Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions

S. E. PERKINS

Department of Physical Geography, Macquarie University, Sydney, Australia

A. J. PITMAN

Climate Change Research Centre, University of New South Wales, Sydney, Australia

N. J. HOLBROOK

Department of Physical Geography, Macquarie University, Sydney, Australia

J. MCANENEY

Risk Frontiers Research Centre, Macquarie University, Sydney, Australia

(Manuscript received 31 August 2006, in final form 25 January 2007)

ABSTRACT

The coupled climate models used in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change are evaluated. The evaluation is focused on 12 regions of Australia for the daily simulation of precipitation, minimum temperature, and maximum temperature. The evaluation is based on probability density functions and a simple quantitative measure of how well each climate model can capture the observed probability density functions for each variable and each region is introduced. Across all three variables, the coupled climate models perform better than expected. Precipitation is simulated reasonably by most and very well by a small number of models, although the problem with excessive drizzle is apparent in most models. Averaged over Australia, 3 of the 14 climate models capture more than 80% of the observed probability density functions for precipitation. Minimum temperature is simulated well, with 10 of the 13 climate models capturing more than 80% of the observed probability density functions. Maximum temperature is also reasonably simulated with 6 of 10 climate models capturing more than 80% of the observed probability density functions. An overall ranking of the climate models, for each of precipitation, maximum, and minimum temperatures, and averaged over these three variables, is presented. Those climate models that are skillful over Australia are identified, providing guidance on those climate models that should be used in impacts assessments where those impacts are based on precipitation or temperature. These results have no bearing on how well these models work elsewhere, but the methodology is potentially useful in assessing which of the many climate models should be used by impacts groups.

1. Introduction

Coupled 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 subcontinental 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 of some variables, past climates, specific perturbations such as volcanic activity, and some key phenomenon (e.g., El Niño–Southern Oscillation, monsoons, and other specific modes of variability).

Corresponding author address: Prof. A. J. Pitman, Climate Change Research Centre, Red Centre Building, University of New South Wales, Sydney NSW, 2052, Australia.
E-mail: a.pitman@unsw.edu.au

DOI: 10.1175/JCLI4253.1

McAvaney et al. (2001) also evaluated the capacity of climate models to simulate extremes including temperature, rainfall, storms, and cyclones. Overall, their conclusion that coupled climate models were useful tools for projecting future climate was rigorously supported by the existing literature.

The latest model evaluations by Collins et al. (2006), Johns et al. (2006), and Delworth et al. (2006) provide detailed assessments of the strengths and weaknesses of three major climate models based principally on seasonal and annual time scales. Some attempts to provide measures of overall climate model skill have been included in recent model evaluation studies. Johns et al. (2006), for example, used a simple weighted nondimensional index of root-mean-square errors compared to present-day climatological means (based on Murphy et al. 2004). Monthly, seasonal, and annual data were used for a range of simulated quantities, and a statistical skill metric, the "Climate Prediction Index," was presented. Other measures of skill have been suggested by Watterson (1996), Taylor (2001), Knutti et al. (2006), Piani et al. (2005), and Shukla et al. (2006) but all, when implemented, use monthly to annual time-scale data, sometimes over ensemble means of climate models with several realizations. While McAvaney et al. (2001) evaluated climate models on a range of time and spatial scales, they were hampered by a lack of literature on the ability of climate models to simulate climate variables on daily time scales. Given that climate on time scales of days has a direct impact on human health (Trigo et al. 2005) and human activities (e.g., agriculture; Luo et al. 2005), an assessment of the capacity of models to simulate climate on time scales of days is clearly valuable. This study attempts such an evaluation for the AR4 models, thereby supplementing the many assessments that evaluate monthly means, variances, as well as specific phenomenon (see McAvaney et al. 2001).

The importance of examining climate statistics other than climate means is not new (Katz and Brown 1992; Boer and Lambert 2001). For example, recent studies by Frich et al. (2002), Kiktev et al. (2003), and Alexander et al. (2006) used climate indices and probability density functions (PDFs) of indices to explore the frequency and severity of climate extremes. These indices-based analyses provide a clear way forward in using climate model results for society-relevant impact assessments. Dessai et al. (2005) calculated PDFs, using seasonal data from climate models, to assess uncertainty in regional climate change projections. They devised a skill score that accounted for model bias and spatial variation to compare models and data for surface air temperature and precipitation.

However, few analyses have directly evaluated the ability of climate models to simulate more than the mean and standard deviation on time scales of days. One of the few reported was conducted by Sun et al. (2006) who used daily data to evaluate how well climate models could simulate precipitation. Monthly, seasonal, or longer averages can hide biases or systematic errors that are identifiable in daily data. Further, the use of statistics like means and standard deviations do not allow for a comparison of the entire data distribution. Indeed, a "good" simulation of the mean does not ensure that other attributes of the data will be captured well (Kharin and Zwiers 2000; Zwiers and Zhang 2003; Schaeffer et al. 2005; Kharin et al. 2005). Given that changes in parts of the simulated distribution other than the mean are likely to affect humans (e.g., the tails), natural ecosystems, agricultural crops, etc., more than changes in the mean (Katz and Brown 1992; Colombo et al. 1999; Easterling et al. 2000), and given there is evidence that extremes may change more than indicated by a change in the mean (Mearns et al. 1984; Schaeffer et al. 2005; Trigo et al. 2005), an evaluation of how well climate models can simulate entire distributions of a simulated variable is clearly warranted.

There is at least one major advantage of evaluating a climate model based on PDFs. If a climate model can simulate an entire PDF, this demonstrates a capability to simulate values that are currently rare and that may become more common in the future. If the distribution of values represented by a PDF shifts due to climate change, it is likely that significant overlap between the new distribution and the present distribution will remain. If a climate model that has been shown to already simulate this region of the distribution that currently exists and will remain in the future (but becomes more likely to occur probabilistically), then we have identified that the model has skill in simulating these future values. Clearly, this confidence declines as the overlap between the present and future PDFs is reduced further into the future. Until the overlap becomes critically small, however, an impacts modeler could use how well a model simulated the whole PDF of a set of variables as criteria for those models to use in future impacts assessments. Further, establishing the skill of a climate model to simulate whole PDFs is a far harder test of a model than (say) the mean and one standard deviation, and thus by succeeding in such a test, we might have more confidence in projections made with this model. We do not claim that a PDF-based assessment of a model is perfect of course. While performing well in a PDF-based assessment is substantially harder than reproducing the mean, key areas of model perfor-

TABLE 1. All climate models with daily data for T_{MAX}, T_{MIN}, and P 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, columns 3, 4, and 5 denote how many realizations from each model could be used, and column 5 is the source of the model (see http://www-pcmdi.llnl.gov/ipvv/about_ipcc.php).

Acronym	Model	T_{MAX}	T_{MIN}	P	Source
BCCR	bccr_bccm2_0	—	1	1	Bjerknes Centre for Climate Research, University of Bergen, Norway
CGCM-h	cccma_cgcm3_1_t63	1	1	0	Canadian Centre for Climate Modeling and Analysis
CGCM-l	cccma_cgcm3_1_t47	4	5	4	Canadian Centre for Climate Modeling and Analysis
CSIRO	csiro_mk3_0	2	3	3	Australian Commonwealth Scientific Industrial and Research Organisation
GFDL2.0	gfdl_cm2_0	—	—	1	Geophysical Fluid Dynamics Laboratory
GFDL2.1	gfdl_cm2_1	—	1	1	Geophysical Fluid Dynamics Laboratory
GISSAOM	giss_aom	1	1	1	NASA Goddard Institute of Space Studies
GISS ER	giss_model_e_r	—	1	1	NASA Goddard Institute of Space Studies
FGOALS	iap_fgoals1_o_g	2	1	3	Institute of Atmospheric Physics, Chinese Academy of Sciences
IPSL	ipsl_cm4	2	2	2	Insitut Pierre-Simon Laplace
MIROC-h	miroc3_2_hires	1	—	—	Centre for Climate System Research, University of Tokyo; National Institute for Environmental Studies; Frontier Research Centre for Global Change
MIROC-m	miroc3_2_medres	2	1	3	Centre for Climate System Research, University of Tokyo; National Institute for Environmental Studies; Frontier Research Centre for Global Change
ECHO-G	miub_echo_g	3	2	3	Meteorological Institute of the University of Bonn
ECHAM	mpi_echam5	—	1	1	Max-Planck-Institut für Meteorologie
MRI	mri_cgcm2_3_2a	3	2	5	Japan Meteorological Agency
CCSM	ncar_ccsm3	—	—	6	National Center for Atmospheric Research

mance such as periods of sustained high temperature or rainfall represented by indices (e.g., Kiktev et al. 2003; Alexander et al. 2006) are not assessed. Further, as an event becomes rarer in both the model and the observed data, failure of the model to simulate these events becomes less important to the skill score. While these represent limits to our methodology, we are confident that a PDF-based evaluation of a climate model is substantially preferable to a mean-based assessment and could simply replace the traditional reliance on evaluating the mean performance.

Thus, the objective for this study is the evaluation of climate models' simulation of daily observations using the original climate model results as the basis for analysis, rather than indices. One obvious requirement is high-quality observed data at suitable spatial and temporal resolution (Peterson et al. 1998; Griffiths et al. 2005). Suitable global-scale datasets remain rare due to gaps in spatial coverage (Kiktev et al. 2003). This provides a rationale for climate model evaluation at continental scales: our study focuses on Australia but we provide a methodology that should be useful elsewhere.

This paper explores the capacity of a large sample of climate models to simulate the PDFs of precipitation (P), minimum temperature (T_{MIN}), and maximum temperature (T_{MAX}). The choice of these three variables was based on available data and on their role in driving many human, industrial, and biological systems (Colombo et al. 1999; Meehl et al. 2000; Christensen and Christensen 2003; Trigo et al. 2005). We utilize the model results submitted to the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory in the United States (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php) as part of the Fourth Assessment Report conducted by the IPCC (AR4).

We have chosen to identify individual models in this paper for two reasons. First, these data are currently being widely used and unless we identify models, users of simulations cannot determine those models with particular strengths or weaknesses. Second, unless model groups know their model performs well/poorly, they cannot build on strengths or address weaknesses in subsequent model development. However, we wish to emphasize that a model that shows skill/weakness over Australia may/may not show comparable skill/weakness in other regions and each user should evaluate the models they choose for their specific region of interest.

In evaluating simulations of PDFs over Australia, we introduce a model metric that is one measure of climate model skill in terms of P, T_{MIN}, and T_{MAX}. Like all existing metrics, this only assesses a fraction of the quantities simulated by a climate model. It cannot be used to determine the "best" model overall, as "best" is dependent on the specific application for which the model will be used.

The remainder of this paper explains our data and analysis methods (section 2). This is followed by an examination of the models' PDF-based performance (section 3). A discussion is presented in section 4 and conclusions in section 5.

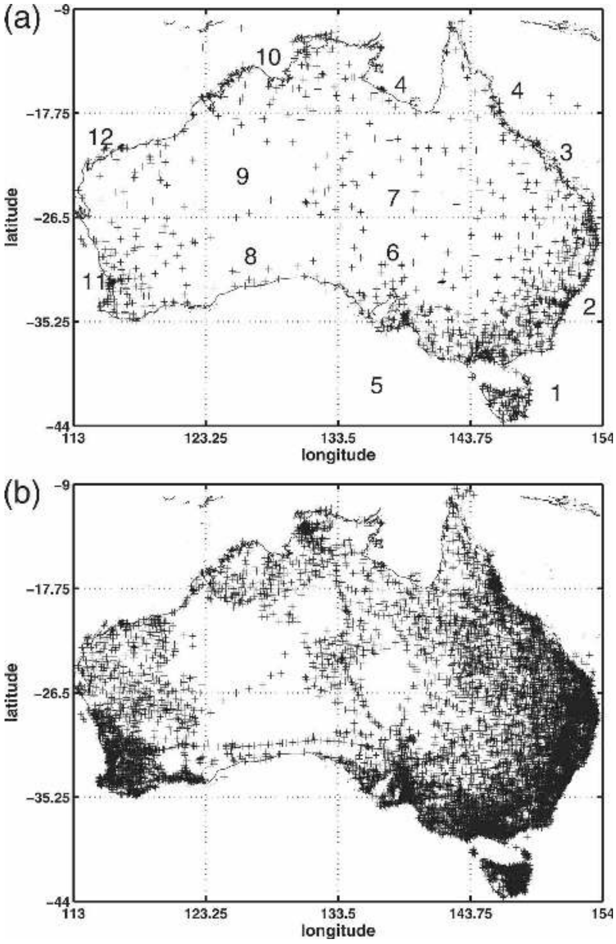

FIG. 1. (a) Locations of observed minimum and maximum temperature stations over Australia with data available between 1961 and 2000; (b) same as in (a), but for precipitation.

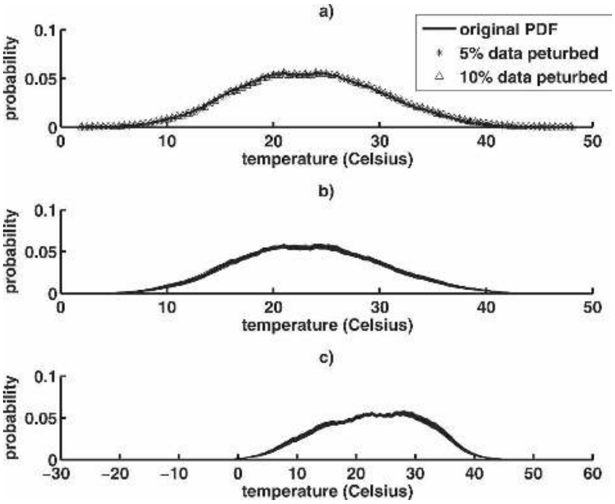

FIG. 2. PDF of observed maximum temperature for (a) unperturbed values, then 5% of values increased 5%, then 10% of values increased 10%. Note that the maximum difference in the PDFs is less than 0.01, between the original and 10% perturbed at 20°C. This PDF is for region 2 but is typical of all regions; (b) 100 PDFs calculated by randomly sampling 75% of the observed stations; (c) same as in (b), but for observed minimum temperature.

2. Data and methods

a. Climate model data

Daily climate model data over Australia for P, T_{MIN}, and T_{MAX} 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. Some datasets contained erroneous data (gaps, periods of repetitive data), or data were not available at the time this study was undertaken and were therefore omitted from subsequent analysis. Table 1 lists all models used, whether daily data were available for a given variable, and if so, the number of independent realizations for each variable. We use each independent realization directly in the initial analysis rather than average these realizations to produce an ensemble result. However, we present ensembles over the available realizations for each climate model in the results for ease of presentation. We found that differences between realizations from a single climate model in the simulated PDFs were negligible. Because of the problem of missing data from some models for some variables, a total of 14 models and 35 individual realizations were applicable for P, 13 models and 23 realizations for T_{MIN}, and 10 models and 21 realizations for T_{MAX}. Model-specific masks were fitted to exclude ocean data.

b. Observed data

Daily observed P, T_{MIN}, and T_{MAX} were obtained from the Australian Bureau of Meteorology (BOM) for the period 1961–2000. A total of 12 525 precipitation and 1529 temperature 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. Figures 1a and 1b show the locations of the temperature and rainfall stations used in this study.

Homogenization and quality control of observed data are common problems in model evaluation (Kiktev et al. 2003; Griffiths et al. 2005; Alexander et al. 2006). Quality control of observed data is vital when means or standard deviations are calculated, since 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

TABLE 2. Latitude and longitude boundaries for all 9.75° × 10.75° regions over Australia, with climate type based on the Köppen classification scheme derived by Australian Bureau of Meteorology.

Region	Lat	Lon	Climate
1	35.25°–44°S	143.75°–154°E	Temperate
2	26.5°–35.25°S	143.75°–154°E	Desert/grassland/temperate
3	17.75°–26.5°S	143.75°–154°E	Desert/grassland/subtropical
4	9°–17.75°S	133.5°–154°E	Tropical/equatorial
5	35.25°–44°S	133.5°–143.75°E	Temperate/grassland
6	26.5°–35.25°S	133.5°–143.75°E	Grassland/desert
7	17.75°–26.5°S	133.5°–143.75°E	Grassland/desert
8	26.5°–35.25°S	133.5°–123.25°E	Grassland/desert
9	17.75°–26.5°S	133.5°–123.25°E	Grassland/desert
10	9°–17.75°S	133.5°–123.25°E	Grassland/tropical
11	26.5°–35.25°S	113°–123.25°E	Temperate/grassland/desert
12	17.75°–26.5°S	113°–123.25°E	Grassland/desert

because they allow a more complete assessment of a climate model's capacity to simulate the complete range of observations at daily time scales.

Figure 2a presents three PDFs for a 10° × 10° region of Australia to illustrate the relative insensitivity of a PDF to errors in the observations (compared to means or standard deviations). First, the original observed data (solid line) has a mean of 23.34°C and a standard deviation of 6.67°C. We then perturb the original dataset by increasing 5% of the values by 5% to obtain the second PDF and to give a revised mean of 23.52°C and standard deviation of 6.77°C. Finally, we perturb the original dataset by increasing 10% of the values by 10% to obtain the third PDF, giving a revised mean of 23.57°C and a standard deviation of 6.75°C. While both the mean and standard deviation change in each case, Fig. 2a shows little change in the shape of the resulting PDFs. We will show later that differences between observed and simulated PDFs are commonly too large to be explained by observational error. A second advantage of using PDFs is that we can safely merge data from multiple stations where the data lengths are different and/or a station samples a small amount of a total time series. We can thus use gap-affected observational data with relative ease provided we assume these gaps (in time) are random in terms of their likelihood of contributing to a particular part of the PDF.

In terms of spatial coverage, Fig. 1 indicates that data coverage is biased toward the coast, in particular in southeastern and southwestern Australia. However, excluding regions 8 and 9, data coverage is quite complete with stations widely distributed even in the sparsely populated areas of the continent. In regions 8 and 9, in particular for precipitation, data coverage is clearly incomplete and/or spatially biased. Overall, however, by using a PDF-based analysis that allows all stations to be used to estimate the probability of a given temperature or precipitation event (as distinct from the mean), our comparison of modeled and observed temperature and precipitation can be based on a more data-rich foundation.

To explore the sensitivity of the observed PDFs to the sampling the observed stations a sensitivity study was conducted. Figure 2b (maximum temperature) and Fig. 2c (minimum temperature) each represent 100 PDFs obtained by sampling the 75% of the individual stations. There is a 0.97 overlap (where 1.0 represents a perfect overlap) between the PDF calculated using all observed data and the most dissimilar PDF calculated using 75% of the data for both the maximum and minimum temperatures (see section 2d for an explanation of

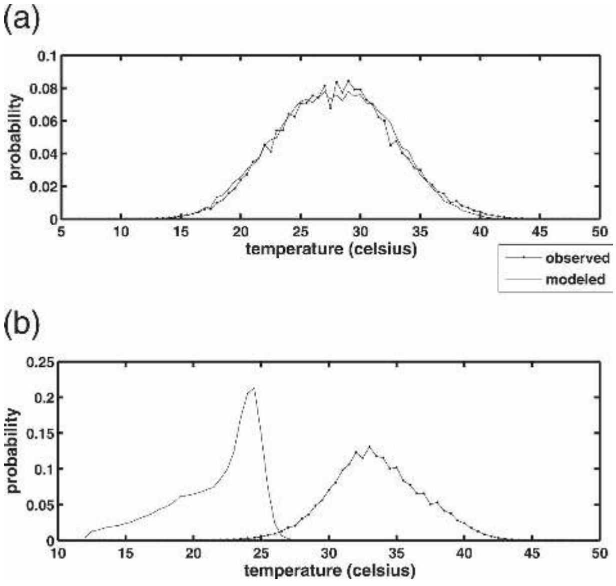

FIG. 3. Diagrams of modeled vs observed PDF illustrating the total skill score in (a) a near-perfect skill score test (0.9) and (b) a very poor skill score (0.02).

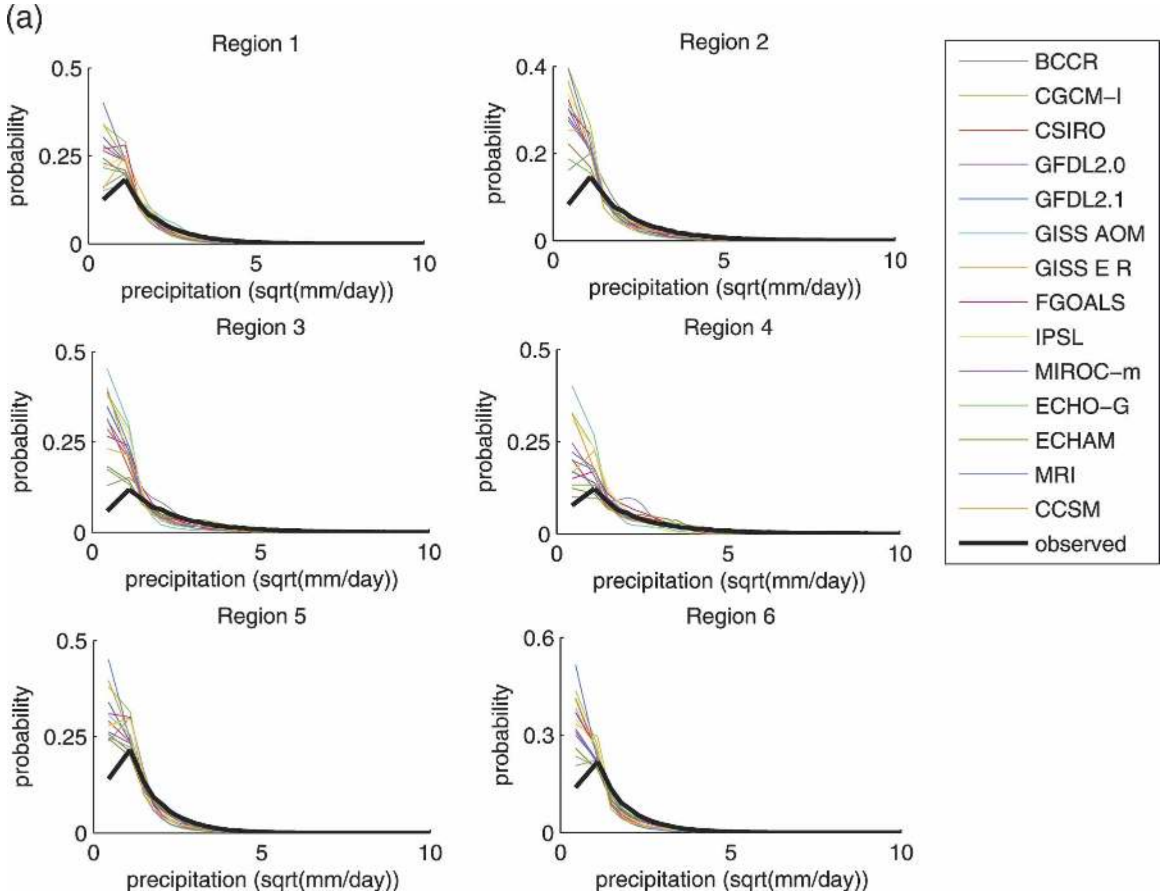

FIG. 4. (a) Probability density functions for precipitation for regions 1–6, defined in Table 2. The observed PDF has been smoothed. (b) Probability density functions for precipitation for regions 7–12, defined in Table 2. The observed PDF has been smoothed.

this overlap statistic). This provides an indication of how robust the observed PDFs are to the spatial sampling of the observed data and provide some measure of how similar a climate model would need to simulate a PDF to be indistinguishable from the observed. We have not attempted to sample the observed stations to achieve a spatially uniform coverage because this leads to the inclusion of individual stations of very different data lengths and the omission of many stations that contain useful samples of data.

c. Gridding and calculation of PDFs

Before evaluating the climate models, Australia was divided into 13 rectangles, each of 9.75° × 10.75°. The two eastern tropical regions were combined, as some climate model grids did not resolve both land areas (both shown as region 4 in Fig. 1a). The final 12 regions and their main climate types are shown in Table 2 (see also Fig. 1a). An advantage of focusing on Australia is that several climate types are represented ranging from tropical to desert environments. This division of the continent into ~10° × 10° regions permits the evaluation of the climate models across various climate types (Table 2), based on a large amount of observational data in most regions (Fig. 1).

In calculating the PDFs, all observed data within each ~10° × 10° region were used to construct the representative distribution. Similarly, for each model, all data contained in the realizations available for a model were used to derive the PDF. We did not average across realizations, nor did we average across individual grid squares contained within a given ~10° × 10° region. The use of the daily data, each grid square, and each realization provides a large sample for calculating the PDF. Further, the use of ~10° × 10° regions also meant that each included at least four climate model grid cells. Our analyses therefore do not attempt to evaluate individual climate model grid elements.

Using MatLab (http://www.mathworks.com), PDFs were calculated for each ~10° × 10° square for P, T_{MIN}, and $T_{\mathrm{MAX.}}$ Observed and model data were binned around centers determined by the range of the ob-

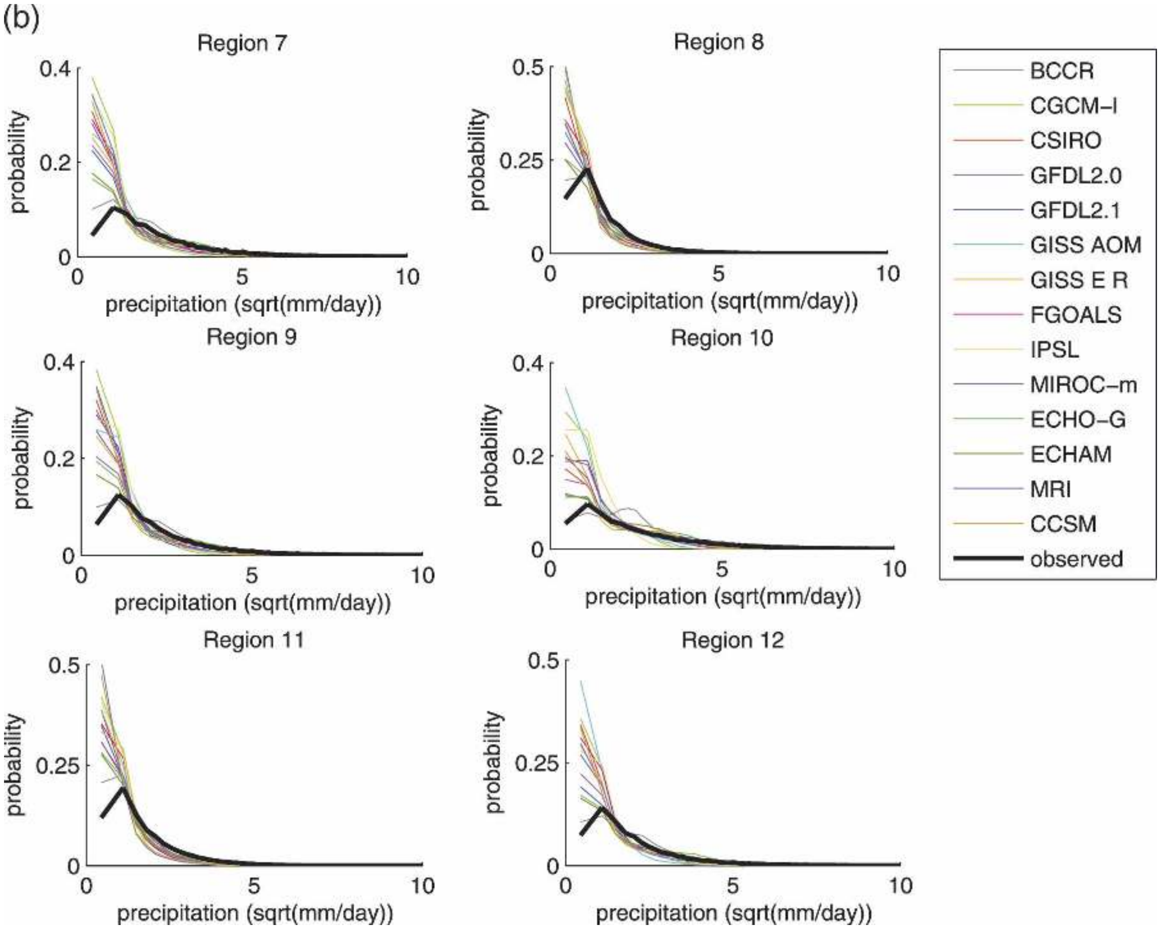

FIG. 4. (*Continued*)

served data for the variable in question, unique to each region. Bin sizes of 0.5°C for T_{MIN} and T_{MAX}, and 1 mm day^{-1} for P were used. All daily values of P below 0.2 mm day^{-1} were omitted because rates below this amount are not recorded in the observations (Parkinson 1986). While Dai (2001) and Sun et al. (2006) noted that precipitation rates below 1 mm day^{-1} contribute little to total daily precipitation over most regions, we used data above 0.2 because over Australia these amounts may be important in some arid regions. 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 conclusions.

An issue with comparing climate model precipitation with observations is whether the generated precipitation is a point or area estimate (Skelly and Henderson-Sellers 1996; Osborn and Hulme 1997; Osborn and Hulme 1998). Osborn and Hulme (1997) introduce a method to avoid biasing a mean-based comparison between simulated and observed precipitation. However, given the large number of observational data used in this paper, and the coverage of stations within each 10° × 10° rectangle, we have not implemented this methodology. Indeed, since our focus is on probability density functions rather than a direct comparison of means we suspect that station coverage within our rectangles would not introduce biases.

d. Skill score

We explored how to create a summed metric using a variety of established statistical tests (e.g., Kolomogorov–Smirnov, Kendall's tau, Cramer von Mises, and Mann–Whiteney). However, it is not clear how to sum across these PDF-based statistics as each test overlaps to some degree in the sense that they examine similar parts of the model-observed statistical space. It is therefore not clear whether various statistical scores should be weighted in some way if they are to be combined.

We therefore explored an alternative metric that appears to be a very simple but very useful measure of

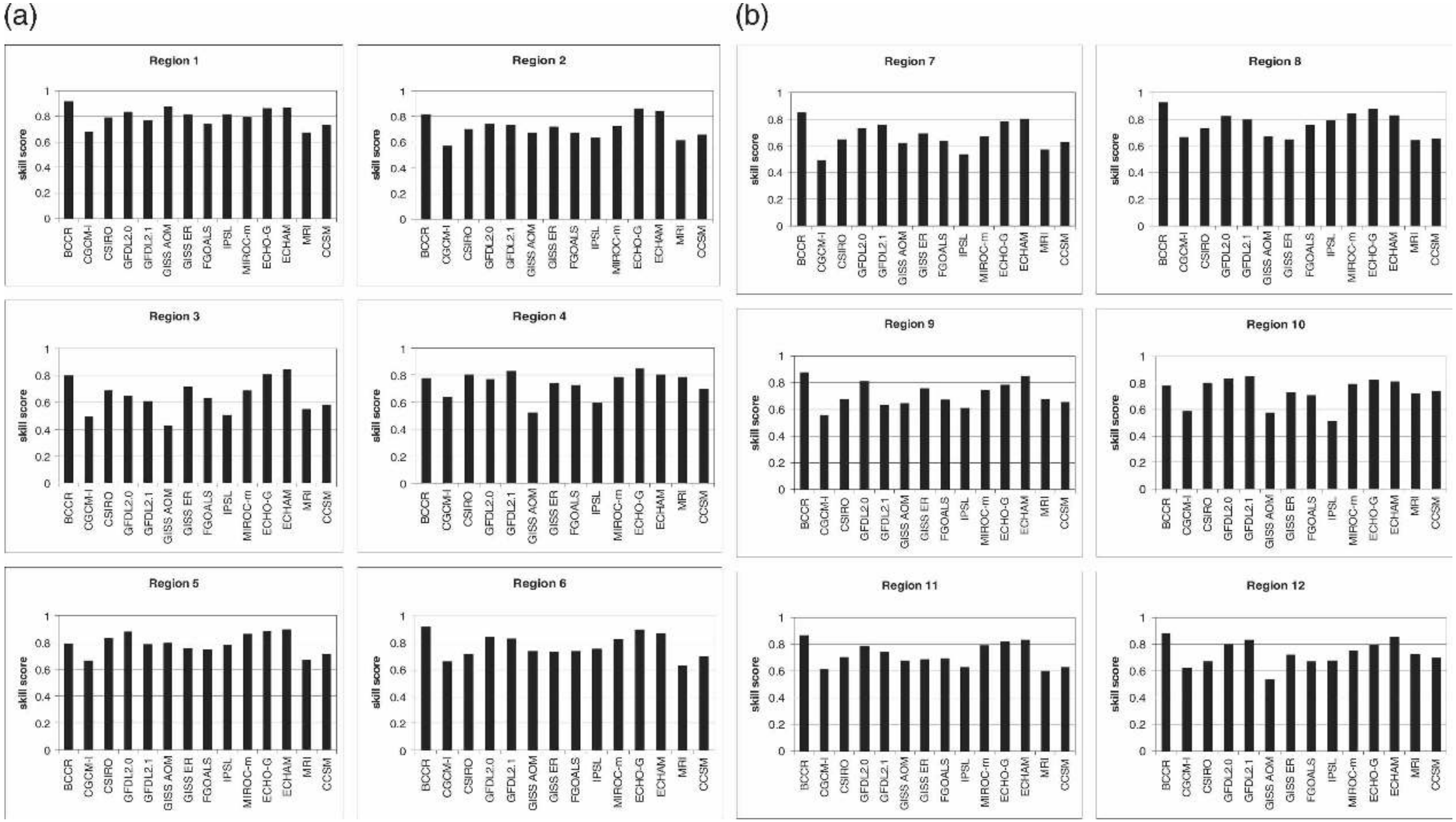

FIG. 5. (a) PDF-based skill scores for precipitation for regions 1–6, defined in Table 2. A perfect score is 1.0; (b) PDF-based skill scores for precipitation for regions 7–12, defined in Table 2. A perfect score is 1.0.

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 center in a given PDF (see Fig. 3a). Expressed formally,

$$S_{\mathrm{score}} = \sum_{1}^{n} \mathrm{minimum}(Z_m, Z_o),$$

where n is the number of bins used to calculate the PDF for a given region, Z_m is the frequency of values in a given bin from the model, and Z_o is the frequency of values in a given bin from the observed data. If a model simulates the observed PDF poorly, it will have a skill score close to zero with negligible overlap between the observed and modeled PDFs (see Fig. 3b). 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.

3. Analysis of model-simulated PDFs

a. Precipitation (P)

Figure 4 shows the PDFs for precipitation for each 10° × 10° region (Fig. 1a; Table 2). We show the PDFs for each model and each region in this section to avoid selectivity. We calculated the PDFs using all independent realizations for a given model as these were virtually indistinguishable in the figures. The x axes in Fig. 4 uses square root of the simulated value since the majority of the simulated P occurs at rates of less than 10 mm day^{-1}. Most models tend to overestimate the amount of "drizzle" that falls in each region (Sun et al. 2006). This is clear in almost all models: the observed is at the low end of the range of simulated values at the intersection of the y axis. In many cases, models overestimate the observed probability of rainfall in the 1–2 mm day^{-1} range by 2–3 times.

It is not easy to interpret the PDFs for rainfall rates greater than about 5 mm day^{-1}, but there is evidence in Fig. 4 that the models do quite well. Figure 5 shows the skill score for P for each model and for each region. This quantification of the model PDFs, in comparison to the observed, provides an objective measure of model skill. Figure 6a shows the ensemble skill score for each model averaged over all 12 regions. Overall, the skill scores for 9 of the 14 models exceed 0.7 (Fig. 6a) and the Bergen Climate Model, version 2.0 (BCCR), ECHAM, and the Meteorological Institute of the University of Bonn ECHO-G are very close to 0.85. This strong performance is clear in most regions: ECHO-G and ECHAM exceed 0.8 in all regions (Fig. 5). The Coupled General Circulation Model, version 3.1 T47 (CGCM-l) and Community Climate System

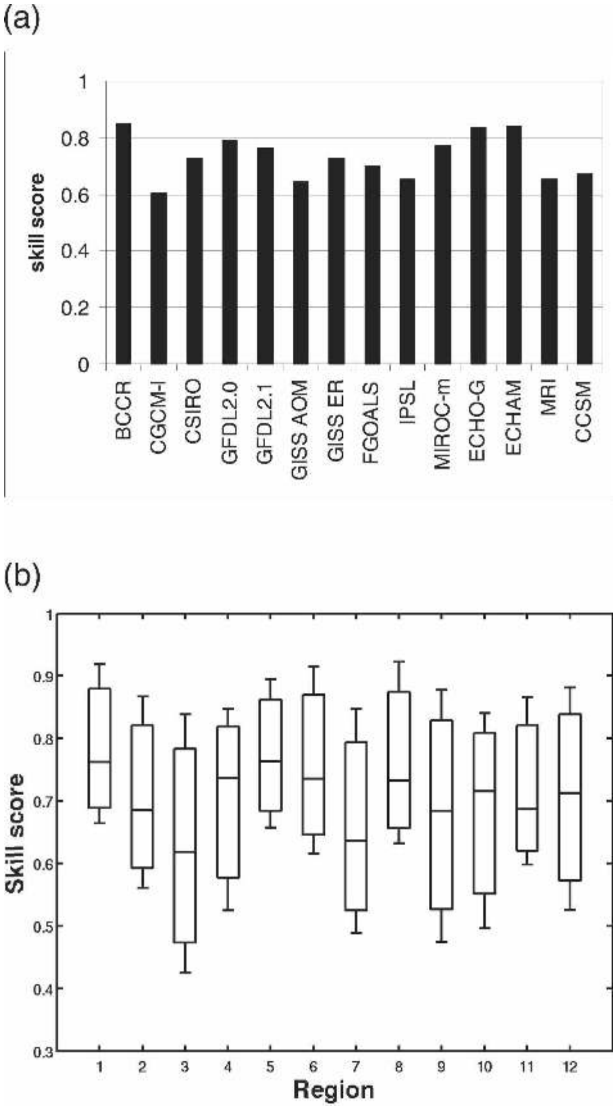

FIG. 6. (a) Ensemble PDF-based skill score for precipitation for each climate model averaged over all regions defined in Table 2; (b) box–whisker diagrams for each region shown in Table 2. The horizontal center bar in each box is the ensemble mean of all models, the upper and lower horizontal bars show the 75th and 25th percentiles, and the extremes of the bars show the maximum and minimum values.

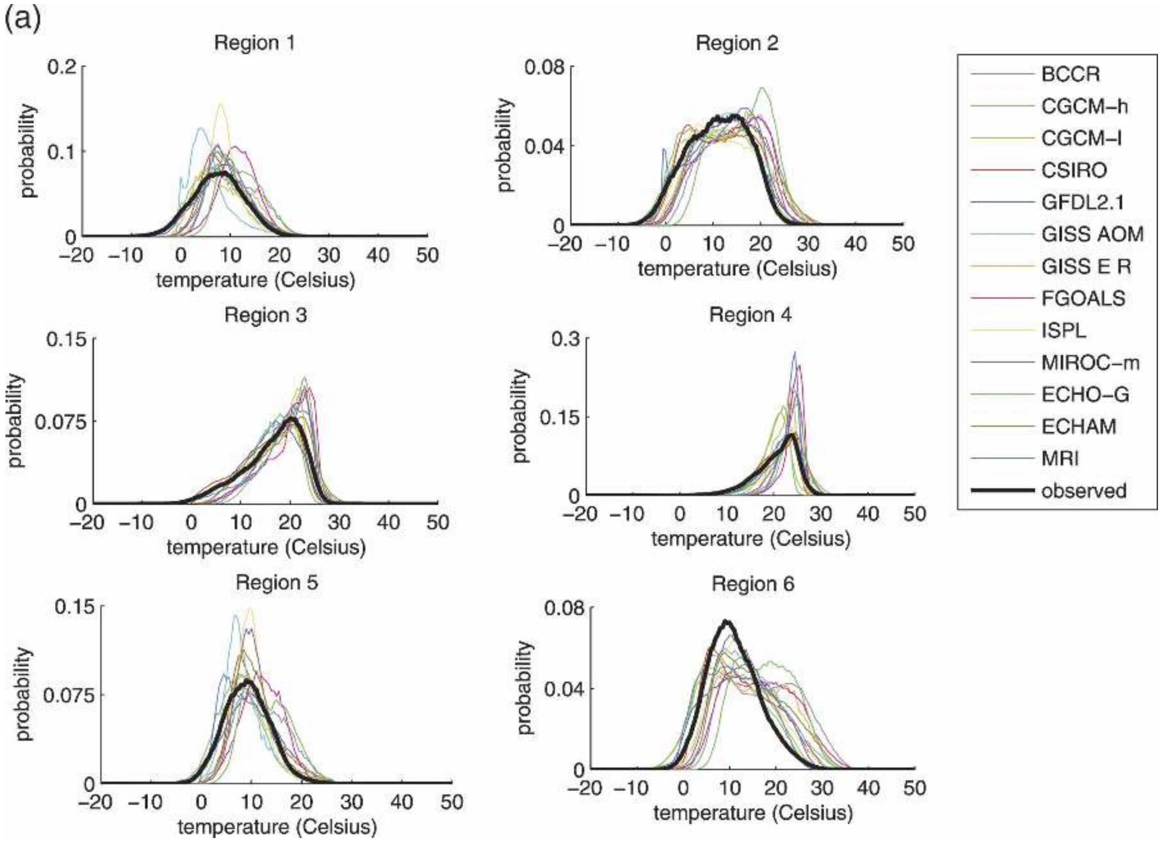

FIG. 7. (a) Same as in Fig. 4a but for minimum temperature. (b) Same as in Fig. 4b, but for minimum temperature.

Model, version 3.0 (CCSM), Meteorological Research Institute CGCM version 3.0 (MRI), Institut Pierre-Simon Laplace Climate Model, version 4 (IPSL), and the Goddard Institute for Space Studies Atmosphere–Ocean Model (GISS AOM) appear to be relatively weak with ensemble skill scores <0.7 or lower in most regions and falling to ~0.5 in some regions (Fig. 5).

Figure 6a shows that, averaged over Australia, three models score higher than 0.8 while no model scores less than 0.6, a result that hides the considerable regional variability. Figure 6b shows the ensemble average performance of all models by region (horizontal line in center of the box) as well as the maximum and minimum skill scores for the models. The worst performing areas are regions 3 and 7 with ensemble skill scores <0.7. These are the subtropical/temperate transition regions. The relatively low skill in these regions contrasts with the better performance in those regions where the rainfall is driven by either monsoons or storm tracks. However, this result is strongly affected by a few models that show very variable performance by region. If models with skill scores <0.7 are omitted, the remaining models show little variability in performance regionally. This strongly supports omitting weak models from an ensemble as these weak models strongly bias the skill of the ensemble but in regionally variable ways. Thus, the apparent overall poorer behavior in regions 3 and 7 (Fig. 6b) is, in fact, a reflection of the poor skill in CGCM-l, GISS AOM, MRI, IPSL, and CCSM.

b. Minimum temperature (T_{MIN})

The simulated and observed PDFs are shown in Fig. 7 for each region. The associated skill scores are shown in Fig. 8. No individual climate model produces systematically anomalous results. Figure 7 shows an encouraging result: at least visually the shape of the models' PDFs varies from region to region in similar ways to the observed. A very tight and pointed observed PDF (regions 4 and 10) can be contrasted with a broader PDF (regions 2, 8, and 12) and it is reassuring that the models do simulate this basic change. The overall skill in the models' simulation of T_{MIN} is >0.8 for 11 of the 13 models (Fig. 9a). Figure 7 shows that the simulated PDFs are typically quite tightly clustered around the observed PDF although there is a tendency toward a horizontal shift in the PDF in some regions (this appears unrelated to the nature of the climate in a given

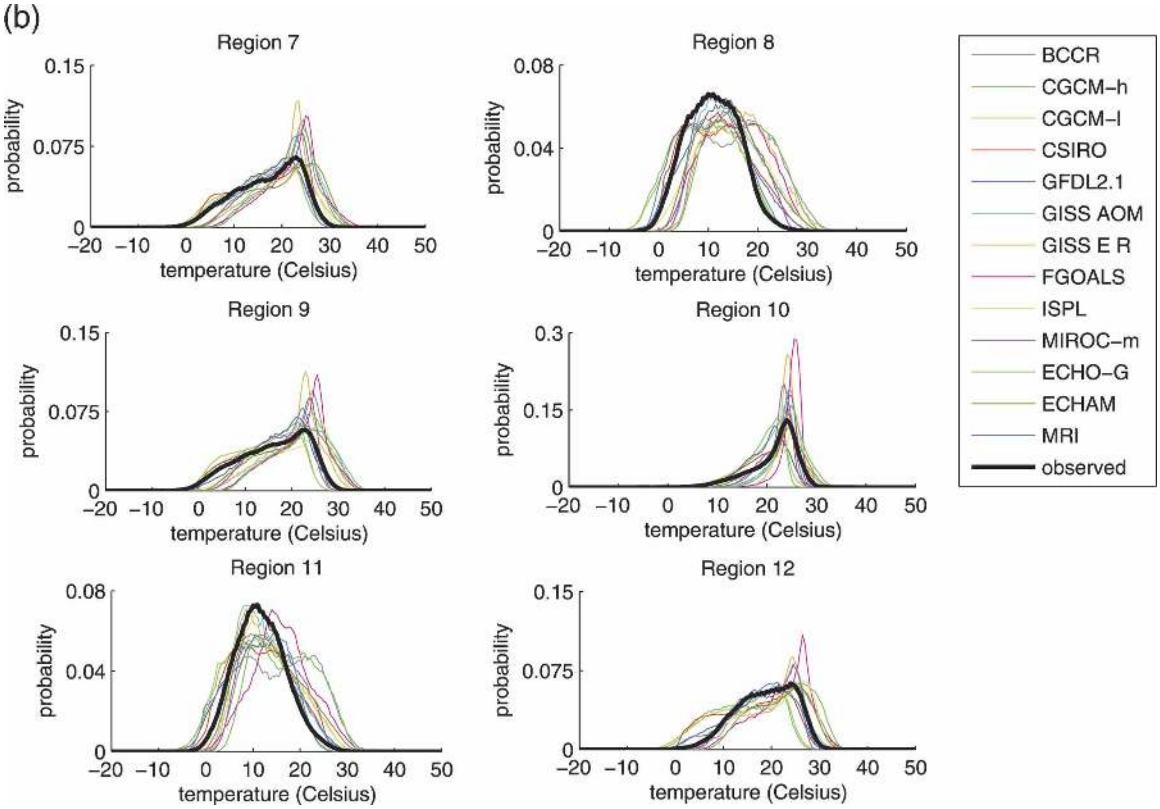

FIG. 7. (*Continued*)

region). In regions 2, 6, 8, and 11 (which are all within the same latitudinal bands) the models tend to overestimate the probability of warmer minimum temperatures but the probability of lower minimum temperatures appears to be quite well represented in most models. Of course, some models perform better than others and overall Commonwealth Scientific and Industrial Research Organisation Mark version 3.0 (CSIRO), the Geophysical Fluid Dynamics Laboratory Climate Model version 2.1 (GFDL2.1), CGCM-l, and MRI perform best (skill scores >0.85; Fig. 9a) although each model also performs relatively weak in some regions. The Flexible Global Ocean–Atmosphere–Land System Model g1.0 (FGOALS) and ECHO-G are relatively poor (skill score <0.7, Fig. 9a) and are substantially weaker (by about 0.1) than most other models.

The regional performance of the ensemble of the models is shown in Fig. 9b. Ensemble mean skill scores are between 0.75 and 0.85 with little variation between regions. However, if those weaker models (FGOALS, ECHO-G, and BCCR) are omitted, the scores range between 0.85 and 0.9 for all regions. This indicates that the better models show very similar skill, but a systematic error of ~0.1 remains in all models.

c. Maximum temperature (T_{MAX})

Figure 10 presents the PDFs of the observed and modeled T_{MAX} for each region and Fig. 11 shows the corresponding skill scores. Overall, most climate models simulate the PDF of observed T_{MAX} well. Many of the models capture the changes in location (with respect to the *x* axis) and shape of the observed PDF well between regions. There are, however, anomalies in all regions. The FGOALS model, for example, is poor in the Tropics (regions 4 and 10) but is highly competitive in many other regions. The CSIRO model grossly overestimates the probability of T_{MAX} toward the lower end of the range in regions 9 and 12. A common problem is an excessive simulation of too many low T_{MAX}s and too many high T_{MAX}s such that the observed peak in the PDF is underrepresented. This is particularly clear in regions 2, 6, 8, 11, and 12. These are the more temperate regions of Australia and likely the most difficult to model well. These are also regions where soil moisture is likely in transition between nonlimiting (to evaporation) and limiting (see Koster et al. 2004). Errors in the modeling of these processes can contribute to errors in the simulation of T_{MAX} and T_{MIN} (see section 4).

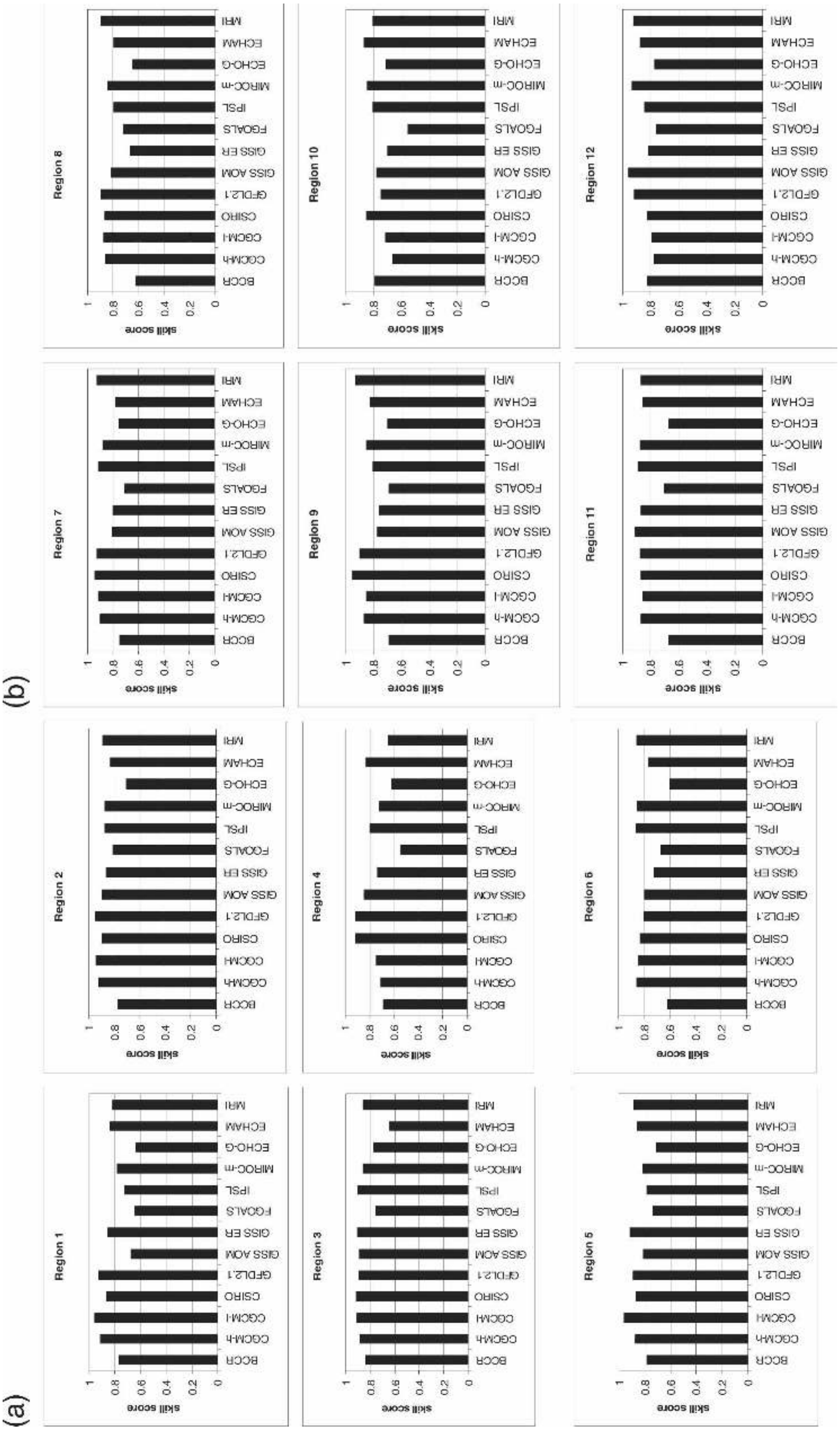

FIG. 8. (a) Same as in Fig. 5a, but for minimum temperature. (b) Same as in Fig. 5b, but for minimum temperature.

The overall performance of the models in representing T_{MAX} is similar to T_{MIN}. Figure 12a shows that 6 of 10 models generate a skill score in excess of 0.8 (in contrast to 11 of the 13 for T_{MIN}). Figure 12b shows the overall region-by-region performance of the models. The poorest simulated regions are 1, 4, and 10, which are the extreme north and south of Australia. However, as with T_{MIN}, if models with skill scores <0.8 are omitted (this excludes CGCM-l, CGCM-h, GISS AOM, and MRI), the regional performance of the remaining models is almost always between 0.85 and 0.95. Thus, the remaining systematic error of ~0.1 is common between T_{MAX} and T_{MIN}.

d. 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 figures showing the PDFs. We calculated continental-scale percentiles to give an overview of the models' capacity. This was achieved by concatenating data from all models for all regions for a given variable and sampling each dataset for the specific percentile.

Figure 13 shows the model results corresponding to each percentile for P, T_{MIN}, and T_{MAX}. Precipitation, at these higher percentiles, is poorly captured by the models. Specifically, the highest values simulated by any model for the 95th percentile is about 24 mm day^{-1} (ECHAM)—a figure close to the observed 90th percentile (~22 mm day^{-1}; see the final data point in Fig. 13a). This is ~10 mm day^{-1} lower than the observed value for the 95th percentile, highlighting that all models systematically underestimate high rainfall rates in addition to the overrepresentation of low rainfall rates. These findings confirm Sun et al.'s (2006) conclusions in a more detailed regional analysis. There are, however, two groups of models represented in Fig. 13a. BCCR, CSIRO, GFDL2.0, GFDL2.1, the Model for Interdisciplinary Research on Climate, version 3.2 T42 (MIROC-m), ECHO-G, and ECHAM are superior to the remaining models in this measure. They produce rainfall for the 95th percentile that compares to the observed 90th percentile. The remaining models' 95th percentile compares more to the observed 80th percentile. This provides one means of discriminating between models in the simulation of the rarer events that are important in impact assessments.

The results for T_{MIN} (Fig. 13b) and T_{MAX} (Fig. 13c) are quite variable. Six of the 13 models are within ±2°C of the observed for T_{MIN}. There is a general bias in the models to overestimate the higher percentiles in both T_{MIN} and T_{MAX}. BCCR is clearly relatively poor for T_{MIN} and GISS AOM, GISS ER, FGOALS, IPSL, ECHO-G, and ECHAM are all about 5°C to warm on each percentile. For T_{MAX}, the Coupled General Circulation Model version 3.1 T63 (CGCM-h), CGCM-l, GISS AOM, and MRI are excessively warm but CSIRO and FGOALS are impressively close to the observed.

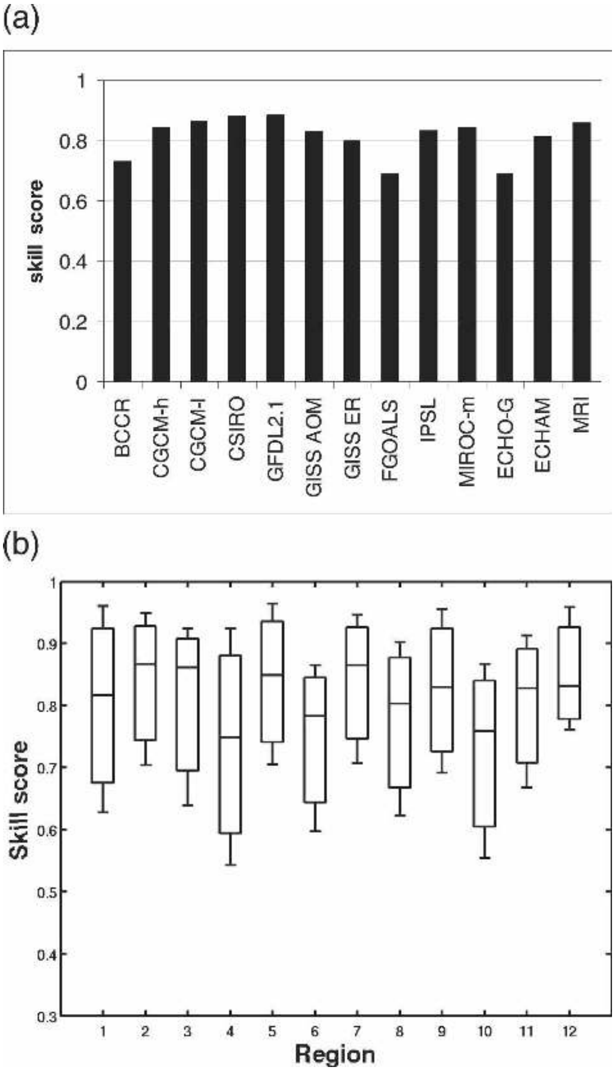

FIG. 9. (a) Same as in Fig. 6a, but for minimum temperature. (b) Same as in Fig. 6b, but for minimum temperature.

The warm bias in T_{MIN} is straightforward to explain. In many parts of Australia, minimum temperatures are orographically induced. The resolution of climate models tends to prevent local minima from being simulated, where these are due to specific topographical features. The warm bias in T_{MAX}, however, is not likely due to orography and is more likely to indicate problems in

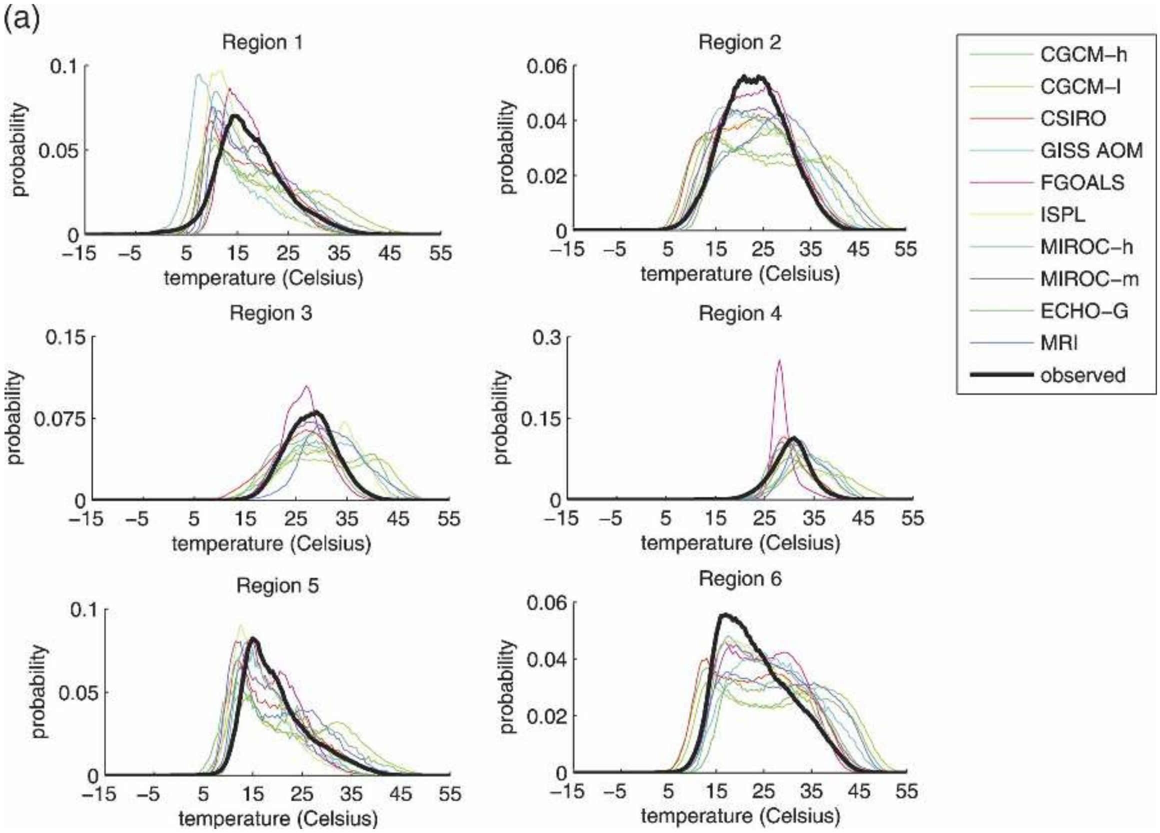

FIG. 10. (a) Same as in Fig. 4a, but for maximum temperature. (b) Same as in Fig. 4b, but for maximum temperature.

the models' simulation of net radiation and/or terrestrial processes (see section 4).

4. Discussion

The development of climate models for future projection has placed increasing demands on models to simulate the present-day climate well. It is a demanding challenge to produce global, fully coupled climate models that show genuine skill in regional climate simulations. While monthly or seasonal means from climate models have traditionally been the most common basis for analysis, they are not necessarily the most useful measures for climate impact assessment. Changes in other statistics like annual maximum temperature, annual daily maximum rainfall, and annual minimum temperature are likely to have a greater impact on a range of biophysical systems than a change in the mean (e.g., Frich et al. 2002). Evaluating indices for each of these is very useful and may underpin many impacts assessments (e.g., Alexander et al. 2006), but this paper focuses on model evaluation and presents a way to evaluate models across a variable's full range utilizing the entire PDF. A climate model with skill across a range of observed PDFs shows a capacity to simulate the full range of climates in different regions. If a climate model can accurately simulate the probability of temperatures two standard deviations from the current mean, this builds confidence that they could simulate the greater proportion of future climates, at least until temperatures rise such that the PDFs overlap with the present day little.

An evaluation of the regional PDFs of P, T_{MAX}, and T_{MIN} for each of the AR4 models show, as expected, a range of performances that were quantified via a skill score that measured the degree of overlap of the PDFs. The skill scores were aggregated by model and by region to enable a quantitative assessment of model performance and to identify regions where the models were particularly good or bad.

First, it surprised us how well most models reproduced the observed PDFs of P, T_{MIN}, and T_{MAX} for each region. It is demanding for a global fully coupled climate model to be able to capture observed regional PDFs. The skill shown by most models strongly supports previous assessments that climate models are useful tools (e.g., McAvaney et al. 2001). However, there were understandably problems with some models in some regions. In terms of rainfall, the tendency of cli-

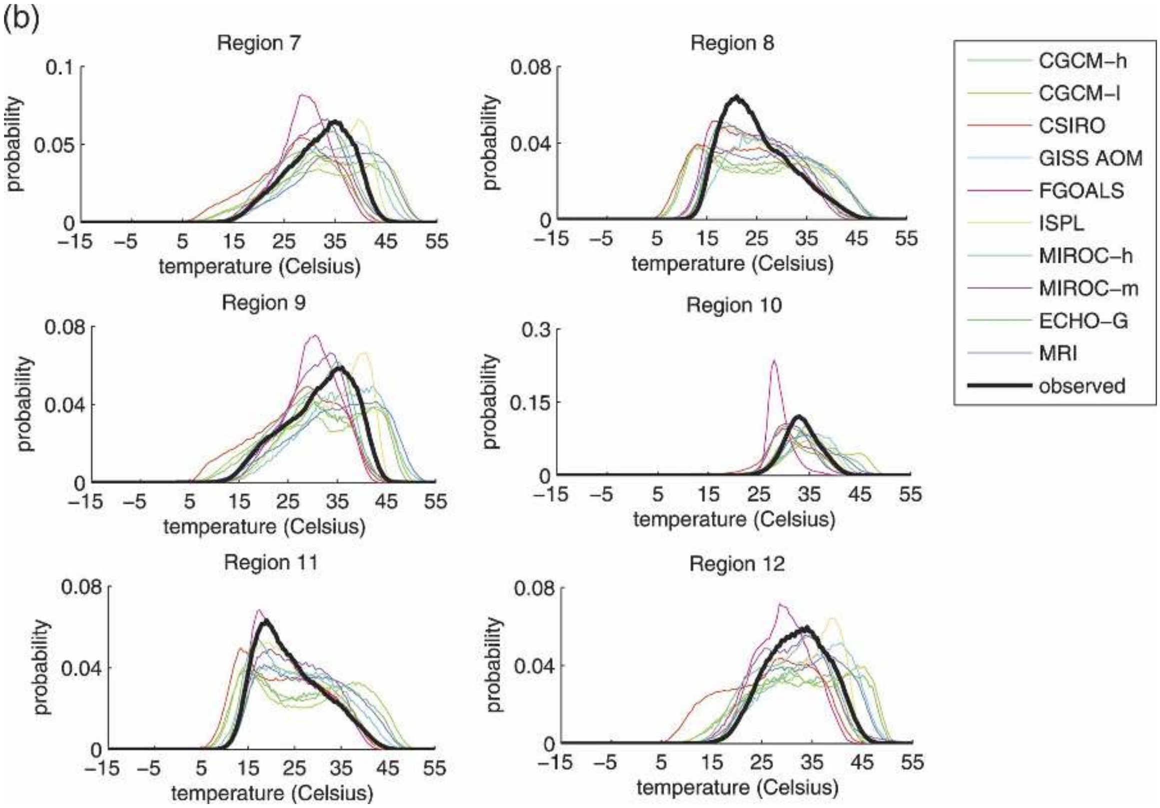

FIG. 10. (*Continued*)

mate models to simulate too much low-intensity rainfall was shown, supporting the findings of Sun et al. (2006). We also identified limits in the models' capacities to simulate the 80th, 90th, and 95th percentiles. We showed that if all models are included, there appears to be greater skill in simulating tropical and storm-track-driven climates. However, this was shown to be biased by a group of models with relatively little skill in the temperate/subtropical regions. If these weaker models were omitted, the remaining models showed high and similar skill in all regions.

One advantage of the skill score used here is that it is comparable between models. It therefore provides a basis for ranking climate models variable by variable or overall by averaging over variables. Table 3 shows, averaged over Australia, that the best model for precipitation is BCCR closely followed by ECHAM and ECHO-G. There is some sensitivity in this ranking to the choice of minimum precipitation in calculating the PDFs. If we omit values below 1 mm day^{-1} (following Sun et al. 2006), ECHAM performs best. The exact ranking is therefore affected by the selection of the minimum value for precipitation. Others using this technique should determine what constitutes a daily rainfall amount that can be ignored since this varies regionally (one would choose a different minimum in the Sahara compared to the Amazon) and depending on the intended application. However, while the exact ranking varies, it is typically by ±1 position and the methodology does not explain the poor scores obtained by some models.

In terms of T_{MIN}, the majority of models (10 of 13) had PDF-based skill scores >0.8, in comparison to 6 of 10 models for T_{MAX}; T_{MIN} is driven by radiative cooling, which in turn is associated with cloud, and atmospheric moisture content. Figure 9a implies that the suite of AR4 climate models are capturing this radiative cooling process well as it would not be possible to simulate the PDF of T_{MIN} otherwise. BCCR, FGOALS, and ECHO-G are, however, relatively weak, suggesting a problem with cloud, radiation, or atmospheric moisture content. Diagnosing the reasons for individual model performance is beyond the scope of this paper. In terms of overall ranking of the AR4 models for T_{MIN}, GFDL2.1, CSIRO, and CGCM-l are best (see Table 3) although the top 10 models vary by only 0.09, indicating that most of the AR4 models perform similarly and well. It is likely that a significant fraction of the remaining skill score error (~0.1) is related to the scale difference between the models and observed. The coarse

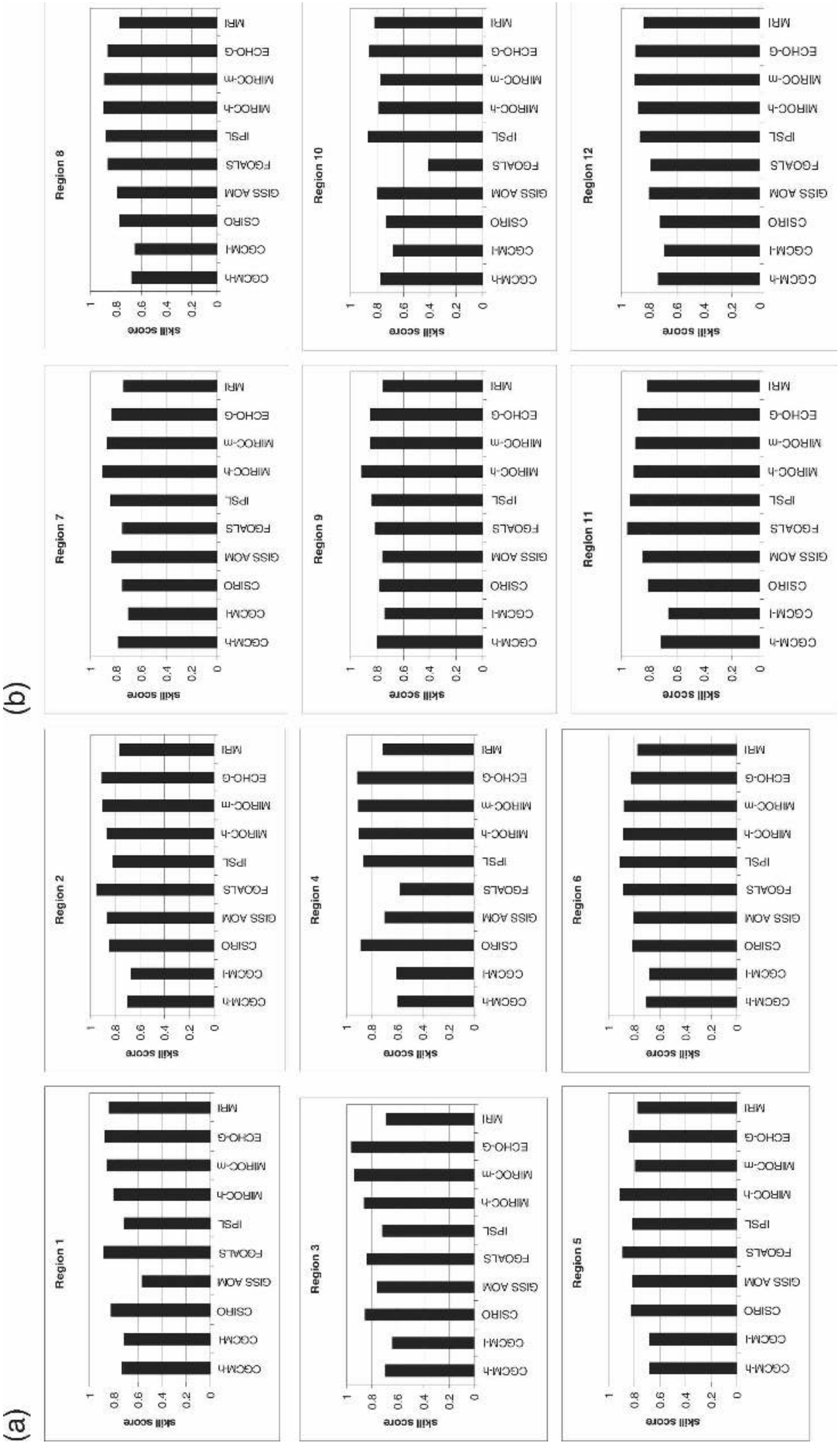

FIG. 11. (a) Same as in Fig. 5a, but for maximum temperature. (b) Same as in Fig. 5b, but for maximum temperature.

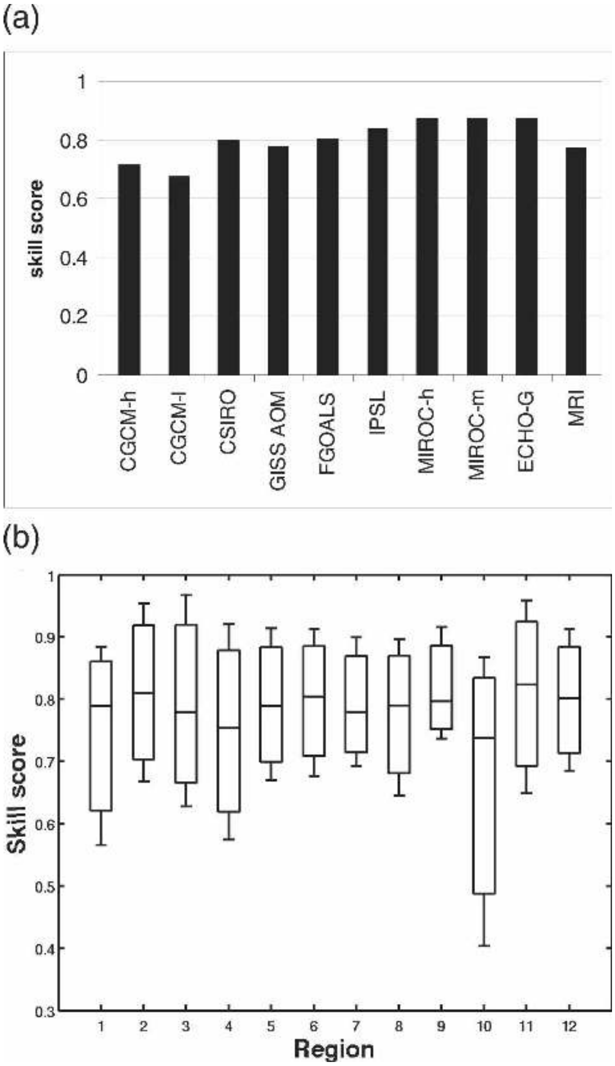

FIG. 12. (a) Same as in Fig. 6a, but for maximum temperature. (b) Same as in Fig. 6b, but for maximum temperature.

model resolution smoothes orography, thereby limiting the capacity of the models to simulate local minima.

The simulation of T_{MAX} requires a larger number of processes to be captured in a climate model. In addition to clouds interacting with incoming solar radiation and water vapor influencing net infrared radiation, T_{MAX} is affected by albedo, which directly controls absorbed solar radiation. The key problem in simulating T_{MAX} is that many models overestimate the probability of high values (see Figs. 7 and 13c). This may be related to an overestimation of net radiation in the models, although this is unlikely since Wild (2005) found climate models to underestimate surface insolation. An alternative possibility is that T_{MAX} is affected by how net radiation is partitioned between sensible and latent heat fluxes and this partitioning is controlled by land surface processes.

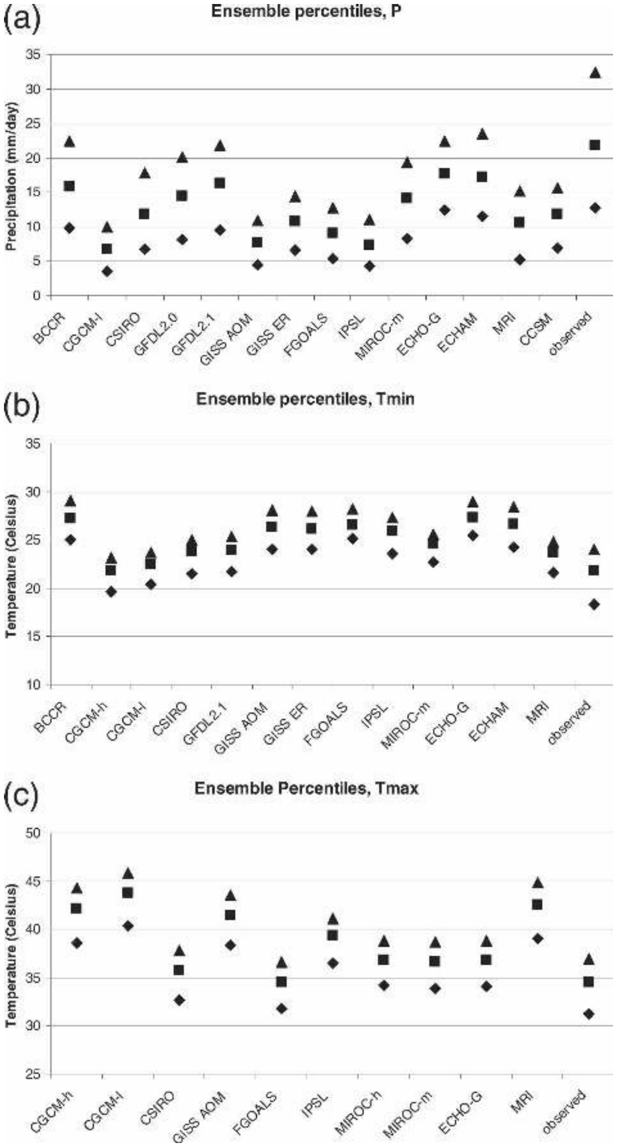

FIG. 13. Cumulative ensemble percentile values for (a) P, (b) T_{MIN}, and (c) T_{MAX}. Diamonds represent the 80th percentile, squares the 90th, and triangles the 95th. The observed is shown at the right-hand side of each figure. Note all daily values <0.2 mm day^{-1} are omitted in calculating the precipitation percentiles.

Variations in any processes that control the supply of water for evaporation (soil moisture, root distribution, stomatal conductance; Pitman 2003) can affect evaporative cooling and therefore T_{MAX} (Collatz et al. 2000). We undertook a first-order examination of climate model skill in simulating T_{MAX} as a function of the complexity of surface parameterization in the climate models. We found no simple association between land surface complexity and skill in simulating T_{MAX} but note it is difficult to determine the complexity of these land surface models implemented in the climate models

TABLE 3. Ranking of climate models for P, T_{MAX}, and T_{MIN} over all regions of Australia (an average of all 12 regions shown in Table 2). The top part of the table includes those models where all data for P, T_{MAX}, and T_{MIN} were available. Only these models are included in the final ranging shown in the right-hand column. Readers interested in the ranking by region can refer to Figs. 5, 8, and 11.

	P	Rank	T_{MAX}	Rank	T_{MIN}	Rank	Overall	Rank
MIROC-m	0.77	5	0.87	3	0.84	5	0.83	1
CSIRO	0.73	7	0.80	6	0.88	2	0.80	2
ECHO-G	0.83	3	0.87	2	0.69	12	0.80	3
IPSL	0.65	12	0.85	4	0.83	7	0.78	4
MRI	0.65	11	0.78	8	0.86	4	0.76	5
GISS AOM	0.64	13	0.78	7	0.83	8	0.75	6
FGOALS	0.70	9	0.81	5	0.69	13	0.73	7
CGCM-l	0.60	14	0.68	10	0.86	3	0.71	8
BCCR	0.85	1			0.73	11	0.79	
CGCM-h			0.72	9	0.84	6	0.78	
GFDL2.0	0.79	4					0.79	
GFDL2.1	0.76	6			0.89	1	0.82	
GISS ER	0.73	8			0.80	10	0.76	
MIROC-h			0.87	1			0.88	
ECHAM	0.84	2			0.81	9	0.83	
CCSM	0.67	10					0.67	

based solely on the literature. Overall, the Model for Interdisciplinary Research on Climate, version 3.2 T106 (MIROC-h), ECHO-G, and MIROC-m simulate T_{MAX} best over Australia (see Table 3).

Another advantage of the PDF skill score used in this paper is that it provides a direct way to omit models from ensembles based on a quantitative threshold. Figure 14 illustrates the use of the skill score to omit indi-

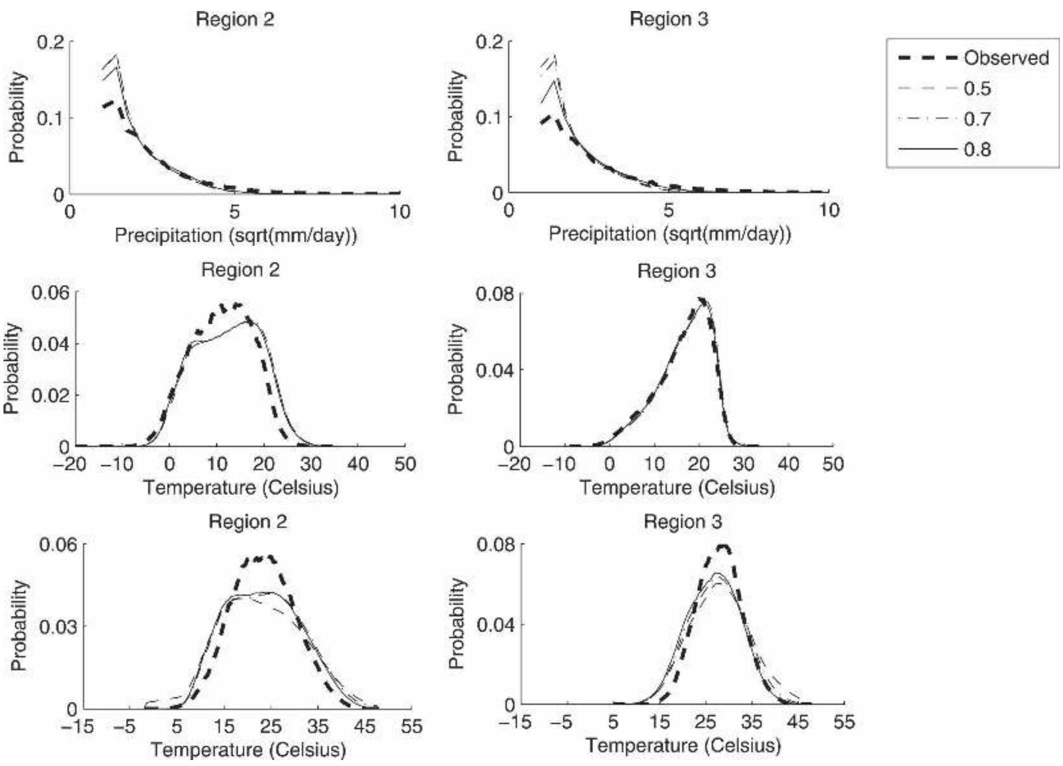

FIG. 14. Ensemble probability density functions for (top) precipitation, (middle) minimum temperature, and (bottom) maximum temperature for (left) region 2 (temperate) and (right) region 3 (tropical). Three simulated PDFs are shown where all model values with a skill score less than 0.5 are omitted, where all values less than 0.7 are omitted, and where all values less than 0.8 are omitted.

vidual models at specific locations. Only regions 2 and 3 are shown (temperate and subtropical climates) since these examples capture the majority of behavior found in other regions. In Fig. 14, models are omitted if their skill score in a given region is <0.5, <0.7, and <0.8. Obviously, as models with poorer skill are omitted, the resulting ensemble PDF becomes more similar to the observed. In the case of precipitation, omitting models with a skill score <0.8 halves the error in the resulting PDF in the subtropical region and reduces the error by about 30% in the temperate region. Overall, the simulation of precipitation at rates >5 mm day^{-1}, and the shape of the PDFs for temperature, are simulated by the better models with considerably more skill than anticipated. A systematic overestimation of higher values of T_{MIN} remains in the temperate region even when only the better models (skill scores >0.8) are included. In contrast, the skill in the tropical region for T_{MIN} is close to perfect. It is not clear why T_{MIN} can be captured so well in the Tropics but not in the temperate regions of Australia, although the coarse orography is a potential explanation. Finally, improvements in T_{MAX} are achieved in both regions, but clearly systematic errors remain in even the best models. This is not possible to attribute to one cause but solar radiation and terrestrial processes are likely factors.

In one sense, Fig. 14 shows how far we have to go to obtain a near-perfect representation of the observed PDFs of these variables. There is clearly considerable room for improvement in P and T_{MAX}. Overall, however, we reiterate that the skill in simulating P, T_{MIN}, and T_{MAX} by most models over all regions of Australia was better than we anticipated. It is worth restating the challenge: to develop a fully coupled global climate model that can simulate the daily observed PDFs of P, T_{MAX}, and T_{MIN} over all 10° × 10° regions of Australia. The best climate models can effectively meet this challenge with skill scores exceeding 0.8.

5. 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 likely the time scales that will most strongly affect human, physical, or biological systems.

To address the issue of mean-based evaluation of climate models we examined the capacity of the AR4 models to simulate the observed PDFs, region by region over Australia using daily data. The skill of each climate model to reproduce the PDF was assessed using a skill score based on the overlap between the observed and modeled PDFs (region by region). While large biases were identified in some models, in general, the AR4 climate models showed considerable skill (commonly $>80\%$) in representing the observed PDFs. These models could capture the changes in the shape of the PDFs as these changed across Australia. This is quite a remarkable achievement given that the complexity of the fully coupled global climate models used here. We suggest that strong performance in a PDF-based evaluation provides more confidence in a climate model that a means-based assessment. However, a model that does well in a PDF (or mean based) assessment could still hide major limitations in (say) the frequency of no-rain days, consecutive days over a threshold temperature, or events that are too rare to significantly contribute to the skill score.

Despite some limitations, we conclude that there are climate models within the AR4 archive that have useful skill in simulating the PDFs of P, T_{MIN}, and T_{MAX} over Australia. We have already noted that BCCR, ECHAM, and ECHO-G simulate P best over Australia, based on our skill score. MIROC-h, ECHO-G, and MIROC-m are best in simulating T_{MAX}; and GFDL2.1, CSIRO, and CGCM-l are best in simulating T_{MIN}. An overall ranking of those models that could be assessed for all of P, T_{MIN}, and T_{MAX} is shown in Table 3. Three models have skill scores over 0.8, averaged over Australia and over the three variables. They are, in order, MIROC-m, CSIRO, and ECHO-G. We note that several models, where data were missing for one or more variables, could not be ranked but are shown in Table 3 to be the best for a single variable (BCCR for P, MIROC-h for T_{MAX}). However, we also note that while BCCR was best for P, it was close to worst for T_{MIN}, highlighting the need for model evaluation to be based on several variables.

McAvaney et al. (2001) concluded that climate models were useful tools, at least down to subcontinental scales. Our analysis, while limited to one continent, suggests that some of the AR4 models show considerable skill at subcontinental 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, some models are demonstrably better than others. While this is not surprising, impacts groups could use our evaluation as a basis for choosing climate models for subsequent study.

Acknowledgments. We acknowledge the international modeling 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, U.S. Department of Energy. We also thank Silicon Graphics for help in porting Matlab.

REFERENCES

Alexander, L. V., and Coauthors, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.,* **111,** D05109, doi:10.1029/2005JD006290.

Boer, G. J., and S. J. Lambert, 2001: Second-order space time climate difference statistics. *Climate Dyn.,* **17,** 213–218.

Christensen, J. H., and O. B. Christensen, 2003: Severe summertime flooding in Europe. *Nature,* **421,** 805–806.

Collatz, G. J., L. Bounoua, S. O. Los, D. A. Randall, I. Y. Fung, and P. J. Sellers, 2000: A mechanism for the influence of vegetation on the response of the diurnal temperature range to a changing climate. *Geophys. Res. Lett.,* **27,** 3381–3384.

Collins, W. D., and Coauthors, 2006: The Community Climate System Model Version 3 (CCSM3). *J. Climate,* **19,** 2122–2143.

Colombo, A., D. Etkin, and B. Karney, 1999: Climate variability and the frequency of extreme temperature events for nine sites across Canada: Implications for power usage. *J. Climate,* **12,** 2490–2502.

Dai, A., 2001: Global precipitation and thunderstorm frequencies. Part I: Seasonal and interannual variations. *J. Climate,* **14,** 1092–1111.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 Global Coupled Climate Models. Part I: Formulation and simulation characteristics. *J. Climate,* **19,** 643–674.

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

Easterling, D. R., G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, and L. O. Mearns, 2000: Climate extremes: Observations, modeling, and impacts. *Science,* **289,** 2068–2074.

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, and T. Peterson, 2002: Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Res.,* **19,** 193–212.

Griffiths, G. M., and Coauthors, 2005: Change in mean temperature as a predictor of extreme temperature change in the Asia–Pacific region. *Int. J. Climatol.,* **25,** 1301–1330.

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noger, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., 2001: *Climate Change 2001: The Scientific Basis.* Cambridge University Press, 881 pp.

Johns, T. C., and Coauthors, 2006: The New Hadley Centre Climate Model (HadGEM1): Evaluation of coupled simulations. *J. Climate,* **19,** 1327–1353.

Katz, R., and B. Brown, 1992: Extreme events in a changing climate: Variability is more important than averages. *Climatic Change,* **21,** 289–302.

Kharin, V., and F. Zwiers, 2000: Changes in the extremes in a ensemble of transient climate simulations with a coupled atmosphere–ocean GCM. *J. Climate,* **13,** 3760–3788.

——, ——, and X. Zhang, 2005: Intercomparison of near surface temperature and precipitation extremes in AMIP-2 simulations. *J. Climate,* **18,** 5201–5223.

Kiktev, D., D. M. H. Sexton, L. Alexander, and C. K. Folland, 2003: Comparison of modeled and observed trends in indices of daily climate extremes. *J. Climate,* **16,** 3560–3571.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate,* **19,** 4224–4233.

Koster, R. D., and Coauthors, 2004: Regions of coupling between soil moisture and precipitation. *Science,* **305,** 1138–1140.

Luo, Q., R. N. Jones, M. Williams, B. Bryan, and W. Bellotti, 2005: Probabilistic distributions of regional climate change and their application in risk analysis of wheat production. *Climate Res.,* **29,** 41–52.

McAvaney, B. J., and Coauthors, 2001: Model evaluation. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 471–524.

Mearns, L., R. Katz, and S. Schneider, 1984: Extreme high-temperature events: Changes in the probabilities with changes in mean temperature. *J. Climate Appl. Meteor.,* **23,** 1601–1613.

Meehl, G. A., F. Zwiers, J. Evans, T. Knutson, L. Mearns, and P. Whetton, 2000: Trends in extreme weather and climate events: Issues related to modeling extremes in projections of future climate change. *Bull. Amer. Meteor. Soc.,* **81,** 427–436.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modeling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Osborn, T. J., and M. Hulme, 1997: Development of a relationship between station and grid-box rainday frequencies for climate model evaluation. *J. Climate,* **10,** 1885–1908.

——, and ——, 1998: Evaluation of the European daily precipitation characteristics from the Atmospheric Model Intercomparison Project. *Int. J. Climatol.,* **18,** 505–522.

Parkinson, G., Ed., 1986: *Atlas of Australian Resources.* 3d series, Vol. 4, *Climate,* Commonwealth of Australia, 60 pp.

Peterson, T. C., and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric data: A review. *Int. J. Climatol.,* **18,** 1493–1517.

Piani, C., D. J. Frame, D. A. Stainforth, and M. R. Allen, 2005: Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.,* **32,** L23825, doi:10.1029/2005GL024452.

Pitman, A. J., 2003: The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol.,* **23,** 479–510.

Schaeffer, M., F. M. Selten, and J. D. Opsteegh, 2005: Shifts in means are not a proxy for changes in extreme winter temperatures in climate projections. *Climate Dyn.,* **25,** 51–63.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophys. Res. Lett.,* **33,** L07702, doi:10.1029/2005GL025579.

Skelly, W., and A. Henderson-Sellers, 1996: Grid box or grid point: What type of data do GCMs deliver to climate impacts researchers? *Int. J. Climatol.,* **16,** 1079–1086.

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

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106** (D7), 7183–7192.

Trigo, R. M., R. García-Herrera, J. Díaz, and I. F. Trigo, 2005: How exceptional was the early August 2003 heatwave in France? *Geophys. Res. Lett.,* **32,** L10701, doi:10.1029/2005GL022410.

Watterson, I. G., 1996: Non-dimensional measures of climate model performance. *Int. J. Climatol.,* **16,** 379–391.

Wild, M., 2005: Solar radiation budgets in atmospheric model intercomparisons from a surface perspective. *Geophys. Res. Lett.,* **32,** L07704, doi:10.1029/2005GL022421.

Zwiers, F., and X. Zhang, 2003: Towards regional-scale climate change detection. *J. Climate,* **16,** 793–797.