Vol. 49: 169–187, 2011
doi: 10.3354/cr01028

CLIMATE RESEARCH
Clim Res

Published online October 20

OPEN ACCESS

Evaluating global climate models for the Pacific island region

Damien B. Irving[1,*], Sarah E. Perkins[1], Josephine R. Brown[2], Alex Sen Gupta[3], Aurel F. Moise[2], Bradley F. Murphy[2], Les C. Muir[4], Robert A. Colman[2], Scott B. Power[2], Francois P. Delage[2], Jaclyn N. Brown[4]

[1]Centre for Australian Weather and Climate Research, CSIRO Marine and Atmospheric Research, Aspendale, Victoria 3195, Australia

[2]Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria 3001, Australia

[3]Climate Change Research Centre, University of New South Wales, Sydney, New South Wales 2052, Australia

[4]Centre for Australian Weather and Climate Research, CSIRO Marine and Atmospheric Research, Hobart, Tasmania 7001, Australia

ABSTRACT: While the practice of reporting multi-model ensemble climate projections is well established, there is much debate regarding the most appropriate methods of evaluating model performance, for the purpose of eliminating and/or weighting models based on skill. The CMIP3 model evaluation undertaken by the Pacific Climate Change Science Program (PCCSP) is presented here. This includes a quantitative assessment of the ability of the models to simulate 3 climate variables: (1) surface air temperature, (2) precipitation and (3) surface wind); 3 climate features: (4) the South Pacific Convergence Zone, (5) the Intertropical Convergence Zone and (6) the West Pacific Monsoon; as well as (7) the El Niño Southern Oscillation, (8) spurious model drift and (9) the long term warming signal. For each of 1 to 9, it is difficult to identify a clearly superior subset of models, but it is generally possible to isolate particularly poor performing models. Based on this analysis, we recommend that the following models be eliminated from the multi-model ensemble, for the purposes of calculating PCCSP climate projections: INM-CM3.0, PCM and GISS-EH (consistently poor performance on 1 to 9); INGV-SXG (strong model drift); GISS-AOM and GISS-ER (poor ENSO simulation, which was considered a critical aspect of the tropical Pacific climate). Since there are relatively few studies in the peer reviewed literature that have attempted to combine metrics of model performance pertaining to such a wide variety of climate processes and phenomena, we propose that the approach of the PCCSP could be adapted to any region and set of climate model simulations.

KEY WORDS: Climate model evaluation · Regional climate projections · CMIP3 · Pacific

1. INTRODUCTION

As the scientific understanding of anthropogenic climate change has increased over recent decades, so has the demand for regional climate change projections. This is particularly true in the Pacific island region, where national government agencies and various international aid initiatives require information for input into risk assessments, in order to decide upon climate adaptation measures that are most deserving of the finite funds available. In an attempt to help guide such risk assessment work, the Pacific Climate Change Science Program (PCCSP, www.cawcr.gov.au/projects/PCCSP/) aims to improve the understanding of climate change in the Pacific, and provide a range of projections for the region. The data for the large-scale projections in this project will be derived primarily from the World Climate Research Project Coupled Model Intercomparison Project Phase 3 (CMIP3; Meehl et al. 2007) data archived at the Program for Climate Model Diagnosis and Intercomparison (PCMDI). In addition, a number of dynamically

*Email: damien.irving@csiro.au

and statistically downscaled projections driven by projections from a small number of selected CMIP3 models will be made available.

A pragmatic and well accepted approach to formulating climate projections involves combining the output from participating CMIP3 climate models to form a multi-model ensemble (Knutti et al. 2010b). However, while the practice of reporting multi-model ensemble projections is well established, many variations on the precise methods used to formulate an ensemble have been reported in the literature (Tebaldi & Knutti 2007). Some authors advocate the relatively simple method of assigning equal value (or weight) to each model, while other more sophisticated approaches assign different weights to different models, with the weights reflecting the respective skill of the models. In an attempt to promote consensus among the climate projection community, Weigel et al. (2010) recently introduced a simple conceptual model of climate change projections, to explore the effects of model-weighting. Consistent with the findings of a number of previous studies (e.g. Gleckler et al. 2008, Pierce et al. 2009, Sen Gupta et al. 2009), the first conclusion presented by Weigel et al. was that equally weighted multi-model ensembles yield, on average, more accurate projections than do the participating single models alone. They also found that in principle, projection errors can be further reduced by optimum weighting, but pointed out that establishing these optimal weights requires accurate knowledge of model skill, for which there is currently no consensus. Since their conceptual model indicated that incorrectly weighted multi-model ensembles generally perform worse than equally weighted ones, they suggested that equal weighting may be a safer and more transparent strategy to obtain optimum results. Within this framework, Weigel et al. suggested that it can be justified to eliminate models from an ensemble if they are known to lack 'key mechanisms which are indispensable for meaningful climate projections'. This approach of equal model weighting with the possible elimination of poor performing models is advocated by a number of other authors (e.g. Räisänen 2007, Stainforth et al. 2007).

While the findings of Weigel et al. (2010) provide a general framework for best practice, they give little guidance as to the finer details of determining regional projections. In particular, it is unclear what constitutes model failure to represent 'key mechanisms', and what model evaluation should be conducted to arrive at such a conclusion. As climate model evaluation is an inherently pragmatic and subjective process, there is in fact no single 'best' method of evaluating model performance. However, since the direct verification of any climate projection is impossible, model agreement with observations of the present climate (i.e. the 20th century) is used as a general approach to assigning model confidence, with the underlying assumption that a model which accurately describes the present climate will make a better projection of the future. While there is no guarantee that this assumption will hold true (e.g. Whetton et al. 2007, Reifen & Toumi 2009, Knutti et al. 2010b), a model that is unable to simulate key mechanisms of the current climate would be less likely to provide a credible projection of the near future (i.e. the 21st century).

The judgement of whether a model is skilful in simulating the present climate may include an assessment of its ability to represent the long term average and seasonal cycle of various model fields (e.g. climate variables such as atmospheric temperature or precipitation), important regional climate features such as the Intertropical Convergence Zone (ITCZ), climate variability on various timescales (e.g. the North Atlantic Oscillation or El Niño Southern Oscillation, ENSO), the observed climate change signal (i.e. warming of the lower atmosphere and upper ocean), and the stability of the modelled climate in the absence of greenhouse gas forcing (i.e. model drift). Since model performance can vary significantly depending on which of these aspects is assessed, recent studies have begun to develop model evaluation and selection strategies that incorporate analyses of multiple climate processes and phenomena (Gleckler et al. 2008, Pierce et al. 2009, Santer et al. 2009). Only when a broad suite of metrics that collectively capture many aspects of a climate model simulation has been calculated, does it become possible to identify models for potential elimination for certain applications. Furthermore, in determining the robustness of a given climate projection, it is common practice to compare and contrast projections arising from different subsets of models. This stratification of the multi-model ensemble is often based on an assessment of model skill (e.g. Smith & Chandler 2010).

In light of these issues, this study presents the results of a comprehensive model evaluation conducted for the Pacific island region and East Timor, which attempts to include a quantitative assessment of all the aforementioned aspects of a climate model simulation. While the results presented will be used to guide the climate projections delivered by the PCCSP, we propose that the same general methodology could be adapted to suit any region and ensem-

ble of climate model simulations. Our methods and associated results are outlined in Sections 2 and 3 respectively, while Section 4 discusses both the methodology and results in the context of the current literature.

2. METHODOLOGY

Since the PCCSP will provide climate projections for the Pacific nations of Palau, Federated States of Micronesia, Marshall Islands, Nauru, Kiribati, Papua New Guinea, Solomon Islands, Vanuatu, Tuvalu, Fiji, Tonga, Samoa, Niue and the Cook Islands, as well as East Timor, the CMIP3 climate models were, unless otherwise stated, evaluated over a geographic region encompassing all 15 countries (25° S to 20° N and 120° E to 210° E but excluding the Australian region south of 10° S and west of 155° E; hereafter referred to as the PCCSP region; see Fig. 1). As previously discussed, many aspects of a climate model simulation of the present climate may be assessed, and these can be grouped into 5 broad categories: (1) climate variables, (2) climate features, (3) climate variability, (4) climate stability (i.e. model drift) and (5) the observed climate change signal (i.e. the warming signal). This section provides a detailed description of our model evaluation process in terms of these 5 categories (summary in Table 1).

2.1. Data

The climate model data were obtained from the CMIP3 data archive at the PCMDI at Lawrence Livermore National Laboratory (www.pcmdi.llnl.gov/). A pre-industrial control run (PIcntrl) incorporating a seasonally varying but annually unchanging forcing indicative of the late 19th century (usually corresponding to 1870 conditions) was used as part of the evaluation of model drift; for all other analyses, the climate model data corresponded to the 20th Century Climate in Coupled Models simulation (20c3m). For the 20c3m runs, modelling groups initiated the models from the PIcntrl (~1870) simulations and then imposed the natural and anthropogenic forcing thought to be important for simulating the climate of the 20th and late 19th centuries. For models where multiple realisations are archived, we used only the first ensemble member for consistency with models that have only one simulation in the CMIP3 archive. The official CMIP3 defined name was used to designate each of the 24 participating climate models. Additional details on each of the models can be found at the PCMDI website listed above and in Table 8.1 of Randall et al. (2007).

Since the meteorological observation record for the PCCSP region suffers from substantial temporal and spatial gaps and inhomogeneities, various global gridded datasets were instead used to represent the observational record. For surface air temperature and surface wind, the European 40 yr reanalysis (ERA-40; Uppala et al. 2005), Japanese 25 yr reanalysis (JRA-25; Onogi et al. 2007) and the joint National Centres for Environmental Prediction and Department of Energy reanalysis (commonly designated NCEP/DOE R-2; Kanamitsu et al. 2002) datasets were used. These are the most modern reanalyses with data dating back to the beginning of the satellite era (i.e. 1979), thus maximising the time period over which comparisons could be made (the scarcity of ground-based data over the PCCSP region necessi-

Table 1. Summary of the statistical tests and observational datasets used to assess 9 key aspects of a climate model simulation. SPCZ: South Pacific Convergence Zone; ITCZ: Intertropical Convergence Zone; WPM: West Pacific Monsoon; ENSO: El Niño Southern Oscillation; SST: sea surface temperature. For observational dataset abbreviations and references, see Section 2.1

Category	Aspect	Observational datasets	Tests
Climate variables	Temperature, wind	ERA-40, NCEP/DOE R-2, JRA-25	Mean state, seasonal cycle (phase & amplitude), spatial features (location & amplitude)
	Precipitation	CMAP, GPCP	
Climate features	SPCZ, ITCZ, WPM	CMAP, GPCP	Location, interannual variability[a]
Climate variability	ENSO	HadISST	Strength, frequency, spatial pattern, link with precipitation
Climate stability	Drift	PIcntrl simulation	Drift magnitude (in surface air temperature and precipitation)
Warming signal	SST	HadSST2, ERSST.v3, Kaplan.v2	Trend

[a]Intensity was not assessed due to the large discrepancy between the CMAP and GPCP observational datasets

tates using only data from the beginning of the satellite era onwards). The Global Precipitation Climatology Project (GPCP; Adler et al. 2003) and Climate Prediction Centre Merged Analysis of Precipitation (CMAP; Xie & Arkin 1997) are generally considered to provide a more accurate representation of global rainfall than current reanalysis products (e.g. Beranger et al. 2006, Bosilovich et al. 2008), and were thus taken to represent the observational precipitation record. While there is some suggestion that the GPCP provides a more credible precipitation climatology over tropical ocean regions, it is not feasible to conclude that it supersedes CMAP, due to the lack of a reliable oceanic reference dataset for validation (Yin et al. 2004). For the analysis of long term trends in sea surface temperature (SST), the second Hadley Centre SST dataset (HadSST2; Rayner et al. 2006), the Extended Reconstruction SST version 3 dataset (ERSST.v3; Smith et al. 2008) and the Kaplan extended SST version 2 dataset (Kaplan.v2; Kaplan et al. 1998) were used. For the analysis of ENSO, observational data were derived from the higher resolution Hadley Centre Sea Ice and SST dataset (HadISST; Rayner et al. 2003). In the following sections, these data will be simply referred to as 'observations', keeping in mind that 'observationally based' or 'reference' data is generally more appropriate.

One limitation of most of these reference datasets is that it is difficult to estimate their associated uncertainties, which arise due to random and bias errors in the measurements themselves, sampling errors, and analysis error when the observational data are processed through models or otherwise altered. In the evaluation of climate models, these uncertainties are usually ignored since they are much smaller than the uncertainty in the model data (Gleckler et al. 2008). If models improve to the point where the associated uncertainty is comparable to that of the reference datasets, then a more rigorous approach will be required. Consistent with the recently released IPCC good practice guidelines on model evaluation (Knutti et al. 2010a), however, we attempted to capture some degree of this uncertainty by repeating our analyses for multiple reference datasets.

Many of the CMIP3 variables are only archived at a monthly timescale and the observational datasets used here often provide little or no daily data. As such, it was deemed most appropriate to evaluate the models only on monthly and longer timescales. This provides a reasonably comprehensive picture of model performance, but excludes the evaluation of some climate aspects, such as the frequency of extreme events that would only be evident in daily or subdaily data. The temporal coverage of the monthly data for each observational dataset spans from the present day back until 1979 or earlier, while the 20c3m model runs terminate at the end of either 1999 or 2000. Hence, we evaluated over the common 21 yr period 1979–1999 for consistency, unless otherwise stated. All atmospheric (oceanic) data were interpolated to a common 2.5° (1.0°) latitude × longitude grid prior to analysis.

2.2. Climate variables

The majority of the PCCSP climate projections relate to surface air temperature, precipitation and surface wind speed and direction. For each of these variables, statistical tests (or metrics) used to compare a climate model simulation with corresponding observational data should assess both the phase and amplitude of the seasonal cycle, in addition to the location and amplitude of any spatial features. Thus, we calculated a grid point average temporal correlation (r_t) and temporal standard deviation (SD) ratio (model / observed; $\sigma_{ratio,t}$) over the PCCSP region, in addition to a monthly time-step average spatial (or pattern) correlation (r_p and spatial SD ratio ($\sigma_{ratio,x}$). At each grid point, the 2 temporal statistics were calculated from a 12-step time series containing the 1979–1999 mean value for each month, while the 2 spatial statistics were calculated from the 1979–1999 mean field for each month. Finally, in order to assess the mean state of a climate model simulation, the average magnitude of the grid point errors in the 1979–1999 annual mean field was calculated (E_{abs}). While a number of studies have proposed various metrics of climate model performance that combine 2 or more of these tests (e.g. Watterson 1996, Taylor 2001), no metric is able to simultaneously assess all temporal and spatial aspects covered here, hence our preference for assessing each fundamental statistic separately.

2.3. Climate features

The major large-scale climate features of the PCCSP region are the South Pacific Convergence Zone (SPCZ), ITCZ and the West Pacific Monsoon (WPM) (see Fig. 1). Since all 3 are characterised by a precipitation maximum, relevant metrics used to define these phenomena can be calculated from precipitation data (e.g. Lin 2007, Brown et al. 2011, Vincent et al. 2011). Any evaluation of these climate features should ideally compare the model simulated location,

intensity and interannual variability with observational data. However, it is well known that there is a large discrepancy between the CMAP and GPCP datasets with respect to the mean intensity of tropical precipitation features such as the ITCZ, SPCZ and WPM (Gruber et al. 2000, Yin et al. 2004). Thus, while it was considered appropriate to assess precipitation intensity over the entire PCCSP region in the climate variables assessment outlined in Section 2.2 (the difference between CMAP and GPCP intensity is less pronounced over broad spatial domains), mean intensity was not assessed for any climate feature. This approach is consistent with a recent SPCZ model evaluation (Brown et al. 2011). Each of the climate features is discussed in more detail below, including a description of the metrics used to assess their location and interannual variability.

2.3.1. South Pacific Convergence Zone

The SPCZ is a band of low-level atmospheric convergence and precipitation, which extends northwest-southeast in a diagonal line from near Papua New Guinea (0°, 150° E) to the south-eastern Pacific (~30° S, 120° W) (e.g. Vincent 1994). This feature is responsible for a large fraction of the climatological precipitation in the South Pacific, particularly during the Austral summer (Dec to Feb, DJF) when it is most active and well defined. Climate model representations of this feature tend to be overly zonal in their orientation, and a small minority actually merges SPCZ and ITCZ precipitation (e.g. Lin 2007, de Szoeke & Xie 2008, Bellucci et al. 2010, Brown et al. 2011).

In evaluating the ability of the CMIP3 climate models to simulate a realistic SPCZ, we adopted the SPCZ region (155° E to 140° W and 0 to 30° S; see Fig. 1) of Brown et al. (2011). The mean location of the SPCZ over this region was assessed by calculating the spatial (or pattern) correlation between the modeled and observed 1979–1999 DJF mean precipitation fields ($r_{p,SPCZ}$). With regard to interannual variability, it is well known that El Niño (La Niña) events are associated with an SPCZ shift to the north and east (south and west) (e.g. Folland et al. 2002). In fact, Brown et al. identified a strong correlation ($r > 0.8$) between the observed 1979–1999 DJF mean latitude of the SPCZ for each year and the corresponding observed Niño 3.4 region (5° N to 5° S and 120 to 170° W) SST anomaly index (Trenberth 1997). Thus, the same correlation (denoted $r_{niño-SPCZ(lat)}$) was calculated here, whereby the mean latitude of the SPCZ was calculated from a linear fit to the latitude of maximum precipitation over the longitudinal extent of the SPCZ region. While the 1979–1999 period was used for the observational data, a longer 50 yr period (1950–1999) was taken for the model data to provide a larger sample of model responses to ENSO.

2.3.2. Intertropical Convergence Zone

The ITCZ is a persistent region of low-level convergence, cloudiness and precipitation, which is located immediately north of the equator and extends roughly zonally across the Pacific. It is marked by the presence of a surface pressure trough, and formed by the convergence of warm and moist Northern and Southern Hemisphere trade winds. The characteristics of the ITCZ vary noticeably with longitude. It is located over the latitudinal SST maximum in the eastern Pacific, but not in the west. Moreover, while the ITCZ is narrow in the central and eastern Pacific (corresponding conceptually to the traditional trade wind 'convergence' model), in the western Pacific it becomes broad due to strong monsoon flows and the latitudinally broad region of the west Pacific warm pool (e.g. Waliser & Gautier 1993). Climate models typically represent a precipitation maximum in the ITCZ region, but as previously mentioned, they do not always separate it well from the SPCZ in the western Pacific. In the central and eastern Pacific, many models suffer from a 'double ITCZ', which tends to be characterised by 2 behaviours: 'persistent double ITCZ' (rain persisting too long in the Southern Hemisphere for at least 4 mo), and an 'alternating ITCZ error', whereby precipitation erroneously migrates into the Southern Hemisphere (de Szoeke & Xie 2008).

Similar to our SPCZ analysis, it was considered appropriate to make a quantitative assessment of the model-simulated location of the ITCZ by calculating pattern correlations. However, given the longitudinally varying characteristics of the ITCZ, these pattern correlations were calculated separately for the 2 distinct sections of the ITCZ that are directly relevant to the climate of the PCCSP region: 165° E to 165° W (western) and 165° W to 135° W (central), both from the equator to 15° N (Fig. 1). Since the latitudinal location of the ITCZ moves throughout the seasonal cycle, with a maximum southward (northward) extent during February (Jul to Sept), these regional correlations were calculated for each of the 4 seasons (DJF, Mar to May MAM, Jun to Aug JJA and Sept to Nov SON), with the seasonal values then averaged to produce a single regional pattern correlation ($r_{p,ITCZ(west)}$ and $r_{p,ITCZ(cen)}$). The average of these 2 re-

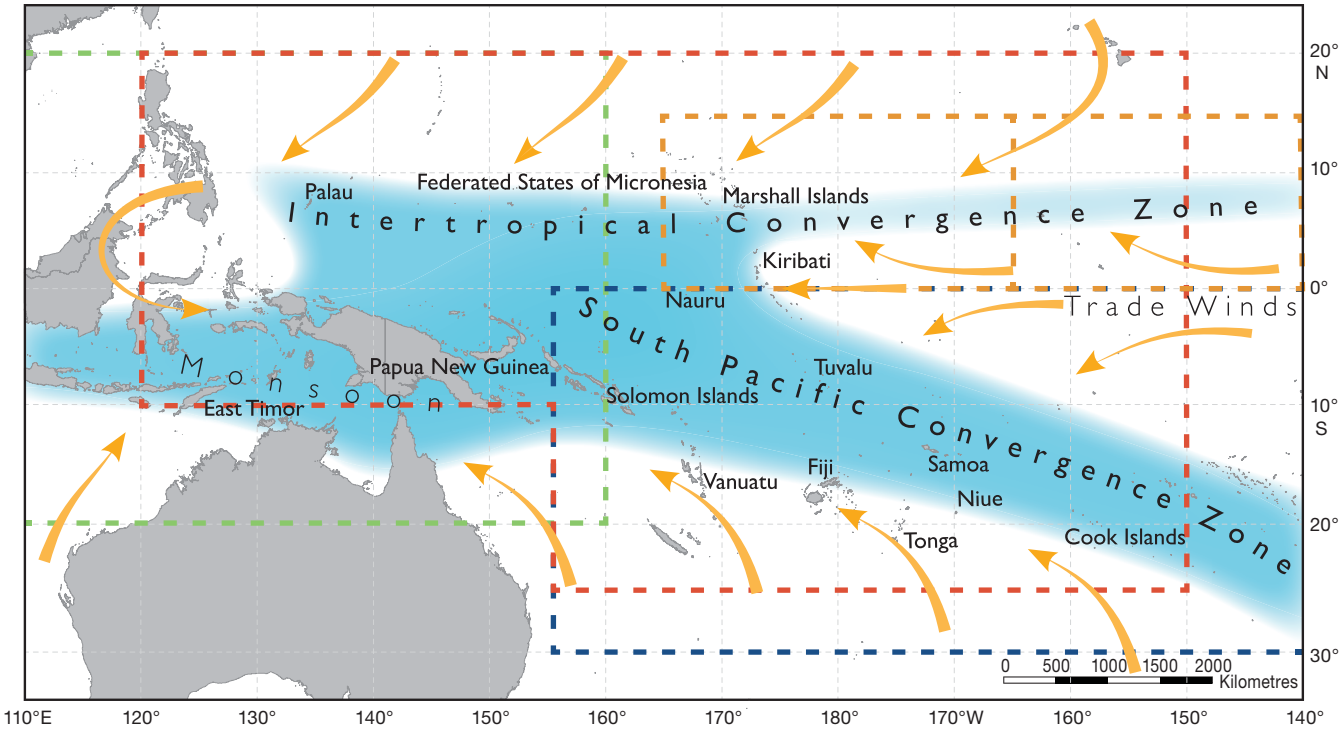

Fig. 1. Geographic location of the Pacific Climate Change Science Program (PCCSP) partner countries. Dominant features of the regional climate (approximately corresponding to the Nov to Apr season) are shown. Dashed red line: boundary of 'PCCSP region' used in the climate variables evaluation; blue, orange and green lines: South Pacific Convergence Zone, Intertropical Convergence Zone (ITCZ), and West Pacific Monsoon regions respectively (the ITCZ region is split into a western and central sector)

gional pattern correlations ($r_{p,ITCZ}$) was taken as the measure of the model simulated location of the ITCZ.

Interannual variability in the location of the ITCZ is known to be closely linked to the ENSO. As the central/eastern Pacific warms (cools) during an El Niño (La Niña) event, the ITCZ is drawn closer (further) from the equator. In order to assess the ability of the models to capture this behaviour, we calculated the composited El Niño minus La Niña location (i.e. mean latitude) of the ITCZ for events over 1950–1999 (for DJF only, as this is when ENSO events are typically strongest; denoted $L_{niño-ITCZ}$). An El Niño and a La Niña were defined as events exceeding +0.75 and –0.75 SDs respectively of the 5 mo running-mean Niño 3.4 index, while the mean latitude of the ITCZ was calculated from a linear fit to the latitude of maximum precipitation over the longitudinal extent of the region spanning 160° E to 120° W and 0° to 20° N. As with our SPCZ analysis, the 1979–1999 period was assessed for the observational data, while a longer 50 yr period (1950–1999) was taken for the model data to provide a larger sample of model responses to ENSO.

2.3.3. West Pacific Monsoon

The WPM is the southern extension of the larger Asian-Australian monsoon system (Wang 2006). In the PCCSP partner countries, the migration of the WPM into the tropical regions of the Southern Hemisphere during the Austral summer months (DJF) is responsible for the majority of wet season rainfall in East Timor and Papua New Guinea. The ability of CMIP3 models to simulate WPM rainfall is quite varied, since the monsoon system is a broad-scale phenomenon incorporating large variability on smaller scales owing to the strongly localised rainfall systems. This characteristic has been very difficult to correctly simulate; thus, many models underestimate factors such as the interannual variability in monsoon precipitation (Kim et al. 2008). There is evidence that some of this deficiency may originate from the relatively low resolution and therefore deficient topography of the CMIP3 models (Qian & Zubair 2010).

The mean location of the WPM was assessed by calculating the pattern correlation between the modeled and the observed DJF precipitation ($r_{p,WPM}$) over

an appropriate WPM region (110° to 160° E and 20° S to 20° N; Fig. 1). With respect to interannual variability, the intensity of the WPM is known to have a strongly inverse relationship with ENSO (e.g. Zhu & Chen 2002); hence, the ability of the models to simulate this relationship was assessed by calculating the correlation between the total DJF precipitation over the WPM region and the model simulated Niño 3.4 index ($r_{niño–WPM(pr)}$).

2.4. Climate variability

Interannual climate variability in the Pacific and surrounding regions is dominated by the ENSO. This coupled ocean–atmosphere phenomenon is a mode of natural climate variability with irregular cycles between large-scale warming (El Niño) and cooling (La Niña) of the central and eastern equatorial Pacific Ocean over periods of 2 to 7 yr. In the atmosphere, the Southern Oscillation manifests as large-scale out-of-phase changes in surface pressure in the east and west Pacific and a weakening (strengthening) of the Walker Circulation during El Niño (La Niña) events. Many global weather systems and climate phenomena respond to the ENSO; in the PCCSP region, these responses include alterations to rainfall patterns and the location of tropical cyclone activity, in addition to regional sea level and surface air temperature changes (e.g. Diaz & Markgraf 2000). Since reproducing ENSO-like behaviour in coupled climate models is a very complex task, it is a significant achievement that most CMIP3 models do simulate an ENSO-like phenomenon (e.g. AchutaRao & Sperber 2006). However, while the simulation of the ENSO has steadily improved with time, a number of issues remain. Common deficiencies include an equatorial cold tongue that is too cold, extends too far into the west Pacific warm pool and is too confined near the equator (commonly known as the 'cold tongue bias'); ENSO events that are either too regular or too frequent; and ENSO events that are not phase-locked correctly to the annual cycle (see Guilyardi et al. 2009 and references therein).

In order to produce a reasonable ENSO simulation, climate models need to realistically reproduce (i) the strength and frequency of ENSO events, (ii) the mean climate and spatial pattern of ENSO, and (iii) the link between ENSO and climate variables such as precipitation. With regard to the first, the strength of model simulated ENSO events was assessed by calculating the ratio of SDs between the modelled and the observed Niño 3.4 index monthly time series ($\sigma_{ratio,niño}$). The frequency of El Niño and La Niña events was assessed by calculating the number of events that occurred over the period 1950–1999 (N_{events}), where an event was defined as exceeding a 5 mo Niño 3.4 running-mean magnitude of 0.4°C for 6 consecutive months (Trenberth 1997). It should be noted that continuous multi-year events were counted only once and that the definition of an El Niño and a La Niña event differed slightly from that in the ITCZ analysis since the entire annual cycle (as opposed to only DJF) was considered. In order to assess the SST pattern associated with ENSO events, the correlation coefficient between the mean July to December Niño 3.4 index and the mean SST anomaly at each grid point over an 'ENSO region' (25° S to 25° N and 120° to 240° E) was first calculated for both model and observed data (HadISST). The pattern correlation between these model and observed correlation fields was then calculated as an indication of the model ability to simulate the mean climate and spatial pattern of ENSO ($r_{p,niño–SST}$).

Similarly, the link between ENSO and precipitation over the ENSO domain was assessed by first calculating the model and the observed temporal correlation between the July to December Niño 3.4 mean value and the corresponding precipitation totals at each grid point. The pattern correlation between the model and the observed (CMAP and GPCP) grid point temporal correlation fields could then be calculated ($r_{p,niño–pr}$). While this pattern correlation on its own captures the ability of the models to simulate the mean pattern of association between the ENSO and precipitation, it does not assess the spatial variance in that pattern. Hence, the SD ratio between the model and the observed temporal correlation fields was also calculated ($\sigma_{ratio,r_{p,niño–pr}}$) and combined with $r_{p,niño–pr}$ as per the *S* statistic proposed by Taylor (2001):

$$S_{ENSO–pr} = \frac{4(1+r_{p,niño\text{-}pr})^4}{\left(\sigma_{ratio,r_{p,niño–pr}} + 1/\sigma_{ratio,r_{p,niño–pr}}\right)^2 (1+r^{*}_{p,niño\text{-}pr})^4} \quad (1)$$

where $r^{*}_{p,niño–pr}$ denotes the perfect $r_{p,niño–pr}$, which is taken to be the highest score obtained among the models. Scores on this metric can range from 0 (no skill) to 1 (perfect skill).

2.5. Climate stability

A persistent problem with coupled climate models is that of drift. This refers to spurious long term trends within climate simulations that are unrelated

to any external forcing (e.g. Covey et al. 2006). These trends are instead a result of discontinuities in surface fluxes during the coupling of climate model components and deficiencies in model physics, which mean that the equilibrium state of the model is different from the initial (usually observationally derived) state. While climate model drift is easiest to identify in unforced PIcntrl experiments, the spurious trend carries through to any corresponding forced simulation. Model drift is often considered to be an ocean model problem, since the relatively slow moving ocean takes hundreds or thousands of years to reach equilibrium (impractically long for most coupled climate simulations), while the atmosphere does so over a few seasons. However, given that the ocean and atmosphere are strongly coupled, particularly in the tropical Pacific, drift can also persist in atmospheric properties. For example, simple arguments suggest that precipitation minus evaporation (P–E) changes scale in proportion to SST changes and the mean P–E (Held & Soden 2006, Cravatte et al. 2009). As P–E can be very large in the tropics, relatively small drifts in SST can create important changes in precipitation.

In order to objectively evaluate drift in each of the models, we calculated the magnitude of the PIcntrl linear surface air temperature and precipitation trend at each grid point over the PCCSP region. Rather than restrict the analysis to the 1979–1999 period, we calculated these trends over a longer 150 yr period where possible to minimise contamination of the drift signal by low frequency natural variability (under the assumption that drift remains constant over this period). While the PIcntrl simulation has no calendar years associated with it, the 150 yr period was selected to align with the 1900–2050 period of the 20c3m simulation (i.e. using the point in time [~1870] at which the 20c3m simulation branched from PIcntrl as a reference point). The spatial average of these trend magnitudes was taken as a measure of model drift (D_{tas} and D_{pr} for surface air temperature and precipitation, respectively).

An important detail to note when considering climate model performance is that ocean model drift is so pronounced in some earlier generation models that many add artificial fluxes of heat, freshwater, and in some instances momentum at the air–sea interface in an attempt to offset the problem. This process is known as flux correction and despite the fact that most of the CMIP3 models no longer require flux correction to avoid model drift, 4 of the CMIP3 models (CGCM3.1(T47), CGCM3.1(T63), ECHO-G and MRI-CGCM2.3.2) use heat and freshwater flux corrections, while another model (INM-CM3.0) uses freshwater flux corrections only. The technique of flux correction attracts concern because of its inherently non-physical nature; however, it may reduce biases in the mean state of the surface ocean, leading to a more realistic present day surface temperature and salinity pattern and associated precipitation distribution.

2.6. Climate change signal

The ability of the CMIP3 models to capture the warming observed in recent decades was assessed by calculating linear SST trends for the 1950–1999 period of the 20c3m simulation, at each grid point over the PCCSP region. These trends were corrected for model drift by subtracting the linear trends from the 150 yr PIcntrl simulations (see Section 2.5), and then compared to observed trends (from the HadSST2, ERSST.v3 and Kaplan.v2 datasets) by calculating the average magnitude of the grid point errors (E_{trend}). There are no PIcntrl data available for the ECHAM5/MPI-OM and PCM models; hence, the trends in these models were not 'de-drifted' prior to comparison with the observational data.

2.7. Combined results

The model evaluation described above can be grouped into 9 key aspects of a climate model simulation: surface air temperature, precipitation, surface wind, SPCZ, ITCZ, WPM, ENSO, drift and (de-drifted) SST trend. Since multiple metrics were calculated for each aspect, a method of combining results across metrics of varying characteristics was required. A number of combination methods have been applied in the literature, the most simple of which involves ranking each model (from 1 to 24 in this case) on each metric (e.g. Perkins et al. 2009). An average ranking can then be calculated as an indication of the overall model performance on any given aspect. While attractive in its simplicity, this method assumes uniform spread among model scores on any particular metric (e.g. it is assumed that the distance between the 1st and 2nd ranked model is the same as that between the 17th and 18th). In order to avoid this assumption of uniform spread, we instead calculated a normalised score for each metric using a method similar to that applied by Santer et al. (2009). For this calculation, a given test score was first converted to an absolute error, E, which indicates its dis-

tance from the perfect (or observed) score (e.g. a pattern correlation of $r_p = 0.87$ becomes $E = 0.13$). As in a number of previous studies (e.g. Gleckler et al. 2008, Santer et al. 2009), in determining the absolute error for SD ratios (denoted here by a generic σ_{ratio}), we calculated a symmetric variability statistic (E_σ), which has the same numeric value for a model that simulates half and twice the observed SD:

$$E_\sigma = \left[\sigma_{ratio} - \frac{1}{\sigma_{ratio}}\right]^2 \quad (2)$$

Once the absolute error was obtained, it was normalised by conversion to an anomaly (i.e. by subtracting the multi-model average score) and then dividing that anomaly by the intermodel SD. Good (poor) performance was therefore indicated by increasingly negative (positive) normalised scores; for each metric, these normalised scores had a multi-model average of 0.0 and an SD of 1.0. To obtain an indication of overall model performance on each of the 9 key aspects of a model simulation, an average over all relevant normalised scores was calculated for each.

Given the wide range of climate projections that will be provided by the PCCSP, it is plausible that for some applications it will be most appropriate to simply sample a subset of the analyses presented. For instance, in producing projections of future SPCZ behaviour, model selection may simply take into account the average normalised test score for the SPCZ alone. However, for other projections it may be considered most appropriate to base model selection on an assessment of multiple aspects of a climate model simulation. With this latter situation in mind, we also calculated the mean of the 9 average normalised test scores corresponding to each key aspect of a climate model simulation, as an indication of overall model performance. There are several implied statistical assumptions associated with calculating such an average, which are explored in detail in Section 4.

3. RESULTS

When considering the results presented in this section, it is important to consider the context in which they are presented. As discussed in Sections 2.3 & 2.4, for some of the aspects we assessed, there is a considerable body of literature devoted to analysing and understanding climate model performance. In order to sufficiently explain our results, we will touch on some of the issues associated with simulating each aspect; however, our main objective is to present a method of statistically quantifying model performance via the calculation of scalar metrics, as opposed to providing a rigorous analysis of the model characteristics and deficiencies responsible for the test scores we obtained (i.e. diagnostics such as spatial maps and distributions are generally not shown and analysed in detail). The consistency of our results with rigorous analyses of this type will be considered in Section 4. Unless otherwise stated, the results presented refer to the multi-model average ± intermodel SD of the all-observation test scores (i.e. the value obtained by averaging across the scores from all relevant observational datasets).

3.1. Climate variables

3.1.1. Surface air temperature

The vast majority of models showed a cold bias throughout much of the PCCSP region (20 out of 24 had a spatial average 1979–1999 annual mean surface air temperature that was less than the observational average), which was largely responsible for the multi-model average mean grid point error magnitude of E_{abs} = 1.02 ± 0.48°C (Fig. 2). The models tended to overestimate the amplitude of the seasonal cycle ($\sigma_{ratio,t}$ = 1.18 ± 0.23), although the phase of this cycle was relatively well represented (r_t = 0.74 ± 0.06). A higher correlation was obtained in assessing the ability of the models to represent the spatial surface air temperature pattern for each month (r_p = 0.89 ± 0.05), while the models tended to slightly overestimate the magnitude of the features within this pattern (e.g. the intensity of the equatorial cold tongue was generally overestimated; $\sigma_{ratio,x}$ = 1.08 ± 0.11). Analysis of the normalised test scores (Fig. 3a) revealed that the PCM and INM-CM3.0 models performed particularly poorly, with the former markedly overestimating the amplitude of the seasonal cycle ($\sigma_{ratio,t}$ = 1.92). No models could be identified as clearly superior to the remainder of the ensemble.

3.1.2. Precipitation

The main large-scale features of the climatological precipitation in the Pacific, including the maxima associated with the SPCZ, ITCZ and WPM, were present in most model simulations (see Section 3.2); however, discrepancies in the precise location of these features as compared to observational data generally conspired to produce relatively large grid point

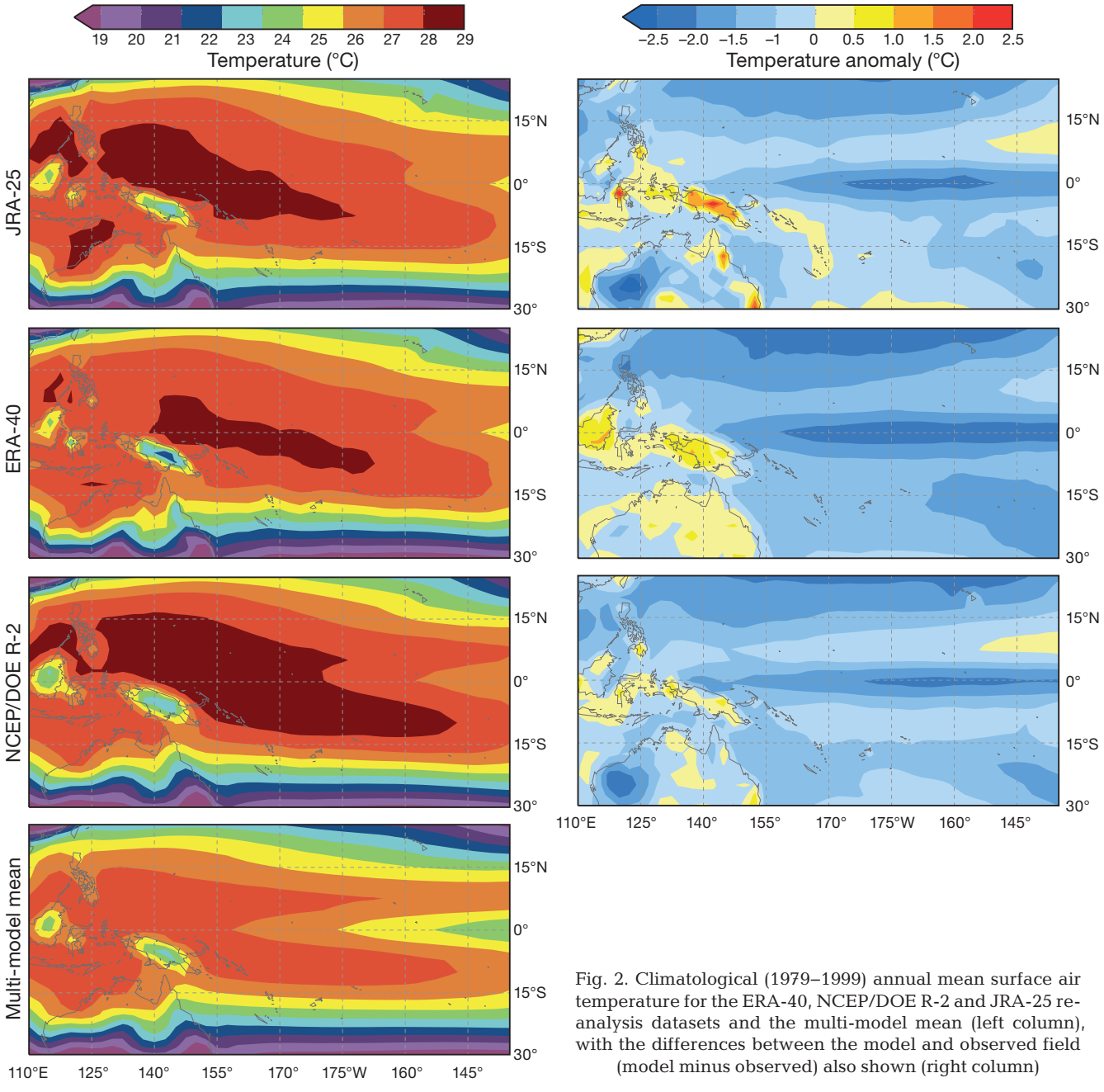

Fig. 2. Climatological (1979–1999) annual mean surface air temperature for the ERA-40, NCEP/DOE R-2 and JRA-25 reanalysis datasets and the multi-model mean (left column), with the differences between the model and observed field (model minus observed) also shown (right column)

errors in the 1979–1999 annual mean precipitation field (E_{abs} = 1.77 ± 0.46 mm d^{-1}; Fig. 4). In addition, these location discrepancies impacted upon the spatial correlation values obtained for each month (r_p = 0.63 ± 0.14). The models demonstrated a similar ability in capturing the phase of the seasonal cycle (r_t = 0.55 ± 0.12). Given the aforementioned differences between the CMAP and GPCP datasets with respect to the intensity of tropical precipitation (Section 2.3; Fig. 4), it was interesting to note that the amplitude of both the modelled seasonal cycle and the spatial distribution of precipitation in the PCCSP region tended to compare much more favourably with CMAP ($\sigma_{ratio,t}$ = 1.06 ± 0.24 and $\sigma_{ratio,x}$ = 1.10 ± 0.20) than with GPCP ($\sigma_{ratio,t}$ = 1.45 ± 0.33 and $\sigma_{ratio,x}$ = 1.43 ± 0.26). As with surface air temperature, no clearly superior models could be identified from the normalised statistics presented in Fig. 3b; moreover, the PCM and INM-CM3.0 models were again particularly poor performers.

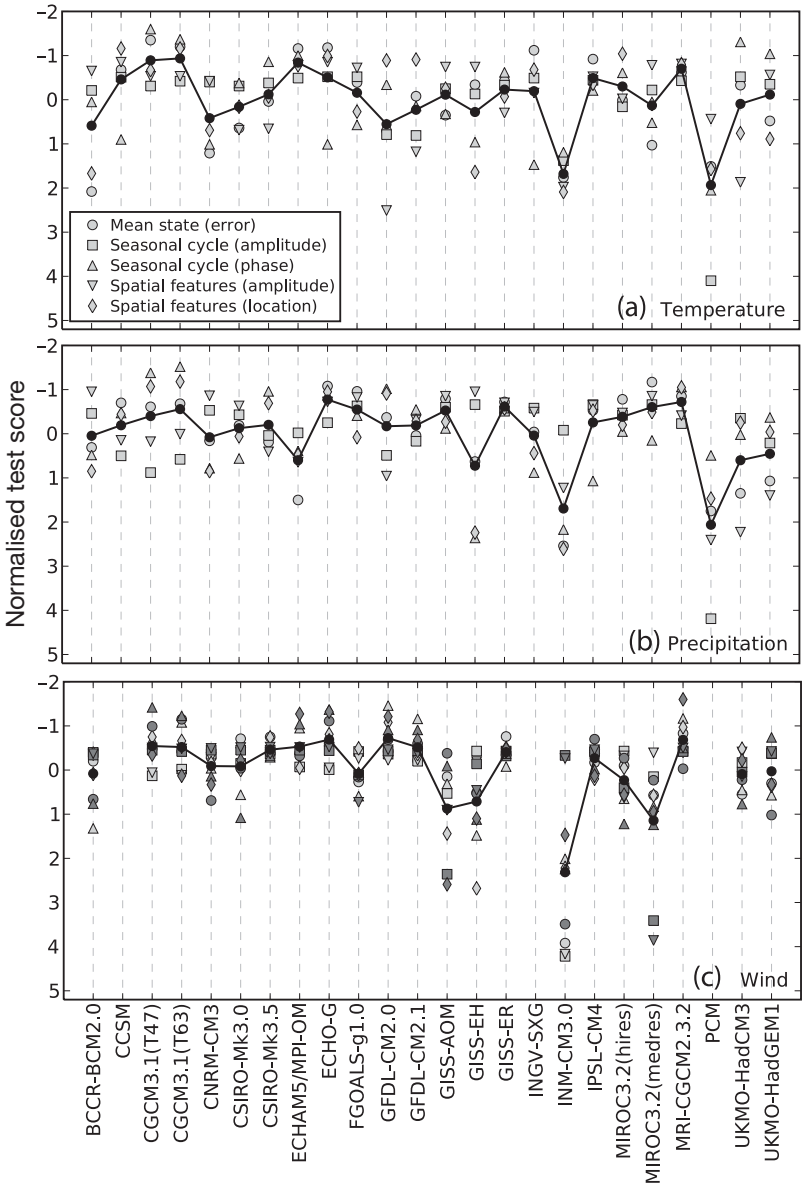

Fig. 3. Normalised test scores (increasingly negative scores indicate better model performance) for each of the analysed climate variables: (a) surface air temperature, (b) precipitation and (c) surface wind. (○) E_{abs}, indicative of the time mean state; (□) $\sigma_{ratio,t}$, amplitude of the seasonal cycle; (△) r_t, phase of the seasonal cycle; (▽) $\sigma_{ratio,x}$, amplitude of spatial features; (◇) r_p, location of spatial features; (—●—) average across all tests. For surface wind, dark (light) grey shading indicates wind direction (speed)

3.1.3. Surface wind

Throughout much of the PCCSP region, the CMIP3 models tended to display a bias towards wind speed overestimation as compared to the observational data, which contributed to a multi-model average grid point error magnitude of $E_{abs} = 0.89 \pm 0.28$ m s^{-1}. The equivalent result for wind direction was $E_{abs} = 14.70 \pm 4.25°$; however, on a subregional scale, errors larger than 10° were typically confined to the region surrounding Indonesia and Papua New Guinea. This result might be due to deficiencies in the ability of the models to correctly capture the large seasonal changes in wind direction associated with the WPM, or could possibly indicate a mismatch between model and reanalysis topography. With respect to the former, the models tended to underestimate the magnitude of the seasonal cycle in wind direction across the PCCSP region ($\sigma_{ratio,t} = 0.86 \pm 0.21°$) and also showed some difficulty in capturing the phase ($r_t = 0.72 \pm 0.09$ [wind speed]; $r_t = 0.63 \pm 0.10$ [wind direction]) and spatial pattern ($r_p = 0.73 \pm 0.07$ [wind speed]; $r_p = 0.45 \pm 0.08$ [wind direction]) of this cycle. The INM-CM3.0 model performed markedly worse than the remainder of the ensemble, while there were no clearly superior models (Fig. 3c).

3.2. Climate features

3.2.1. South Pacific Convergence Zone

The multi-model average pattern correlation for DJF rainfall over the SPCZ region ($r_{p,SPCZ} = 0.68 \pm 0.14$) was likely influenced by the aforementioned overly zonal orientation of the SPCZ in most models (Section 2.3.1; Fig. 4). A similar multi-model average (temporal) correlation was found between the location of the SPCZ and the Niño 3.4 index, although the spread in model performance was much greater ($r_{niño-SPCZ\,(lat)} = 0.61 \pm 0.30$). Given that CMAP and GPCP had corresponding correlations of $r_{niño-SPCZ\,(lat)} = 0.83$ and 0.86 respectively, it is noteworthy that the best 10 or so performing models had correlations that closely matched these observational values. The normalised statistics revealed no outstanding models, while the GISS-AOM, GISS-ER, GISS-EH, INM-CM3.0, MIROC 3.2(hires), MIROC3.2(medres) and UKMO-Had GEM1 models performed poorly (Fig. 5a).

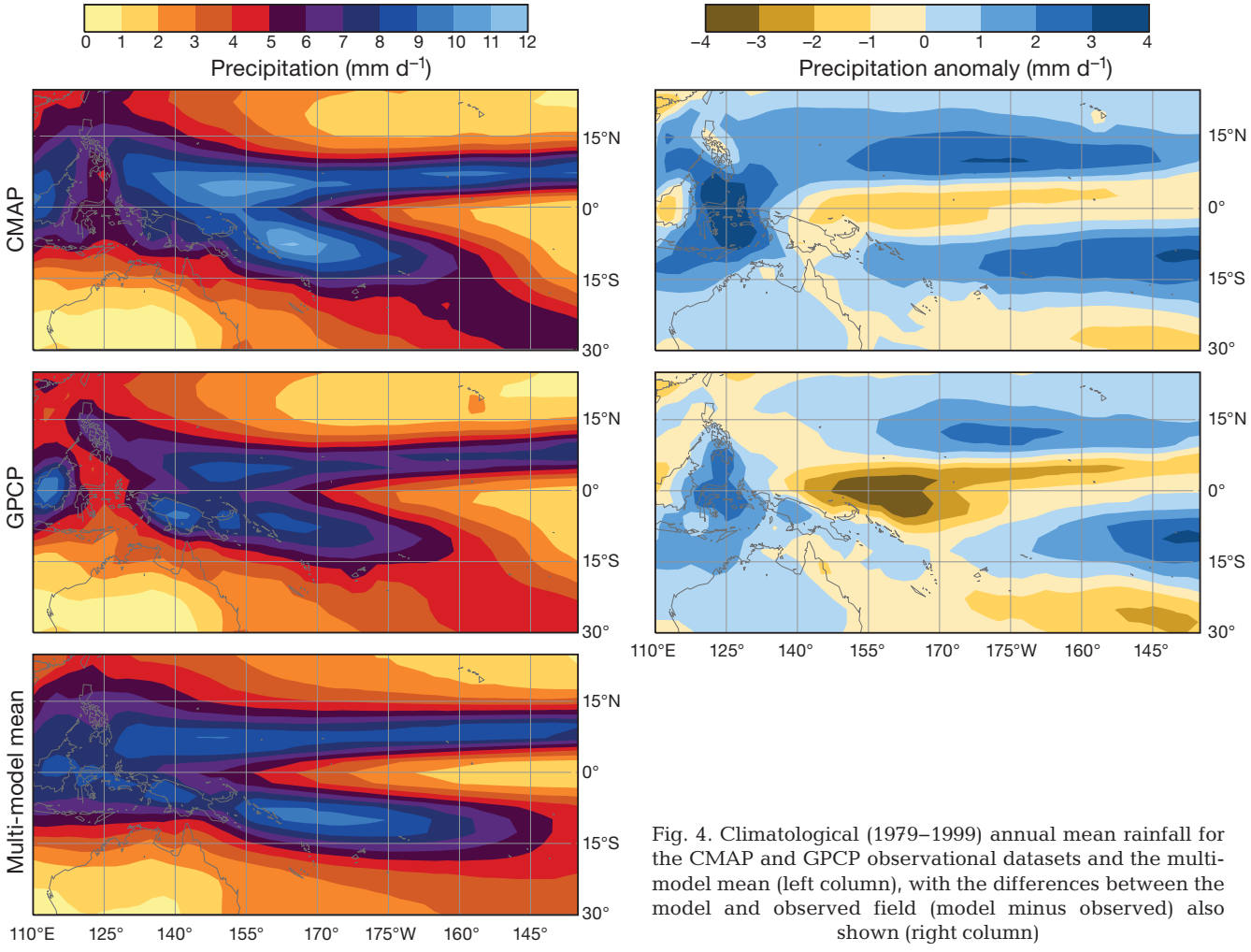

Fig. 4. Climatological (1979–1999) annual mean rainfall for the CMAP and GPCP observational datasets and the multi-model mean (left column), with the differences between the model and observed field (model minus observed) also shown (right column)

3.2.2. Intertropical Convergence Zone

The multi-model average seasonal precipitation pattern correlation across the central and western ITCZ regions was $r_{p,ITCZ} = 0.64 \pm 0.22$. Models tended to perform better in the central region ($r_{p,ITCZ(cen)} = 0.69 \pm 0.19$ vs. $r_{p,ITCZ(west)} = 0.59 \pm 0.26$), although the order of the models from best to worst remained relatively unchanged between the regions, consistent with the zonal nature of the ITCZ. In the western region, model correlations peaked in DJF (on average) and were lowest in JJA (on average only half as large). The latter is partly the result of the model ITCZ's being generally too far poleward in JJA. These seasonal correlation discrepancies were similar but less pronounced in the central region. The modelled shift in the mean latitude of the ITCZ in response to ENSO was generally too small ($L_{niño-ITCZ} = -2.23 \pm 1.52$ compared to the CMAP value of −2.90); however, all models showed the correct direction of movement. The normalised statistics revealed the MRI-CGCM2.3.2 to be the best performing model, while GISS-EH and INM-CM3.0 performed poorly (Fig. 5b).

3.2.3. West Pacific Monsoon

The multi-model average DJF precipitation pattern correlation for the WPM region was similar to those obtained for the SPCZ and ITCZ ($r_{p,WPM} = 0.65 \pm 0.14$), while the GISS-EH model was a pronounced outlier ($r_{p,WPM} = 0.23$). The ability of the models to capture the strongly inverse relationship between the intensity of the WPM and the ENSO varied greatly ($r_{niño-WPM(pr)} = -0.24 \pm 0.4$). While the CSIRO-Mk3.5 ($r_{niño-WPM(pr)} = -0.85$), GFDL-CM2.0 ($r_{niño-WPM(pr)} = -0.82$), ECHO-G ($r_{niño-WPM(pr)} = -0.80$) and MRI-

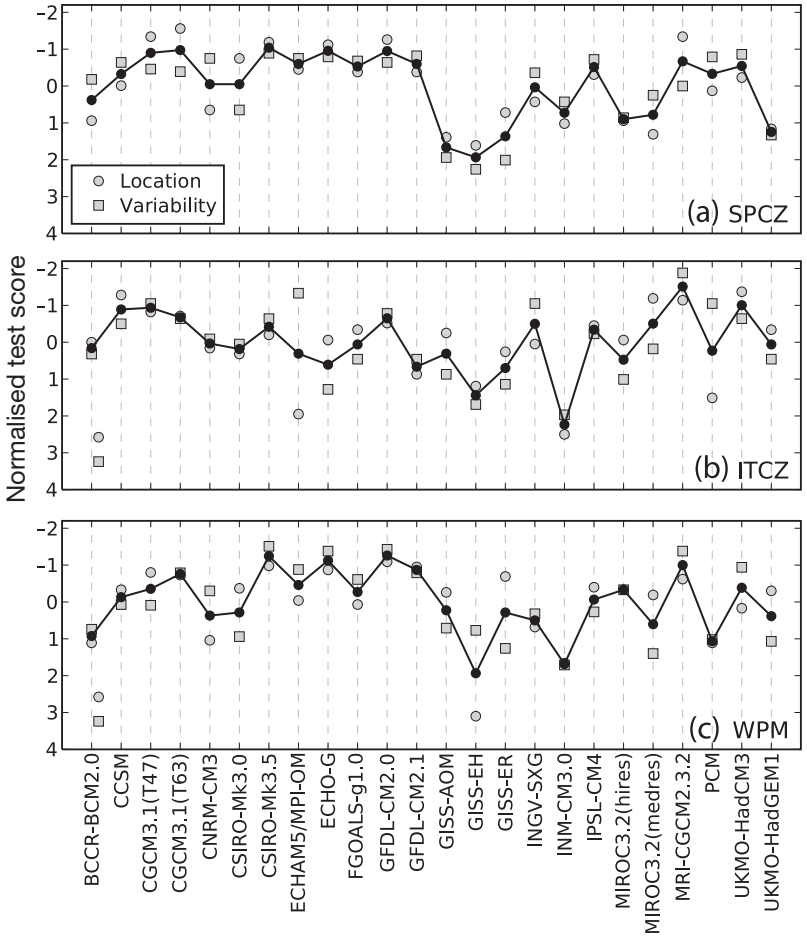

Fig. 5. Normalised test scores (increasingly negative scores indicate better model performance) for each of the analysed climate features: (a) South Pacific Convergence Zone (SPCZ), (b) Intertropical Convergence Zone (ITCZ) and (c) the West Pacific Monsoon (WPM). (◯) metrics relating to the mean location ($r_{p,SPCZ}$, $r_{p,ITCZ}$ and $r_{p,WPM}$ respectively); (□) metrics that assess interannual variability ($r_{niño-SPCZ(lat)}$, $L_{niño-ITCZ}$ and $r_{niño-WPM(pr)}$ respectively); (–●–) average across all tests

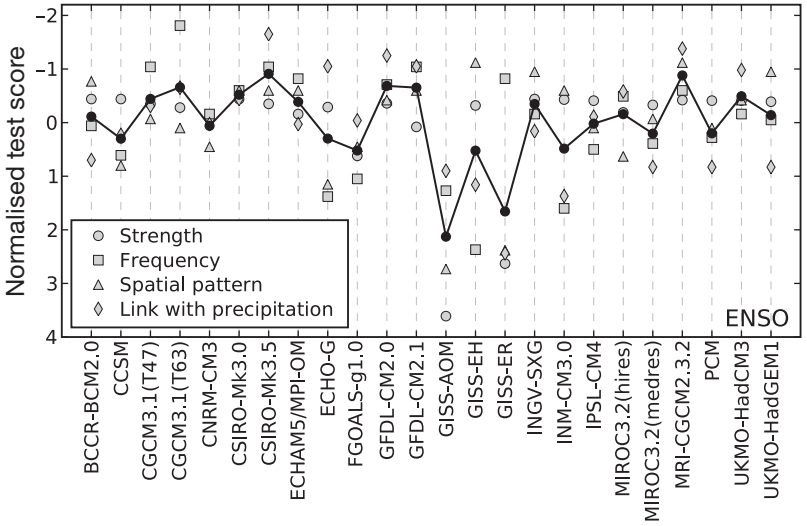

Fig. 6. Normalised test scores (increasingly negative scores indicate better model performance) for the El Niño Southern Oscillation (ENSO). (◯) $\sigma_{ratio,niño}$, indicative of the strength of ENSO events; (□) $r_{p,niño-SST}$, mean climate and spatial pattern; (△) N_{events}, frequency of ENSO events; (◇) $S_{ENSO-pr}$, link between ENSO and precipitation; (–●–) average across all tests

CGCM2.3.2 ($r_{niño-WPM(pr)}$ = −0.80) models had correlations that were very similar to the CMAP and GPCP datasets ($r_{niño-WPM(pr)}$ = −0.84 for both), 9 of the models had a positive correlation, indicating that they failed to capture this inverse relationship. The normalised statistics (Fig. 5c) revealed the BCCR-BCM2.0, GISS-EH, INM-CM3.0 and PCM models as particularly poor performers, while it was again difficult to identify a clearly superior subset of models.

3.3. Climate variability

There was a large spread in results related to the model simulated strength of ENSO events, varying from pronounced underestimation ($\sigma_{ratio,niño}$ = 0.23 for GISS-AOM) to overestimation ($\sigma_{ratio,niño}$ = 2.51 for FGOALS-g1.0) of the observed temporal variance in the Niño 3.4 index ($\sigma_{ratio,niño}$ = 1.11 ± 0.53). Similarly, the number of modelled El Niño and La Niña events over the 1950–1999 period ranged from 0 (GISS-AOM) to 37 (ECHO-G), as compared to the observed count of 23 (N_{events} = 23.3 ± 9.5). It can be seen from Fig. 6, however, that despite the large spread in model performance on these metrics, the relative ordering of the models from best to worst showed some consistency from one metric to the next.

Results were much less varied when considering the link between ENSO and precipitation ($S_{ENSO-pr}$ = 0.38 ± 0.15); however, even the high-

est model score of $S_{ENSO-pr}$ = 0.59 (MRI-CGCM2.3.2) was considerably less than value of $S_{ENSO-pr}$ = 0.94 obtained when comparing the GPCP and CMAP datasets (i.e. this value can be taken to represent the maximum possible score after taking observational uncertainty into account). Relatively little intermodel variation was also evident in the ability of the models to simulate the spatial SST pattern associated with ENSO ($r_{niño-SST}$ = 0.77 ± 0.09). As previously mentioned, majority of the models suffer from large biases in the equatorial Pacific due to the overly westward extension of the equatorial cold tongue (Section 2.4; Fig. 2); hence, it was not surprising that models with a relatively low $r_{niño-SST}$ tended to have large cold tongue biases.

The normalised statistics revealed the GISS-AOM and GISS-ER models as particularly poor performers with respect to their overall simulation of ENSO (Fig. 6). It is also noteworthy that in general, models with either very weak or very strong ENSO variability (as indicated by N_{events}; FGOALS-g1.0 is an exception) or relatively poor teleconnections between Niño 3.4 and precipitation ($S_{ENSO-pr}$) also performed poorly in the precipitation analysis (Section 3.1.2). This highlights the strong link between ENSO and precipitation in the tropical Pacific and the possible detrimental effect that a deficient model simulation of ENSO can have on related model fields.

3.4. Climate stability

The grid point averages of the PIcntrl trend magnitudes (i.e. the unforced model drift) in surface air temperature and precipitation were generally <0.2°C century^{-1} and <0.2 mm d^{-1} century^{-1} respectively for any individual model (D_{tas} = 0.10 ± 0.08°C century^{-1}; D_{pr} = 0.14 ± 0.06 mm d^{-1} century^{-1}); however, the INGV-SXG model was a pronounced outlier for both variables (D_{tas} = 0.42°C century^{-1}; D_{pr} = 0.36 mm d^{-1} century^{-1}). To put these results in perspective, the 1950–1999 multi-model grid point average (non-de-drifted) trend magnitude in surface air temperature was 1.04 ± 0.42°C century^{-1} for the 20c3m experiment. This means that model drift typically introduces an error of 9.9 ± 8.9% to the local 20c3m surface air temperature trend, while the error for the INGV-SXG model was ~44%. The latter result is particularly concerning, as the only PIcntrl data available for INGV-SXG spans the period 1760–1860, which means that drift cannot be removed from any 21st century projections arising from this model. As expected, the flux corrected models were generally associated with a smaller drift in both surface air temperature and precipitation than those without flux correction (Fig. 7a).

3.5 Climate change signal

The 1950–1999 linear grid point trends in SST averaged over the PCCSP region for the HadSST2, ERSST.v3 and Kaplan.v2 datasets were 0.78, 0.60 and 0.46°C century^{-1} respectively. In general, the models overestimated these trends, as evidenced by the multi-model average (de-drifted) trend of 0.94 ± 0.44°C century^{-1}. With regard to the average grid point error magnitude statistic used to assess the long term trends in SST, the CGCM3.1(T47), CGCM3.1(T63) and CNRM-CM3 models were pronounced outliers (E_{trend} = 1.4, 1.26 and 1.06°C century^{-1} respectively). As can be noted from Fig. 7b, the remainder of the models scored very similarly (E_{trend} = 0.30 ± 0.08°C century^{-1}). Since some models do not include all relevant forcings

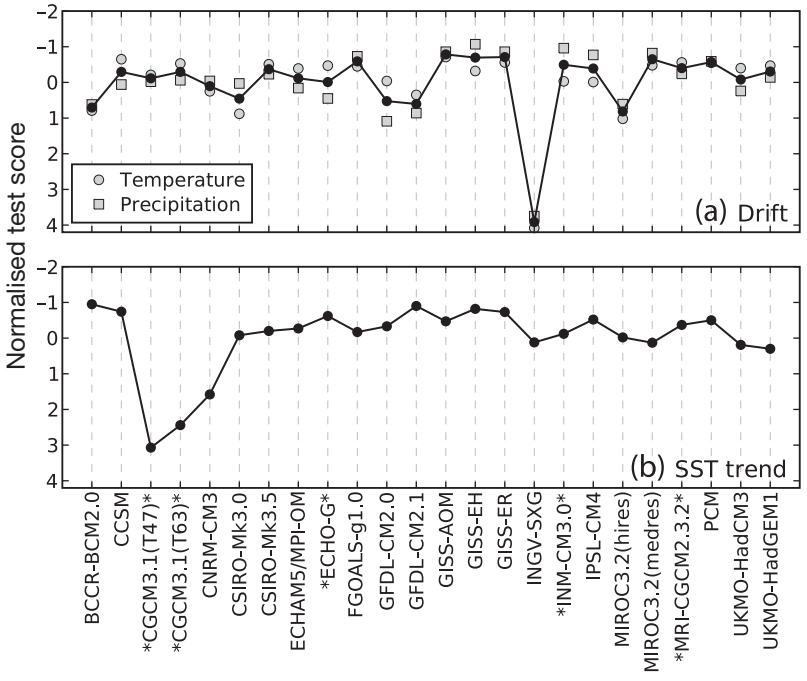

Fig. 7. Normalised test scores (increasingly negative scores indicate better model performance) for (a) drift and (b) sea surface temperature (SST) trend. For drift, (○) surface air temperature drift (D_{tas}); (□) precipitation drift (D_{pr}); connected black dots: average across both tests. For the SST trend, (–●–) average magnitude of the grid point error (E_{trend}). *Flux adjusted models

(e.g. indirect aerosols, solar and volcanic), it is important to note that any discrepancy between modelled and observed trends may only party reflect a problem in the model physics. In addition, there is substantial spread in observational estimates of long term trends in the tropical Pacific (Deser et al. 2010).

3.6. Overall model performance

A summary of overall model performance is shown in Fig. 8. It can be seen that the average normalised score across all 9 key aspects of a climate model simulation gradually rises (indicating poorer model performance) from the best performing model (MRI-CGCM2.3.2) to the 23rd ranked model (GISS-EH), before rising more sharply to the INM-CM3.0 model (the 24th and last ranked). For a given model, the average normalised score for each aspect tended to cluster around the overall average value, although there were some examples of extreme outliers. The relatively large drift associated with the INGV-SXG model, for instance, was inconsistent with the performance of this model on all other aspects. Similarly, the relatively large error in simulating long term SST trends shown by the CGCM3.1(T47) and CGCM3.1 (T63) models contrasted with their performance on the other 8 aspects. These examples serve to highlight the importance of assessing model performance across a broad range of climate aspects. A simple analysis of the mean climate would have failed to identify these deficiencies in model performance.

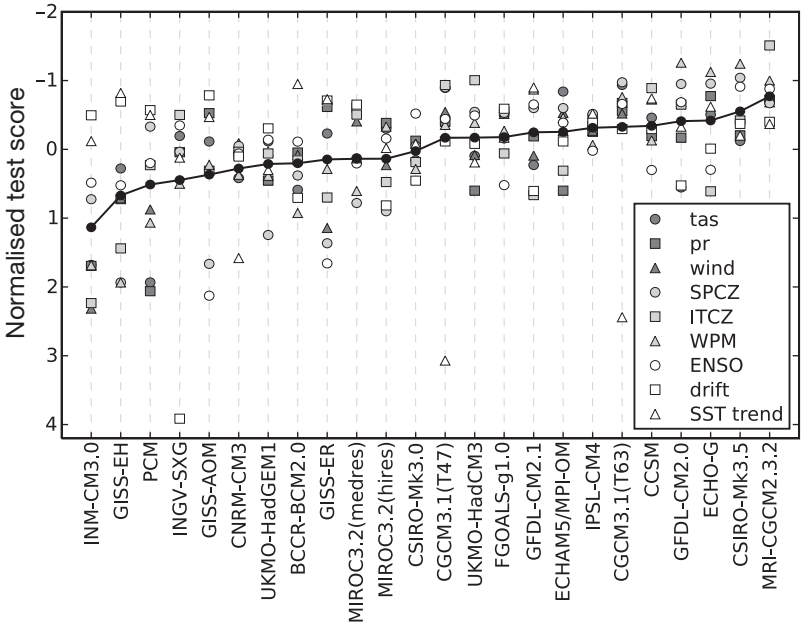

Fig. 8. Summary of the average normalised test scores (increasingly negative scores indicate better model performance) for all 9 key aspects of a climate model simulation: surface air temperature (tas), precipitation (pr), surface wind, South Pacific Convergence Zone (SPCZ), Intertropical Convergence Zone (ITCZ), West Pacific Monsoon (WPM), El Niño Southern Oscillation (ENSO), drift and sea surface temperature (SST) trends. (–●–) average across all aspects

3.7. Model elimination

In order to recommend an appropriate subset of models for use in calculating PCCSP climate projections, a degree of expert judgement was required (i.e. it was not considered possible to define an objective threshold of acceptable model performance). After careful consideration of both the projections information required by the 15 PCCSP partner countries and the model evaluation results obtained for each of the 9 key aspects of a model simulation, the following CMIP3 models are recommended for elimination in calculating all PCCSP climate projections (e.g. Perkins et al. in press), for the reasons outlined:

INM-CM3.0, PCM and GISS-EH: These models perform particularly poorly with respect to their simulation of many aspects of the present day climate over the PCCSP region (Section 3.6; Fig. 8).

INGV-SXG: This model is associated with strong drift and does not provide the required control simulation data to remove this drift from projected changes (Section 3.4; Fig. 7a).

GISS-AOM and GISS-ER: These models perform particularly poorly with respect to their simulation of the present day ENSO (Section 3.3; Fig. 6), which was considered to be a critical aspect of the PCCSP regional climate.

In addition, the following models are candidates for elimination in calculating specific projections, due to critical deficiencies related to isolated climate features:

MIROC3.2(medres) and MIROC3.2 (hires): These models are candidates for elimination in determining projections of future SPCZ activity, as they perform particularly poorly in simulating the present day characteristics of the SPCZ (Section 3.2.1; Fig. 5a).

MIROC3.2(hires): This model is a candidate for elimination in determining projections of future ITCZ activity, as it performs particularly poorly in simulating the present day characteristics of the ITCZ (Section 3.2.2; Fig. 5b).

4. DISCUSSION

As in many regional climate studies, the PCCSP will draw upon the CMIP3 data archive in determining a range of climate change projections for the Pacific island region. As a first step in this process, we present an objective assessment of the ability of the CMIP3 models to represent the 9 key aspects of a climate model simulation that are most relevant to the region and the intended PCCSP projections. This information was used to recommend certain models for elimination from the multi-model ensemble (see Section 3.7), in calculating projections for the 15 PCCSP partner countries (e.g. Perkins et al. in press).

While it may be deemed appropriate to use the results presented here to inform other non-PCCSP studies focused on the broader Pacific region, the relative performance of a climate model can in some cases vary markedly across the globe (e.g. Gleckler et al. 2008), thus limiting the wider applicability of our results. However, despite the regionally specific nature of the results, we propose that our general approach could be adapted to any region and set of climate model simulations. In fact, there are relatively few studies in the peer reviewed literature that have attempted to combine metrics of model performance pertaining to such a wide variety of climate processes and phenomena (e.g. Gleckler et al. 2008, Pierce et al. 2009, Santer et al. 2009). In order to explore the practicalities of such an approach, it would be desirable to have many different attempts documented in the literature. As discussed by Knutti (2010), the climate projection community would benefit from a larger set of proposed model evaluation methods as the basis for the assessment in the Fifth Assessment Report of the IPCC. We learn from the diversity of models, and we will learn from different ways to evaluate them.

Any model evaluation process involves making a number of subjective decisions that may require justification. In the current study, the use of a large number of model performance metrics is one such decision, given the possibility of redundancy between tests. For instance, the models were evaluated on their ability to simulate the long term mean and seasonal cycle of precipitation over the entire PCCSP region, but precipitation data were also used to evaluate the SPCZ, ITCZ and WPM. It could also be argued that the flux adjusted models should have been assessed separately from the remainder of the ensemble, while the choice of method for combining results from multiple metrics (and equally, the choice of metrics used to assess each key aspect of a model simulation) is potentially even more controversial.

For some of these subjective decisions, there is no single correct answer and a number of approaches can be justified. For example, with regard to the issue of overlap between metrics, we err on the side of ensuring that all aspects of a simulation are captured by our analysis. Similar to the process followed by Pierce et al. (2009), we may have been equally justified in retaining only the leading modes from empirical orthogonal functions constructed from the test score array. This would provide the most compact and orthogonal set of metrics for differentiating the models from each other (i.e. there would be very little overlap), at the expense of describing the absolute model skill (i.e. the selected set of metrics would be unlikely to capture all aspects of the simulation, and may therefore give the false impression of a large spread in model performance). The fact that precipitation data were used extensively in our evaluation might also mean that our analysis was biased towards this variable; however, this may be justified considering that precipitation is highly relevant to societal impacts in the Pacific. Finally, we did not discriminate between flux and non-flux adjusted models since they are treated this way in most projection studies; however, arguments supporting their separation may be equally justified owing to the 'unfair' advantage that flux corrected models have on some tests of performance.

One of the benefits of the method of normalising and ultimately combining test scores is that it does not assume uniform spread in model performance (see Section 2.7). However, since the scores on each metric are transformed to have a mean of 0.0 and an intermodel variance of 1.0, the method implicitly assumes that the collective model performance on each test is approximately equal, both in terms of the mean performance and the distribution of the models about that mean. A hypothetical situation where this assumption may be invalid is when the CMIP3 models have an impressively high average pattern correlation for a particular variable (e.g. $r_p = 0.92 \pm 0.05$) but are far less impressive in simulating the observed spatial variance (e.g. $\sigma_{ratio,x} = 1.40 \pm 0.11$). Using the methods applied in the current study, average model performance on these 2 metrics would be associated with a similar normalised test score. However, it could be reasonably argued that the normalisation process should penalise the spatial variability scores more harshly than the pattern correlation scores. While it may be considered safe to assume that the mean and distribution of model performance on each metric would be relatively similar within any particular aspect of a model simulation (e.g. ENSO, drift), it

would be less safe to make this assumption when combining results from multiple aspects. Thus, the 'overall' results presented in Fig. 8 should be interpreted with particular caution.

Avoidance of these implicit assumptions would require a normalisation method that is independent of the performance of the remainder of the ensemble. For instance, the recently updated reliability ensemble averaging method (Xu et al. 2010) normalises test scores using an estimate of the natural (observed) variability of the statistic being measured. The shortcoming of this method, however, is that it requires an observational record of sufficient temporal extent (Xu et al. calculated the difference between the maximum and minimum values of the 20 yr moving averages in a century long observational record). While some of the observational datasets used here have sufficient temporal extent to determine a representative value of the natural variability, most only provide data from 1979 onwards. In addition, there are many metrics for which a relevant measure of natural variability is not obvious, which means that the types of metrics that can be used for this method are also limited. Thus, our preference was to instead pursue the intermodel SD approach outlined in Section 2.7.

Unlike these methodological decisions regarding redundancy between metrics, flux corrected subsets and test score normalisation, there exist other decisions where it is much easier to distinguish between a 'correct' and an 'incorrect' choice. For instance, in objectively evaluating the ability of the models to simulate each of the 3 climate features (SPCZ, ITCZ and WPM) and ENSO, we used relatively simple metrics. In order to determine whether these metrics are able to sufficiently capture the ability of the models to simulate these phenomena, it is necessary to consider the more detailed analyses presented in the literature. In particular, these detailed analyses consider a vast array of diagnostics such as maps, time series and distributions that do not always lend themselves to objective interpretation, but from which important inferences can be drawn (i.e. attempts to derive objective metrics from diagnostics often result in a condensation of the original information). With regard to the SPCZ, Brown et al. (2011) found that the GISS-AOM, GISS-ER, MIROC(medres) and MIROC(hires) models failed to simulate a distinct SPCZ in the austral summer. The average normalized test scores calculated in the present study for the SPCZ ranked these models as 23rd, 22nd, 19th and 20th out of 24 models respectively, indicating that our simple metrics did a relatively good job of identifying the deficiency in these models.

Similarly, in a detailed analysis of ENSO, van Oldenborgh et al. (2005) concluded that the GISS-AOM and GISS-ER models show no ENSO-like variability, which is consistent with the identification of these models as poor performing outliers in the present study. Slight differences in the relative order of ENSO related model performance can be identified between this study and others documented in the literature (e.g. AchutaRao & Sperber 2006, Joseph & Nigam 2006); however, these may partly reflect the fact that our analysis focused on SST and precipitation patterns (which were considered most appropriate for a climate impacts study like the PCCSP), whereas many other studies have adopted a more dynamic approach. Thus, we would contend that our ENSO results represent a plausible ranking of the CMIP3 climate models, which is clearly able to differentiate particularly poor performing models from the remainder of the ensemble.

When interpreting the results from any model evaluation, there is an obvious need to be cognisant of the subjective decisions made, as different choices will cause slightly different results. However, in practice, the specific details of these decisions (provided they can be justified as reasonable) may in some cases have very little influence on the ultimate selection of models for inclusion into the multi-model ensemble. For instance, it can be seen from the results presented that while it was difficult to determine a clearly superior subset of models for any given test, poor performing models were relatively easy to identify and would be likely to show up in the results regardless of the specific methodological decisions made. In fact, as a somewhat extreme example, Gleckler et al. (2008) examined CMIP3 model performance over the global tropical region (20° S to 20° N) by combining root mean square error results calculated for 26 climate model variables and identified the INM-CM3.0, PCM and GISS-EH models as particularly poor performers, consistent with the results presented here. This observation of the relative ease (difficulty) with which poor (superior) models could be identified may be partially explained by the fact that in a relative sense, a metric such as a correlation (which was used widely here) penalises poor performance more harshly than it rewards good performance (e.g. as opposed to a metric such as the explained variance, r^2). It is also consistent with a number of previous studies (e.g. Santer et al. 2009) and supports the preference of Weigel et al. (2010) and the PCCSP for simply eliminating poor performing models, while equally weighting the rest. In this sense, our

results may add more credence to the argument for equally weighting models (while eliminating poor performers), as opposed to determining individual model weights.

With regard to future work, it would be important to consider whether a more objective definition of unacceptable model performance (i.e. failure to represent key mechanisms which are indispensable for meaningful climate projections) could be achieved. For the purposes of the PCCSP climate projections, 6 of the 24 CMIP3 models were recommended for elimination based on expert judgement (Section 3.7). While there were justifiable reasons for eliminating each of these models, it would be desirable to try and remove some of the subjectivity associated with the threshold between acceptable and unacceptable performance. Data retrieval and homogenisation efforts associated with the PCCSP will also hopefully mean that future model evaluation can make comparisons against station-based observational data. In the evaluation presented here, a degree of fortuitous agreement between the climate model and reanalysis data cannot be discounted, owing to possible common biases in both climate and reanalysis models (Tebaldi & Knutti, 2007).

As previously mentioned (Section 3), it should also be noted that the method of model evaluation presented here aims to rank models or identify those that perform particularly poorly, but does not provide qualitative information about specific biases in the model climate, such as the equatorial 'cold tongue' or 'double ITCZ' biases. For model projections in regions influenced by these biases, the user will need to carefully consider the impact of such model biases. More detailed treatment of these issues is provided elsewhere in the literature, including publications arising from the PCCSP (e.g. Brown et al. 2011, Perkins et al. in press).

Acknowledgements. This research was conducted with the support of the PCCSP, a program funded by AusAID, in collaboration with the Department of Climate Change and Energy Efficiency, and delivered by the Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). We thank the PCMDI and the World Climate Research Program's Working Group on Coupled Modelling for their roles in making available the CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, US Department of Energy. More details on model documentation are available at the PCMDI website (www-pcmdi.llnl.gov). We also thank J Richmond for drafting the map shown in Fig. 1 and J. Bhend for his insightful comments on many aspects of the study methodology.

LITERATURE CITED

AchutaRao K, Sperber KR (2006) ENSO simulation in coupled ocean–atmosphere models: Are the current models better? Clim Dyn 27:1–15

Adler RF, Huffman GJ, Chang A, Ferraro R and others (2003) The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). J Hydrometeorol 4:1147–1167

Bellucci A, Gualdi S, Navarra A (2010) The double-ITCZ syndrome in coupled general circulation models: the role of large-scale vertical circulation regimes. J Clim 23: 1127–1145

Beranger K, Barnier B, Gulev S, Crepon M (2006) Comparing 20 years of precipitation estimates from different sources over the world ocean. Ocean Dyn 56:104–138

Bosilovich MG, Chen JY, Robertson FR, Adler RF (2008) Evaluation of global precipitation in reanalyses. J Appl Meteorol Climatol 47:2279–2299

Brown JR, Power SB, Delage FP, Colman RA, Moise AF, Murphy BF (2011) Evaluation of the South Pacific Convergence Zone in IPCC AR4 climate model simulations of the 20th century. J Clim 24:1565–1582

Covey C, Gleckler PJ, Phillips TJ, Bader DC (2006) Secular trends and climate drift in coupled ocean–atmosphere general circulation models. J Geophys Res 111:D03107 doi:10.1029/2005JD006009

Cravatte S, Delcroix T, Zhang DX, McPhaden M, Leloup J (2009) Observed freshening and warming of the western Pacific Warm Pool. Clim Dyn 33:565–589

de Szoeke SP, Xie SP (2008) The tropical eastern Pacific seasonal cycle: assessment of errors and mechanisms in IPCC AR4 coupled ocean–atmosphere general circulation models. J Clim 21:2573–2590

Deser C, Alexander MA, Xie S, Phillips AS (2010) Sea surface temperature variability: patterns and mechanisms. Annu Rev Mar Sci 2:115–143

Diaz HF, Markgraf V (2000) El Niño and the Southern Oscillation: multiscale variability and global and regional impacts. Cambridge University Press, Cambridge

Folland CK, Renwick JA, Salinger MJ, Mullen AB (2002) Relative influence of the Interdecadal Pacific Oscillation and ENSO on the South Pacific Convergence Zone. Geophys Res Lett 29:1643 doi:10.1029/2001GL014201

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res 113:D06104 doi:10.1029/2007JD008972

Gruber A, Su XJ, Kanamitsu M, Schemm J (2000) The comparison of two merged rain gauge–satellite precipitation datasets. Bull Am Meteorol Soc 81:2631–2644

Guilyardi E, Wittenberg A, Fedorov A, Collins M and others (2009) Understanding El Niño in ocean–atmosphere general circulation models: progress and challenges. Bull Am Meteorol Soc 90:325–340

Held IM, Soden BJ (2006) Robust responses of the hydrological cycle to global warming. J Clim 19:5686–5699

Joseph R, Nigam S (2006) ENSO evolution and teleconnections in IPCC's twentieth-century climate simulations: realistic representation? J Clim 19:4360–4377

Kanamitsu M, Ebisuzaki W, Woollen J, Yang SK, Hnilo JJ, Fiorino M, Potter GL (2002) NCEP-DOE AMIP-II reanalysis (R-2). Bull Am Meteorol Soc 83:1631–1643

Kaplan A, Cane MA, Kushnir Y, Clement AC, Blumenthal MB, Rajagopalan B (1998) Analyses of global sea surface temperature 1856–1991. J Geophys Res 103:18567–18589

Kim HJ, Wang B, Ding QH (2008) The global monsoon variability simulated by CMIP3 coupled climate models. J Clim 21:5271–5294

Knutti R (2010) The end of model democracy? Clim Change 102:395–404

Knutti R, Abramowitz G, Collins M, Eyring V, Gleckler PJ, Hewitson B, Mearns L (2010a) Good practice guidance paper on assessing and combining multi model climate projections. IPCC Working Group I Technical Support Unit, University of Bern, Bern

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010b) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758

Lin JL (2007) The double-ITCZ problem in IPCC AR4 coupled GCMs: ocean–atmosphere feedback analysis. J Clim 20:4497–4525

Meehl GA, Covey C, Delworth T, Latif M and others (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Meteorol Soc 88: 1383–1394 doi:10.1029/2009GL037293

Onogi K, Tslttsui J, Koide H, Sakamoto M and others (2007) The JRA-25 reanalysis. J Meteorol Soc Jpn 85:369–432

Perkins SE, Pitman AJ, Sisson SA (2009) Smaller projected increases in 20-year temperature returns over Australia in skill-selected climate models. Geophys Res Lett 36: L06710 doi:10.1029/2009GL037293

Perkins SE, Irving DB, Brown JR, Power SB, Moise AF, Colman RA, Smith I (in press) CMIP3 ensemble climate projections over the western tropical Pacific based on model skill. Clim Res doi:10.3354/cr01046

Pierce DW, Barnett TP, Santer BD, Gleckler PJ (2009) Selecting global climate models for regional climate change studies. Proc Natl Acad Sci USA 106:8441–8446

Qian JH, Zubair L (2010) The effect of grid spacing and domain size on the quality of ensemble regional climate downscaling over South Asia during the northeasterly monsoon. Mon Weather Rev 138:2780–2802

Räisänen J (2007) How reliable are climate models? Tellus 59:2–29

Randall DA, Wood RA, Bony S, Coleman R and others (2007) Climate models and their evaluation. In: Solomon S, Qin D, Manning M, Chen Z and others (eds) Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407 doi: 10.1029/2002JD002670

Rayner NA, Brohan P, Parker DE, Folland CK and others (2006) Improved analyses of changes and uncertainties in sea surface temperature measured *in situ* since the mid-nineteenth century: the HadSST2 dataset. J Clim 19: 446–469

Reifen C, Toumi R (2009) Climate projections: past performance no guarantee of future skill? Geophys Res Lett 36: L13704 doi:10.1029/2009GL038082

Santer BD, Taylor KE, Gleckler PJ, Bonfils C and others (2009) Incorporating model quality information in climate change detection and attribution studies. Proc Natl Acad Sci USA 106:14778–14783

Sen Gupta A, Santoso A, Taschetto AS, Ummenhofer CC, Trevena J, England MH (2009) Projected changes to the Southern Hemisphere ocean and sea ice in the IPCC AR4 climate models. J Clim 22:3047–3078

Smith I, Chandler E (2010) Refining rainfall projections for the Murray Darling Basin of south-east Australia — the effect of sampling model results based on performance. Clim Change 102:377–393

Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). J Clim 21: 2283–2296

Stainforth DA, Allen MR, Tredger ER, Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc Lond A 365: 2145–2161

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106: 7183–7192 doi:10.1029/ 2000JD900719

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc Lond A 365:2053–2075

Trenberth KE (1997) The definition of El Niño. Bull Am Meteorol Soc 78:2771–2777

Uppala SM, Kallberg PW, Simmons AJ, Andrae U and others (2005) The ERA-40 re-analysis. QJR Meteorol Soc 131:2961–3012

van Oldenborgh GJ, Philip SY, Collins M (2005) El Niño in a changing climate: a multi-model study. Ocean Sci 1: 81–95

Vincent DG (1994) The South Pacific Convergence Zone. Mon Weather Rev 122:1949–1970

Vincent EM, Lengaigne M, Menkes CE, Jourdain NC, Marchesiello P, Madec G (2011) Interannual variability of the South Pacific Convergence Zone and implications for tropical cyclone genesis. Clim Dyn 36:1881–1896

Waliser DE, Gautier C (1993) A satellite derived climatology of the ITCZ. J Clim 6:2162–2174

Wang B (2006) The Asian monsoon. Springer, Berlin

Watterson IG (1996) Non-dimensional measures of climate model performance. Int J Climatol 16:379–391

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multi-model climate projections. J Clim 23:5162–5182

Whetton P, Macadam I, Bathols J, O'Grady J (2007) Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. Geophys Res Lett 34:L14701 doi:10.1029/ 2007GL030025

Xie PP, Arkin PA (1997) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull Am Meteorol Soc 78:2539–2558

Xu Y, Gao XJ, Giorgi F (2010) Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections. Clim Res 41:61–81

Yin XG, Gruber A, Arkin P (2004) Comparison of the GPCP and CMAP merged gauge–satellite monthly precipitation products for the period 1979–2001. J Hydrometeorol 5:1207–1222

Zhu YF, Chen LX (2002) The relationship between the Asian/Australian monsoon and ENSO on a quasi-four-year scale. Adv Atmos Sci 19:727–740

Editorial responsibility: Peter Gleckler, Livermore, California, USA

Submitted: November 16, 2010; Accepted: July 1, 2011
Proofs received from author(s): October 7, 2011