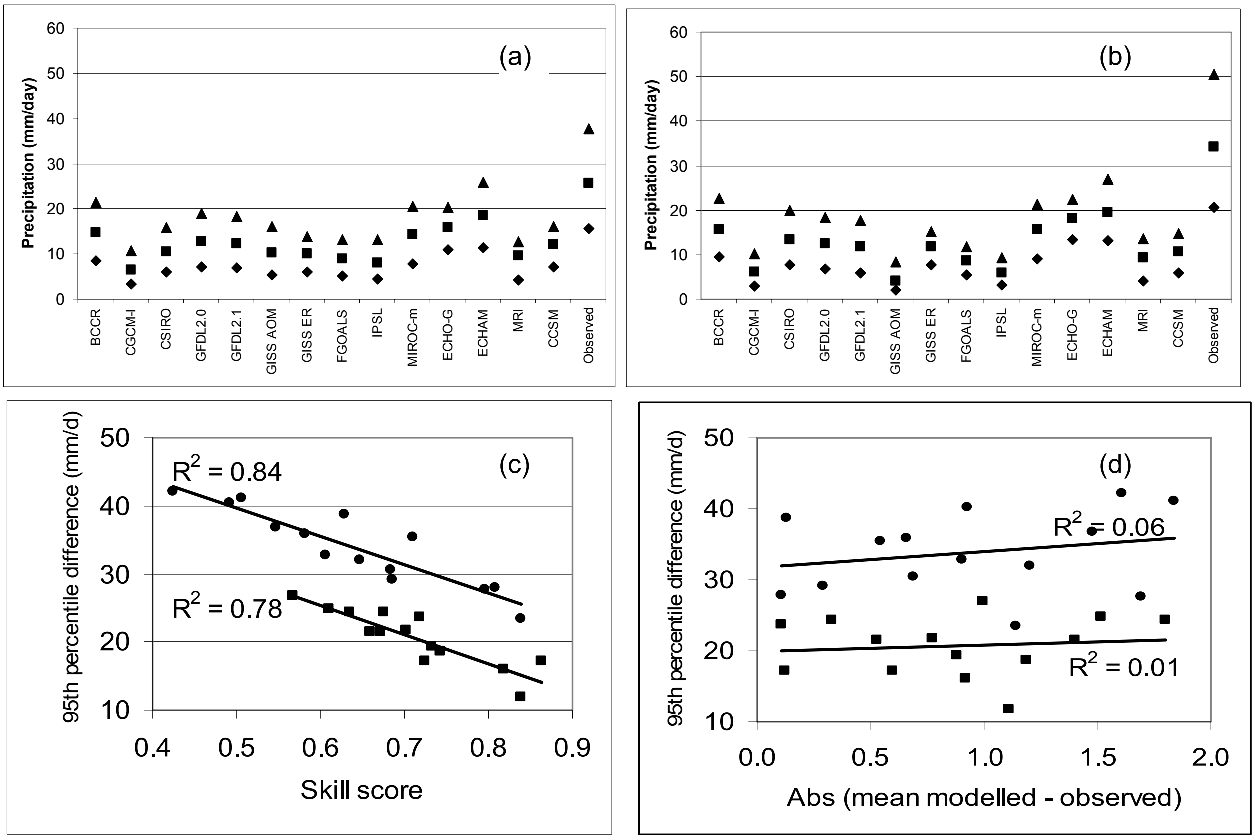

Fig. 6 Percentile values for precipitation for: (a) Region 2, and (b) Region 3. Diamonds represent the 80th percentile, squares the 90th and triangles the 95th. The observed is shown at the right-hand side of each figure. (c) Difference between the modelled 95th percentile and the observed for Region 2 (squares) and Region 3 (circles) *vs* the regional skill score. The line is a line of best fit with an associated r^2 value plotted. Note that precipitation amounts less than 0.2 mm d^{-1} are omitted from the calculation of the return intervals. (d) Difference between the modelled 95th percentile and the observed for Region 2 (squares) and Region 3 (circles) *vs* the absolute value of the modelled mean minus observed.

First, it was reassuring how well several models reproduced the observed PDFs. It is demanding for a global fully-coupled climate model to undertake such a task with considerable confidence. The skill shown by most models strongly supports previous assessments that climate models are useful tools (e.g. McAvaney *et al.*, 2001). Figure 3 shows, for precipitation, that the best models are ECHO-G, ECHAM and BCCR. These three models are the only ones with a skill score >0.8 and these models achieve 0.8 in both regions.

In terms of the ability of the models to simulate the percentiles, all models underestimated the observed 95th percentile. However, the models that were closest to the observed were ECHAM, BCCR, ECHO-G and MIROC-m. These were also the four models that were best ranked in the simulation of the percentage of days with total precipitation occurring at rates over 1 mm d^{-1}.

The key issue in our results is the little similarity between those models one would choose for impacts modelling given an evaluation based on a mean as distinct from a PDF. The best models in terms of the mean when combined over both regions were

MIROC-m and FGOALS. In terms of ranking, GISS ER, CCSM and ECHO-G also performed well. Of these models, MIROC-m is mid-range on the PDF skill score, CCSM and FGOALS are relatively poor and GISS ER is mid-range. ECHO-G, however, ranks 1st for Region 2 and 2nd for Region 3. ECHAM (ranked 2nd, Region 2 and 1st Region 3), ECHO-G and BCCR are consistently ranked top three in PDF-score, 95th percentile score, and percentage of rainfall occurring at rates over 1 mm d^{-1}. Indeed, there is no relationship between mean performance ranking and PDF-based ranking ($r^2 = 0.12$) indicating that an assessment based on the mean does not inform regarding the capacity of the model to simulate other aspects of precipitation magnitude or frequency.

CONCLUSIONS

The evaluation of climate models against observed data is an important step in building confidence in their use for impact assessment. While climate models can be evaluated in many ways, the most common methods explore model performance in annual, seasonal or monthly means. These are not the time scales that will most strongly affect hydrological systems. Despite limitations in the evaluation of climate models via means, we show that the best five models are GISS ER, FGOALS, MIROC-m, ECHO-G and CCSM. These rank in the top five for mean precipitation for both regions. However, on other measures, such as the overall PDF, the 95th percentile and the percentage of rainfall occurring at rates >1 mm d^{-1}, ECHO-G (and mostly MIROC-m) is the only one of these models that performs well.

To address the limitations of mean-based evaluation of climate models, we examined the capacity of the AR4 models to simulate the observed PDFs using daily data. The skill of each climate model to reproduce the PDF was assessed using a skill-score. While large biases were identified in some models, in general, several of the AR4 climate models showed considerable skill in representing the observed PDFs. The best three models are ECHO-G, ECHAM and BCCR, which are ranked in the top three for both regions. These same three models ranked top three in simulating the 95th percentile and the percentage of rainfall occurring at rates >1 mm d^{-1}. Thus, given these are likely more significant precipitation statistics than the mean for impacts assessment, a decision on which models to use based on mean performance would omit one of the best models.

It is useful to assess each model that ranks highly on one or more measures in turn. For hydrological impacts assessment over the two regions discussed in this paper, the use of GISS ER, FGOALS and CCSM is not recommended because their strong mean-based performance is countered by a poor simulation of the overall PDF, the 95th percentile and the percentage of rainfall occurring at rates >1 mm d^{-1}. The strong performance of BCCR and ECHAM in all PDF-related quantities is countered by a poor simulation of the mean, and it would need to be determined whether the mean rainfall was central to the hydrological system being explored before using these models. This leaves two models: MIROC-m and ECHO-G. These two models rank well in all measures: they are always within the top five on mean, PDF-score, 95th percentile and the percentage of rainfall occurring at rates >1 mm d^{-1}.

McAvaney *et al.* (2001) concluded that climate models were useful tools, at least down to sub-continental scales. Our analysis, while limited to two sub-continental regions, suggests that some of the AR4 models show considerable skill at these scales, even when assessed using daily data. This builds confidence in the use of these models for regional assessment. However, we also note that some models show major biases that need to be addressed. All of the models reported here are included in the AR4 assessment and clearly, at least over Australia, all models are not equally good. An important aspect of model evaluation is the timescale used. In our view, using PDFs based on daily data is relatively straightforward and as shown in Fig. 6(c), the resulting score provides guidance on the model's ability to simulate more extreme values. The model mean does not provide this guidance and thus we conclude that for applications where the mean is not the key driver, an assessment based on the PDF is likely to be a more reliable guide to selecting models. However, it is of course not enough to simply assess climate models based on their 20th century PDF of precipitation. Other metrics that evaluate the simulation of atmosphere and ocean dynamics, for example, as well as other specific phenomena (see McAvaney *et al.*, 2001 for other examples) have to be used. However, in exploring hydrological systems, and as a first step, the capacity of a climate model to simulate the PDF of 20th century precipitation is a very useful evaluation method that is preferable to assessing models via the simulation of the mean.

APPENDIX: DATA AND METHODS

Climate model and observed data

Daily climate model data over Australia for were taken from the PCMDI archive (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). Data from 1961–2000 from the *Climate of the Twentieth Century* simulations were used as this time period was common to all models. We found that differences between realizations from a single climate model in the simulated PDFs were negligible and we present ensembles over the available realizations for each climate model. Due to the problem of missing data from some models for some variables, a total of 14 models and 35 individual realizations were available. Model specific masks were fitted to exclude ocean data.

Observed data

Daily observed precipitation (*P*) data were obtained from the Australian Bureau of Meteorology (BOM) for the period 1961–2000. A total of 12 525 precipitation stations contained data for all, or a part of, this time period. Some individual stations contained missing data but remaining data for an incomplete station were included in the calculation of the PDFs.

Homogenization and quality control of observed data is a common problem in model evaluation. Quality control of observed data is vital when means or standard deviations are calculated, as common and/or large outliers can significantly affect these statistics. We use PDFs as the basis of our analysis in part because they are less likely to be affected by observation errors than the mean or standard deviation, and in part because they allow a more complete assessment of a climate model's capacity to simulate the complete range of observations at daily time scales.

Calculation of PDFs

Using MatLab (http://www.mathworks.com), PDFs were calculated for each of 12 ~10° × 10° squares over Australia for *P*. Observed and model data were binned around centres determined by the range of the observed data for the variable in question, unique to each region. Bin sizes of 1 mm d^{-1} for *P* were used. All daily values of P below 0.2 mm d^{-1} were omitted because rates below this amount are not recorded in the observations. The PDF of the observed values was smoothed to remove artificial variability caused by observer biases (values immediately after the decimal point tended to be biased to either zero or five). This did not affect the resulting skill scores to an extent that affect the conclusions.

Skill-score

We devised a metric that appears to be a very simple but very useful measure of similarity between two PDFs, which allows a comparison across the entire PDF. This metric calculates the cumulative minimum value of two distributions of each binned value, thereby measuring the common area between two PDFs. If a model simulates the observed conditions perfectly, the skill-score (S_{score}) will equal one, which is the total sum of the probability at each bin centre in a given PDF. This is a very simple measure that provides a robust and comparable measure of the relative similarity between model and observed PDFs, and is likely preferable to ad hoc weightings based on statistical tests. We base our analysis on this statistic because it is clear, easily interpreted and directly comparable across variables. It also has the virtue of providing a quantitative measure of similarity comparable to what would be assessed by eye.

Acknowledgements We acknowledge the international modelling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modeling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, US Department of Energy.

REFERENCES

Arnell, N. W. (2004) Climate change and global water resources: SRES emissions and socio-economic scenarios. *Global Environ. Change* **14,** 31–52.

Arora, V. K. (2001) Streamflow simulations for continental-scale river basins in a global atmospheric general circulation model. *Adv. Water Resour.* **24**, 775–791.

Chiew, F. H. S. & McMahon, T. A. (2002) Modelling the impacts of climate change on Australian streamflow. *Hydrol. Processes* **16**, 1235–1245.

Cubasch, U., Meehl, G. A., Boer, G., J., Stouffer, R. J., Dix, M., Noda, A., Senior, C. A., Raper, S. & Yap, K. S. (2001) Projections of future climate change. In: *Climate Change, 2001, The Scientific Basis* (Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change, ed. by J. T. Houghton, Y. Ding, D. J. Griggs, M. Noger, P. J. van der Linden, X. Dai, K. Maskell & C. A. Johnson), 525–582. Cambridge University Press, Cambridge, UK.

Houghton, J. T., Ding, Y., Griggs, D. J., Noger, M., van der Linden, P. J., Dai, X., Maskell, K. & Johnson, C. A. (eds) *Climate Change, 2001: The Scientific Basis* (Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change). Cambridge University Press, Cambridge, UK.

McAvaney, B. J., Covey, C., Joussaume, S., Kattsov, V., Kitoh, A., Ogana, W., Pitman, A. J., Weaver, A., Wood, R. A. & Zhao, Z.-C. (2001) Model evaluation. In: *Climate Change, 2001, The Scientific Basis* (Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change; ed. by J. T. Houghton, Y. Ding, D. J. Griggs, M. Noger, P. J. van der Linden, X. Dai, K. Maskell & C. A. Johnson), 471–524. Cambridge University Press, Cambridge, UK.

McGuffie, K. & Henderson-Sellers, A. (eds) (1997) *A Climate Modelling Primer* (second edn). John Wiley and Sons Ltd, Chichester, UK.

Nijssen, B., *et al.* (2003) Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2(e). 2: Comparison of model results with observations. *Global Planet. Change* **38**, 31–53.

Nohara, D., Kitoch, A, Hosaka, M. & Oki, T. (2006) Impact of climate change on river discharge projected by multimodel ensemble. *J. Hydromet.* **7**, 1076–1089.

Perkins, S. E., Pitman, A. J., Holbrook, N. J. & McAneney, J. (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions. *J. Climate* (accepted).

Pitman, A. J., Henderson-Sellers, A. & Yang, Y.-Z. (1990) Sensitivity of the land surface to sub-grid scale precipitation in AGCMs. *Nature* **346**, 734.

Schlosser, C. A., *et al.* (2000) Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). *Mon. Weath. Rev.* **128**, 301–321.

Seckler, D., Barker, R. & Amarasinghe, U. (1999) Water scarcity in the twenty-first century. *Water Resour. Dev.* **15**, 29–42.

Sun, Y., Solomon, S., Dai, A. & Portmann, R. W. (2006) How often does it rain? *J. Climate* **19**, 916–934.

Vörösmarty, C. J., Green, P., Salisbury, J. & Lammers, R. B. (2000) Global water resources: vulnerability from climate change and population growth. *Science* **289**, 284–288.

Wood, E. F., *et al.* (1998) The project for intercomparison of land-surface parameterisation schemes (PILPS) Phase 2(c) Red-Arkansas River basin experiment: 1. Experiment description and summary intercomparison. *Global Planet. Change* **19**, 115–135.