Australian Meteorological and Oceanographic Journal 63 (2013) 181–190

A skill score based evaluation of simulated Australian climate

Ian G. Watterson, Anthony C. Hirst and Leon D. Rotstayn
Centre for Australian Weather and Climate Research (CAWCR),
a partnership between CSIRO and the Bureau of Meteorology, Aspendale, Australia

(Manuscript received July 2012; revised December 2012)

The three Australian submissions to CMIP5, based on two versions of the ACCESS coupled climate model and the upgraded CSIRO model Mk3.6, are evaluated using skill scores for both global fields and for features of Australian climate. Global means of surface air temperature and precipitation are similar to those from observational datasets, except for a cool bias in Mk3.6. The agreement of the climatological seasonal mean fields for these and eleven other quantities, as measured by the *M* score, is comparable to the Australian CMIP3 models, in the case of Mk3.6, but mostly improved on by both ACCESS1.0 and ACCESS1.3. The ACCESS mean sea-level pressure and winds are notably better. An overall skill score, calculated for both global and Australian land, shows that the ACCESS models are among the best performing of 25 CMIP5 models. ACCESS1.0 improves slightly on the U.K. Met Office submissions. A suite of tests developed for the CAPTIVATE project is applied, and again ACCESS1.0 matches the U.K. Met Office reference model. ACCESS1.3 has improved cloud cover and a better link between equatorial sea surface temperatures (SSTs), using the Pacific-Indian Dipole index, and Australian rainfall. Mk3.6 has excessive variability in the SST index. In all, both versions of ACCESS are highly successful, while the ensembles of simulations of Mk3.6 make it also a very worthwhile submission to CMIP5.

Introduction

Over the past five years, the Centre for Australian Weather and Climate Research (CAWCR) has developed the new ACCESS coupled climate model, which uses as its atmospheric component the U.K. Met Office's Unified Model (Davies et al. 2005). Over this time, methods have been developed for objectively assessing the performance of the model. In late 2010 these were used in the Met Office Hadley Centre's (MOHC) project 'CAPTIVATE' (standing for Climate Processes, Variability and Teleconnections; see Scaife et al. 2011), for objectively evaluating the Australian climate of versions of the Hadley Centre's experimental model, HadGEM3, that also incorporates the Unified Model. A suite of tests was developed, along with a summary chart in colour-coded 'traffic light' form to highlight model improvements. The tests and results were described by Watterson et al. (2011). The main purpose of the present study is to apply these same tests to the three Australian models submitted to CMIP5 (Taylor et al. 2012). These are the CSIRO-BOM ACCESS1.0 and ACCESS1.3 models, described in this issue by Bi et al. (2013), and the CSIRO-QCCCE submission CSIRO-Mk3.6 (Rotstayn et al. 2012). In the CAPTIVATE study, model results were compared to those from an earlier present day climate simulation of the standard HadGEM2-AO model (HadGEM2 Development Team; Martin et al. 2011) and we continue to use it as the 'reference model' with regard to the colour-coded grading of skill. For this study, we include results for two model versions submitted by MOHC to CMIP5 (HadGEM2-CC and HadGEM2-ES) in a simpler, previously-used, skill analysis for both the globe and Australia applied to a total of 25 CMIP5 models.

The models that are the focus of the study are briefly described next, along with means of precipitation (pr, following the CMIP5 nomenclature) and surface air temperature (tas). Simple skill assessments are then presented for 13 global quantities from the three Australian models, and for three variables (tas, pr and psl (MSLP)) from an ensemble of CMIP5 models. The CAPTIVATE tests are then applied, targeting features of the mean state climatology (CLIM, retaining the short names used extensively in the project), variability (VAR) and teleconnections (TELE) that are important to Australian climate. A further assessment of the CMIP5 models by Smith et al. (2013) can be found in this issue.

Corresponding author address: Ian Watterson, CSIRO Marine and Atmospheric Research, Private Bag 1, Aspendale, Vic. 3195. Email: ian.watterson@csiro.au.

The models and their mean temperature and precipitation

The Mk3.6 model is an upgrade of the CMIP3 model Mk3.5 (Gordon et al. 2010), with the main improvements being the inclusion of an interactive aerosol scheme and a new radiation code (see Rotstayn et al. 2012 for details). The resolution is unchanged, with 18 levels in the vertical, and a horizontal grid spacing of approximately 1.9°. The ocean is the GFDL MOM2.2 with 31 levels, on the same grid, but with halved latitude spacing.

Like HadGEM2-AO, both versions of ACCESS have an atmospheric component with a grid spacing of 1.25° in latitude and 1.875° in longitude, and with 38 levels. ACCESS1.0 uses the HadGEM2 r1.1 atmosphere, while ACCESS1.3 uses an atmosphere similar to the Met Office's subsequent Global Atmosphere 1.0, and with a different land surface scheme (see Bi et al. 2013 for details and further references). Both ACCESS models use the GFDL MOM4p1 ocean with a tri-polar grid of 1° or less (a similar resolution but different model to HadGEM2-AO).

While the four models simulate mostly similar processes in the climate system, Mk3.6 typically has simpler representations and at lower resolution. It would not be expected to be as detailed, but being highly efficient has allowed ten simulations for each of the CMIP5 experiments, compared to only one from each ACCESS model.

The four models are listed in Table 1, along with their simulated mean tas and pr over two domains. For comparison, and to give an indication of uncertainty, two observational results (abbreviated when convenient as 'Obs') are given in each case. As elsewhere, the primary result is denoted Obs-1 and the alternative Obs-2. Slightly abbreviated names for the models are also used, including 'Ref' for HadGEM2-AO.

Like those from Mk3.6, the ACCESS data are from the 'historical' simulations submitted to CMIP5 (see Bi et al. 2013), and we use averages from the 30 years (1975–2004). The averages of the ten Mk3.6 results are given in Table 1. A 30-year climatology is also used for HadGEM2-AO.

Means for the HadCRUT3 observational dataset (for 1961–1990, on a 5° grid, Jones et al. 1999) are Obs-1 for tas over the globe in Table 1 (exact matching of years is not essential, given that both unforced variability and measurement error lead to uncertainties). For Obs-2 we use the 1979–2008 means from the ERA-Interim reanalysis data (on a 1.5° grid, Dee et al. 2011). The same period is available for pr from the GPCP dataset (Adler et al. 2003) used as Obs-1 (globe). For Australia, we use Bureau of Meteorology (BoM) gridded (0.25°) monthly fields for tas and pr averaged over 1958–2001 (which matches the ERA-40 reanalysis used later). The BoM tas is the average of the daily maximum and minimum fields. The Australian means are from fields interpolated to the BoM grid (the corresponding mean for GPCP, whose field appears as a low-resolution version of that from BoM, is 1.47 mm d^{-1}.) While the ERA-Interim precipitation is generated by the reanalysis (weather) model, and not directly based on observations, Dee et al. (2011) show that it is a viable alternative over data-sparse regions. In any case it provides an interesting comparison with the climate models (see below).

Table 1. Models considered, and their simulated global and Australian mean temperature and rainfall. Observational data from two sources are also given. Here Obs-1 is HadCRUT3, GPCP (globe) and Bureau of Meteorology (Aust.). Obs-2 is ERA-Interim.

Model	*tas (°C)*		*pr mm d^{-1}*	
	Globe	*Aust.*	*Globe*	*Aust.*
Mk3.6	13.1	21.1	2.88	1.44
ACCESS1.0	14.1	22.3	3.07	1.16
ACCESS1.3	14.3	22.0	3.15	1.50
HadGEM2-AO	13.5	21.3	3.05	1.23
Obs-1	13.9	21.8	2.67	1.36
Obs-2	14.4	22.1	2.91	1.19

All model results are within 1 °C or 0.3 mm d^{-1} of one of the 'Obs'. These are levels of agreement typical of CMIP5 coupled models, and indeed the deviations are barely greater than the differences between the Obs pair. Mk3.6 is evidently cooler than observed. ACCESS1.3 appears wetter for both the globe and Australia. An indication of the uncertainty due to internal variability is given by the range across the ten Mk3.6 results. This is 0.01 (0.06) mm d^{-1} for globe (Aust) pr and 0.2 °C for both tas values.

Maps of the pr fields over Australia from the four models and the two observational fields are shown in Fig. 1. All four models simulate the basic pattern, although none resolve the orographic peaks in the southeast. These are barely evident in the 1.5° ERA-Interim field, also. The ACCESS1.0 result is marred by relative dryness in the northeast. Rainfall along the wetter coasts, including the southwest, tends to be too light in all the models. Compared to the Bureau, the climate models and ERA-Interim are mostly too dry in the subtropical west.

Simple skill assessments

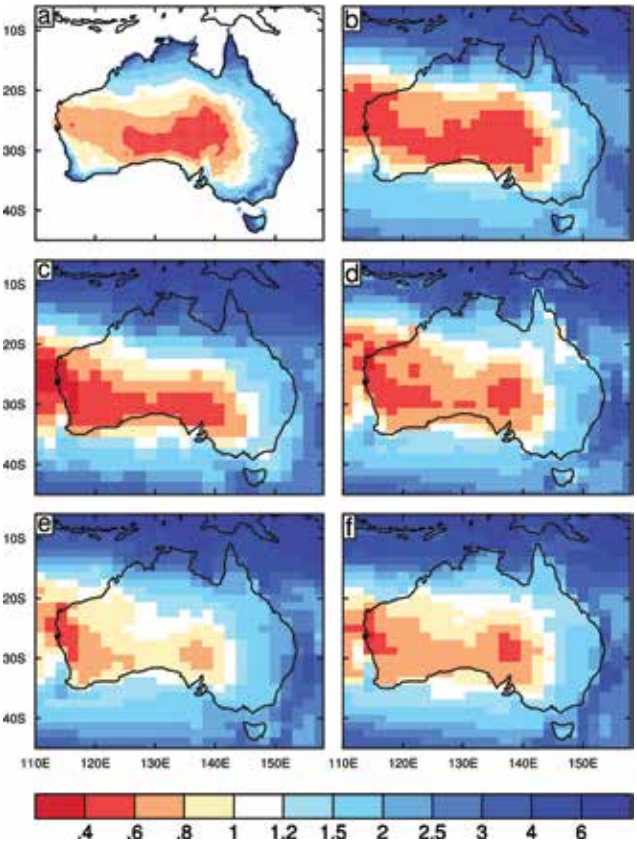
M scores

In studies of the earlier CSIRO models and of CMIP3, Watterson and Dix (1995) and Watterson (2008) used simple skill scores to quantify the agreement between simulated and observed climatological fields. The non-dimensional statistic M allows scores for different variables and seasons to be averaged. For the model field X and observed field Y, this statistic of agreement is given by

$$M = (2/\pi)\ \arcsin[1 - mse / (V_X + V_Y + (G_X - G_Y)^2)],$$

with mse the mean square error between X and Y, and V and G are spatial variance and domain mean of the fields (as subscripted). As an example, M is used to compare the pr fields in Fig. 1, with the results for similarity with Obs-1

Fig. 1. Annual mean precipitation (in mm d^{-1}) over Australia. The BoM observations are shown as (a), followed by (b) ERA-Interim. The M score for agreement over land with BoM is 0.70. The four coupled model fields are (c) Mk3.6 (run 1) (M=0.57), (d) ACCESS1.0 (0.53), (e) ACCESS1.3 (0.61), and (f) HadGEM2-AO (0.59).

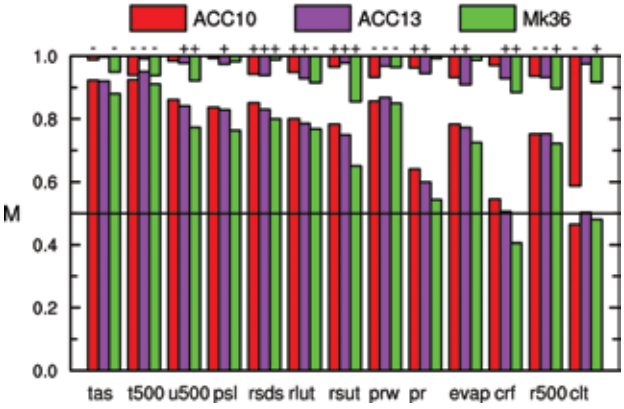


(BoM) given in the caption. Note that the score for GPCP with BoM pr is 0.75, so only a little higher than for ERA-Interim and the models. Indeed the range of the ten Mk3.6 results (0.05) is similar to the differences. Interestingly, all four models are more similar (by M) to ERA-Interim and GPCP than BoM, in part due to data resolution. Over the globe the models are more similar to ERA-Interim than GPCP.

A histogram of scores

In assessing the Mk3.0 and Mk3.5 models, Gordon et al. (2010) presented a histogram of M scores for a set of 13 quantities of interest with respect to both climate impacts (including tas and pr) and feedbacks to forced climate change (such as precipitable water or prw, and TOA upward radiative fields for LW or rlut, and SW or rsut). Also included are evaporation and SW down at the surface (rsds), and temperature, zonal wind and relative humidity at 500 hPa (t500, u500, r500). The analysis has been repeated for the new Australian models and the results are shown in Fig. 2. The averages of 1975–2004 (extended through February 2005) are again used (for Mk3.6, using run 1). For each quantity, the M score for the global domain has been calculated for each of the four seasons and the results averaged. In each case the

Fig. 2. Skill scores for 13 quantities (with CMIP5 names) over the globe, for three models: ACCESS1.0 (red), ACCESS1.3 (purple), and Mk3.6 (green). Here bars up show the M score averaged over the four seasons. The bar down shows the average of the score in which mse is replaced by the mean bias squared. The sign is shown if this bias is the same from each season. The Obs-1 used here are from ERA-Interim, except for rsut, rlut, crf, and clt (satellite-based, following Pincus et al. 2008).



scores are considered acceptable, given that uncertainties in the observations, particularly for clouds and precipitation, leads to scores between Obs-1 and an alternative set Obs-2 (see caption, and Gordon et al. 2010, p19) that are well short of the perfect match score ($M = 1$). In most cases ACCESS1.0 scores highest, followed by ACCESS1.3 and Mk3.6. The average of the 13 results for Mk3.6 (0.71) is similar to that for Mk3.5 (Gordon et al. 2010), while those for ACCESS1.0 (0.77) and ACCESS1.3 (0.76) are creditable improvements. The average for Obs-2 (relative to Obs-1) was 0.83.

The bars down in Fig. 2 give an indication of the bias in the global means. Signs are shown for cases in which the same sign of bias occurs in each season, consistent with the cases with the larger annual biases in tas and pr for globe in Table 1 (relative to ERA-Interim). The bars down also indicate that ACCESS1.3 has a similar global mean total cloud (clt) amount to Obs-1 (the ISCCP dataset), averaging 66 per cent, while ACCESS1.0 has less, only 54 per cent. Note, though, that ACCESS1.0 has a slightly better result for cloud radiative forcing (crf, see also Bi et al. 2013).

Skill scores from CMIP5 models

Only a limited set of quantities is readily available to us from the non-Australian submissions to CMIP5, so we consider the three (tas, pr and psl) that were used in two previous assessments of 23 CMIP3 models. For the 2007 Australian climate projections (CSIRO and the Bureau of Meteorology 2007) M values for Australia were averaged over the four seasons and three variables to give a single value indicative of the overall skill of a model. Watterson and Whetton (2011) used scores for the globe (restricting tas and pr to land, excluding Antarctica, with climate impacts in mind). We have used the same approach here, and the same observational data except that psl is now from ERA-Interim. Both tests

have been applied to 25 CMIP5 models (which are used also for 21st century simulations), as well as the reference. The scores (with M multiplied by 1000 for convenience) and ranks (among the 25 values) for our four models (Table 1) are included in Table 2, along with values from the top eight models in each domain. Histograms with all the models are shown in Fig. 3.

The highest scoring model for the domain globe is from MPI (as it was for CMIP3). The top score for the domain Australia is from ACCESS1.0. In both lists, the four HadGEM based models (from Australia and the U.K.) rank within the top eight (and the reference also scores highly). These scores are mostly close, however. The average of ten results is given for Mk3.6, and while below the average in both cases, they are well above the lowest value (given). The range across the ten results for Mk3.6 is rather small (0.002, 0.02 for globe, Aust.). If runs of Mk3.6 are compared against each other, then the average result for Aust is 0.89 for pr, 0.95 for psl and 0.97 for tas, an effective upper limit. Naturally, uncertainty associated with observations can also lead to differences in the scores and ranks.

Table 2. Skill score ($M \times 1000$) and rank for the Australian models and other leading CMIP5 models for the globe and Australia based on three quantities. Here the Obs tas and pr are from CRU (globe, for land only; New et al. 1999) and BoM (Aust). The Obs psl is from ERA-Interim (full globe and Aust). Also given are the lowest and average scores from 25 CMIP5 models and the score for the reference model, each with its position within the order (between two values, as for the average).

Model	*Rank*		*Score*	
	Globe	*Aust.*	*Globe*	*Aust.*
MPI-ESM-LR	1	2	788	722
CNRM-CM5	2	4	786	706
ACCESS1.0	3	1	777	727
HadGEM2-ES	4	3	777	720
GFDL-CM3	5	9	771	672
HadGEM2-CC	6	6	770	698
CCSM4	7	15	765	639
ACCESS1.3	8	7	755	690
BCC-CSM1.1	11	8	748	687
CanESM2	13	5	743	704
Mk3.6	18	18	722	617
Lowest of 25	25	25	640	553
Average of 25	14.5	12.5	736	644
HadGEM2-AO	2.5	3.5	778	715

CAPTIVATE tests

Climatology tests (CLIM)

The CAPTIVATE tests were designed to target important features of the Australian climate, including the circulation drivers of rainfall depicted by Risbey et al. (2009, their Fig. 1). The CLIM tests concerned the ten features listed in Table 3 and given the short names used in the original CAPTIVATE summary charts and in the chart that follows. These names include '1.5m T' for (surface air) temperature and 'Ppn' for precipitation. For practicality only seasonal climatologies of available variables (and pressure levels) were used (see Watterson et al. 2011 for further discussion). Thus 'monsoon onset' (Mons. Ons.) could only be assessed in a very approximate way, using seasonal low-level winds over north Australia. A proposed assessment of 'northwest cloud band' became simply a test 'Cloud' of the total cloud field (or clt, which was omitted by Watterson et al. 2011 because the reference data were unavailable). The test of the subtropical jet (SubTJet) assessed mean westerlies over eastern Australia. A further simplification was the testing of only spatial fields over suitable domains, using M ($\times$ 1000) scores.

The analysis used four domains:

- Aust.—the Australian land mass;
- Region—105–165°E, 50–0°S;
- N. Aust.—120–150°E, 20–10°S;
- E. Aust.—140–150°E, 40–15°S.

The variables (using the CMIP5 names) and domains for each test are given in Table 3. All four standard seasons were used, except for monsoon onset, for which September–November and December–February results were averaged.

The model fields are compared to the primary observational data, Obs-1, while a second set is used to indicate feasible limits to the skill of a model simulation (as quantified by the M scores), given observational uncertainty and variability. For tas and pr the data are as in Table 1 (Aust). For the other variables, Obs-1 is ERA-Interim and Obs-2 is ERA-40, except for clt (Obs-1 from ISCCP, Obs-2 ERA-Interim, as for Fig. 2). The scores for Obs-2 (versus Obs-1) are mostly above 850, which indicates strong similarity between these datasets. For pr and clt (especially), there are considerable differences between the datasets.

The scores achieved by the four models in these tests are given in Table 3. As noted by Watterson et al. (2011) the Reference model performs well (scores there differ slightly because the shorter 1989–2008 period was used for Obs-1). Here, for total cloud we use data from HadGEM2-ES as Ref.

The aim of the original CAPTIVATE tests was to see if the new MOHC model versions had achieved the skill of HadGEM2-AO. With that aim, colour grades were applied to the scores, and we use them here also. These were assigned as follows:

- Green > average (Obs 2, 850);
- Yellow > Ref. +10;
- Amber within Ref. ±10;
- Red < Ref. −10.

The grade green was ideally 'fit for purpose' or 'within the range of observations', but the above simple formula representing a good score in simulating the present climate was used. Amber is the default grade for Ref.

Given the similarity of the atmospheric and land

Table 3. CLIM tests for ten features of Australian climate, with variables and domain given, followed by M (× 1000) scores for Obs-2 and four models, each compared with Obs-1. The colours indicate how the test result compares to the reference model (HadGEM2-AO): better (yellow), similar (amber), or worse (red). Green indicates a 'fit for purpose' or good result (see text for details).

Feature (short name)	*Fields CMIP5 name*	*Domain*	*Obs-2*	*Mk36*	*ACC10*	*ACC13*	*Ref.*
1.5 m T	tas	Aust.	876	744	832	792	810
850 wind	ua, va 850 hPa	Region	907	679	760	693	731
200 wind	ua, va 200 hPa	Region	906	650	749	686	750
SLP	psl	Region	920	657	834	798	805
500 GPH	zg 500 hPa	Region	948	859	923	928	892
10 m wind	uas, vas	Aust.	820	697	760	704	775
Ppn	pr	Aust.	697	482	552	544	573
Cloud	clt	Aust.	355	089	254	406	231
Mons. Ons.	uas, vas ua 850 hPa	N. Aust.	872	668	665	557	719
SubTJet	ua, 850, 500, 200 hPa	E. Aust.	922	612	795	741	780

components it is not surprising that the scores for ACCESS1.0 are rather similar to Ref. Consistent with the Aust rank in Table 3, ACCESS1.0 scores a little higher in five features and does less well in two (including Ppn, as in the annual pr result, Table 1). Again, ACCESS1.3 is not quite as successful, although its increased cloud amount (noted earlier) is an improvement according to Obs-1. Not surprisingly, given its resolution, Mk3.6 scores are relatively low, with red for all features. Its cloud score is low because the subtropical minimum (in the zonal mean over Australia, not shown) is located some 7° too far south.

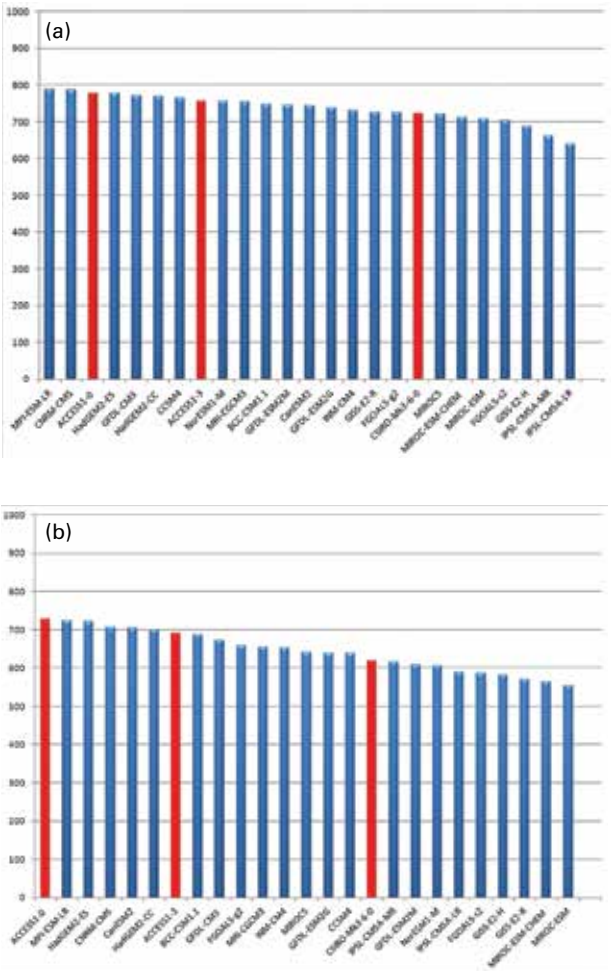
Variability assessment (VAR)

The assessment of variability and teleconnection patterns in CAPTIVATE focuses on Australian rainfall and its 'drivers', specifically tropical SST anomalies and atmospheric circulations, represented in somewhat novel ways, but also with concessions made to practicality. Time series of rainfall (over 1900–2010, here) are readily available from the BoM website (www.bom.gov.au/cgi-bin/climate/change/timeseries.cgi) for seven key regions. Some of the detail in rainfall patterns is lost by this choice, but the global climate models are not expected to simulate this detail.

The first region is the all-Australia average (All), and clearly there is some overlapping of regions. Two regions are 'north' (Nth) and 'south' (Sth), separated by the latitude 26°S. The eastern region is the four eastern states (in a model this is simplified to be east of 141°E). The southeastern (SE) region is south of 33°S and east of 135°E (see Fig. 4 for the region as represented on the reference model grid). Southwestern (SW) is the far southwest of Western Australia (in a model, the land bounded by 115–120°E, 35°S–31°S). The Murray Darling Basin (MDB) is the seventh region, simplified to 139–150°E, 37–25°S.

Watterson (2012) argued that the gradient of SSTs from the northwest to northeast of Australia is a basic potential feature of climate change (although still unknown) that

Fig. 3. Histograms of skill scores (M × 1000) for (a) the globe and (b) Australia, as in Table 2, for 25 CMIP5 models. The results from the three Australian models are highlighted.



models should simulate, as this was closely related to the 21st century change in Australian rainfall in CMIP3 simulations. This gradient was quantified by a Pacific-Indian Dipole index (PID), or alternatively, the west Pacific minus east Indian temperature difference. Analysis of HadISST observational data (Rayner et al. 2003, see the Appendix) has shown that with regard to interannual variability in the present climate, this index is related to both the ENSO and Indian Ocean Dipole (IOD) indices (see Risbey et al. 2009), and for the most part PID displays comparable correlations with the regional rainfall series. As noted by Watterson et al. (2011), tests indicate that stronger correlations are obtained with the Indian Ocean region extended to the north of Australia, with the band becoming 15°S–5°N, 85–135°E (IND, in Fig. 4). The Pacific region (PAC) remains 10°S–10°N, 150–200°E. Averages of temperature over ocean points in each domain are taken, and the simple difference forms PID, with unit K.

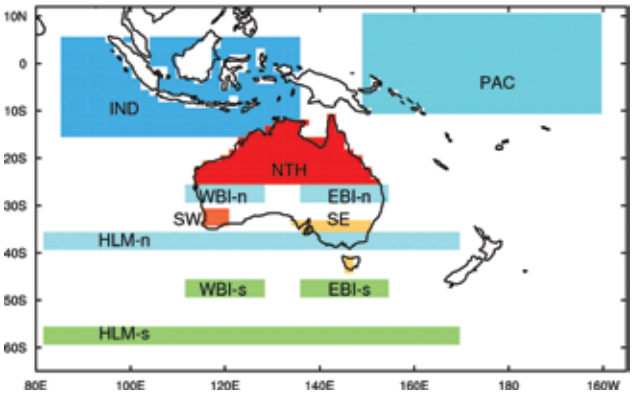
Atmospheric circulation features that are associated with Australian rainfall are here limited to a sectoral southern annular mode or 'high latitude mode' (HLM) and two sub-domain blocking indices (BIs). All three are formed from box averages of the zonal wind at 500 hPa. The HLM index is based on longitudes 80–170°E, and the box over 60–55°S minus that over 40–35°S (see Fig. 4). For the BIs, the longitude spans are 135–155°E (EBI) and 110–130°E (WBI). In each case, the latitudes of the boxes are over 50–45°S and 30–25°S, and again the index is simply the southern value minus the northern value. The EBI relates especially to southeastern regions, while the WBI relates to the southwest. These indices are similar to those used by Risbey et al. (2009) and others, except that a third latitude band (60–55°S) is not used, as this already features in HLM (see Fig. 4). For simplicity, all the tests use time series of yearly values of area averages, of either annual (calendar years) or seasonal means (of three consecutive months). Since all five cases are of interest, the results are simply averaged (following Watterson et al. 2011) in making a single final grading.

The observational data and their statistics were discussed by Watterson et al. (2011), but some basic results are included here in the Appendix. For VAR/TELE 100-year monthly series (from 1900–1999) for the three fields needed were used for the CMIP5 models. Only 50 years were available for Ref, similar to the number of years available for the reanalysis wind series.

In the VAR tests we consider the variability of rainfall in the seven regions and the variability of the four indices (PID, HLM, EBI and WBI) in each of five seasons (counting the annual mean as one). There is overlap in the time period, as in the regions, but ultimately we average results to get a single objective value. Further, we average the BI's for this presentation.

The only statistic used is the standard deviation (SD) of each series. For rainfall, we consider the SD as a percentage of the mean rainfall. To focus on the contribution from year-to-year variability the SD is calculated from de-trended series. De-trending removes part of the variation forced through global warming in both the observations and CMIP5 simulations.

Fig. 4. Some regions used in the VAR and TELE assessments, on the reference model grid. Three Australian land regions are shown (SE, SW and NTH). The tropical ocean regions (PAC, IND) are used for PID. The regions for the wind indices HLM, EBI and WBI (on the surface grid) are plotted, with s and n denoted the southern and northern components. SE extends under HLM-n.



For brevity we present only the averages, starting with the SDs given in Table 4. The overall variability of rainfall, as a percentage, is remarkably well matched by the Ref model, as well as ACCESS1.0 and ACCESS1.3. Note though that larger SDs (in per cent) for the SW region, which is drier than observed (see Fig. 1), tend to compensate for smaller SDs in some other regions, and the average would not be a reliable test statistic in itself. Variability in rainfall is excessive in Mk3.6. Overall variability in the wind indices is also quite realistic in all four models, although somewhat larger than observed for HLM, while Mk3.6 has less variability for BI. The PID SD values differ widely across the four models, with the ACCESS1.3 SD being similar to Obs, while the ACCESS1.0 and Ref values are relatively low and Mk3.6 high. Maps (not shown) of SD values at grid points show that the variability of the index is strongly influenced by the extent across the equatorial Pacific of ENSO-like sea surface temperature variability. As with Mk3.5 (Fig. 14 of Gordon et al. 2010), this is too far westward in Mk3.6.

The score statistic is simply the magnitude of the difference ('bias') between the model SD and the observed SD, for each case. We could consider only the larger SD cases, but, as with TELE below, it seems also important that the model produce small values, when they occur in Obs. For our single grade we average over all cases, and assign colours. Naturally, there is statistical uncertainty in such a mean, but this is relatively small in averages over the 35 rainfall cases, for example. A green grade is given when the bias is sufficiently small. Without an objective target, we use the same criterion value for each test as used by Watterson et al. (2011), which was simply half the largest of those results. For amber, the mean difference is to be within 20 per cent of the Ref result, allowing for statistical uncertainty in

Table 4. Variability (VAR) of rainfall (Ppn) and of three indices from observations (Obs) and models, including mean bias of the standard deviation (SD). In each case the result is an average over the five seasons (including annual). The rainfall is given as a percentage of the mean, and the average is over the seven regions as well. For BI, the average is over the EBI and WBI. The bias is calculated by averaging the absolute differences from the Obs results. The colours indicate grades as in Table 3.

	Ppn SD per cent	*Ppn bias per cent*	*HLM SD m/s*	*HLM bias m/s*	*BI SD m/s*	*BI bias m/s*	*PID SD K*	*PID bias K*
Obs	31.3	0.0	3.45	0.0	3.34	0.0	0.353	0.0
ACC10	30.1	4.3	3.96	0.69	3.19	0.40	0.215	0.138
ACC13	32.9	4.3	4.17	0.72	3.30	0.49	0.362	0.020
Mk36	41.0	10.2	4.05	0.61	3.05	0.55	0.812	0.459
Ref	31.7	4.2	3.76	0.40	3.26	0.44	0.294	0.059

both Obs and model SDs. This criterion is again subjective, but is rather consistent with the range of values found when assessing different 50-year periods from earlier ACCESS runs. Yellow (red) indicates as bias smaller (larger) than the range for amber.

The bias values and colour grades are given in Table 4. For Ref, the average bias was within 17 per cent of the average SD in each (Obs) quantity. Given that statistical uncertainty will produce positive (absolute) bias for each case, this seems very good. From the new results, only the ACCESS1.3 PID score is rated green, while the others are amber or red. Mk3.6 scores lower for Ppn and BI.

This assessment regarding equatorial SSTs is specific to PID, as Rashid et al. (2013) found that ACCESS1.0 performed at least as well as ACCESS1.3 in simulating the ENSO cycle in the eastern Pacific, although with somewhat smaller amplitude than observed.

Teleconnection assessment (TELE)

Here we focus on the links between the drivers and rainfall in the seven regions. The statistic is the correlation (r) between each pair of de-trended series, and the test based on the average bias. Only the simultaneous (zero time lag) relationship is considered. With 50 or more yearly values in each case (as in VAR), correlations need only be 0.2 or so to be of statistical significance (assuming the values are independent). We need not dwell on this, given that values would need to be larger to be of practical significance, but uncertainty is allowed for in the grading. While not the focus here, the observational results are of considerable interest and are described by Watterson et al. (2011) and in the Appendix. The correlations between drivers and seasonal rainfall are as large as –0.77 for PID and –0.68 for EBI, supporting the relevance of the tests.

Turning to the model results, we again present only the averages in Table 5, and given that both signs of r occur in the wind cases, for a representative value the magnitudes of r are averaged. (Naturally, this averaging hides some of the larger correlations that are of more practical interest). The HLM relationships are stronger overall in Ref than in Obs. The average bias is also larger for Ref than for the Australian models, despite it having the best VAR result. Using the same colour criteria as for VAR, we rate both ACCESS1.0 and ACCESS1.3 as improved, with yellow, while Mk3.6 warrants a green (although by the barest of margins). The two blocking indices also correlate rather modestly with rainfall overall, with similar averages in each model (Table 5). The model bias in r (again, averaged over both strong and weak relationships) is also similar to Ref for ACCESS1.0 and ACCESS1.3, but evidently worse for Mk3.6, which scores a red.

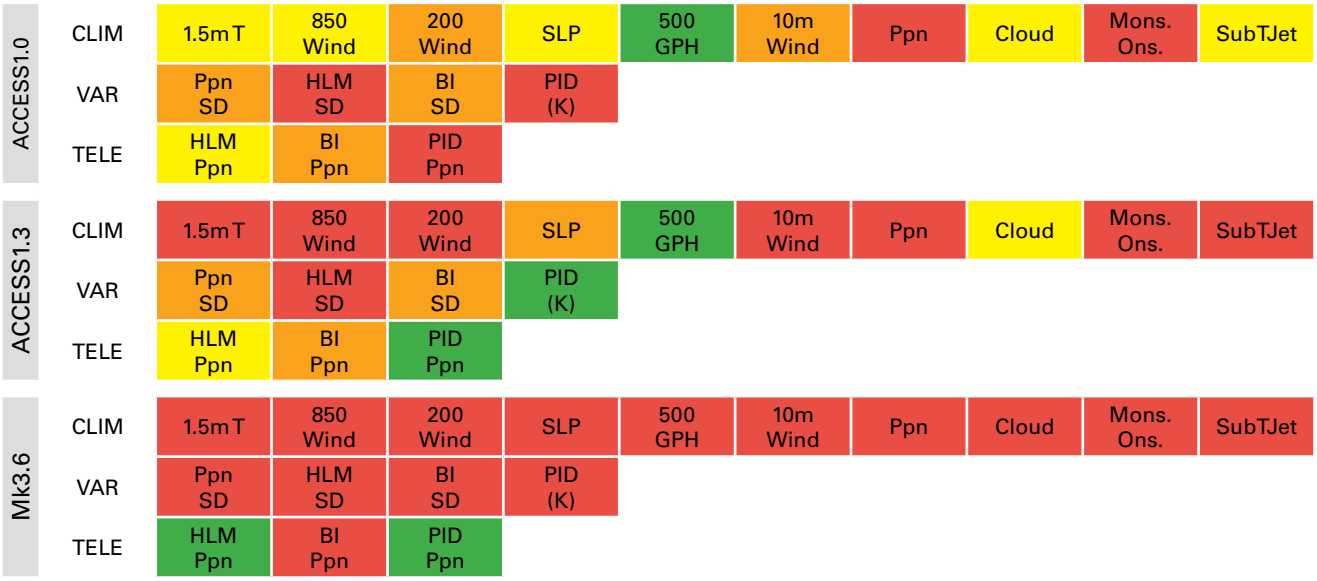
Table 5. Teleconnection (TELE) between indices and rainfall from Obs and models, quantified using the correlation coefficient r (× 100), including the bias. In each case the result is an average over the five seasons (including annual) and seven rainfall regions. For the HLM and BI cases, the values are the average of the magnitude of r. For BI, the average is over the EBI and WBI. The bias is calculated by averaging the absolute differences from the Obs results. The colours indicate grades as in Table 3.

Data	*HLM Ppn abs(r) × 100*	*HLM Ppn bias*	*BI Ppn abs(r)*	*BI Ppn bias*	*PID Ppn r × 100*	*PID Ppn bias*
Obs	21	0.0	27	0.0	–42	0.0
ACC10	30	15	20	15	–21	30
ACC13	28	13	23	19	–37	13
Mk36	28	12	19	21	–55	15
Ref	39	24	25	16	–25	19

The observed PID-rain relationship is most closely matched by ACCESS1.3. Consistent with the smaller PID SD, ACCESS1.0 and Ref have weaker correlations on average. They have a larger overall bias in r, also. Mk3.6 has stronger correlations than observed, but the mean bias is relatively small. It and ACCESS1.3 are graded green, and ACCESS1.0 red.

These grades have been based on averages of the correlation biases from all seasons and rainfall regions. An alternative would be to consider only the cases in which there is a significant relationship between the driver and rainfall. Simply excluding the cases in which the observed r has magnitude less than 0.2 produces only small differences to the values in Table 5. The only colour change would be for the HLM influence in ACCESS1.3, which improves to green.

Fig. 5. Australian regional summary charts for the CAPTIVATE tests applied to the CMIP5 models: ACCESS1.0 (top), ACCESS1.3 (middle), Mk3.6 (bottom). The order in each row matches that in the corresponding Table (3, 4 and 5), as do the colour grades.



Summary charts

In this CAPTIVATE section we have applied ten tests to assess simulated features of the climatology, four tests for the variability of rainfall and its drivers, and three tests for the associated teleconnections. Figure 5 shows summary charts for the analysis, matching those for the CAPTIVATE models shown by Watterson et al. (2011). Overall ACCESS1.0 has matched, and in some cases, exceeded, the performance of the HadGEM2-AO reference. Its main deficiency is in the variability of gradients of equatorial SSTs (and the PID index), and the associated teleconnection to Australian rainfall. The ACCESS1.3 and Mk3.6 models perform better in that test. The three models simulate the zonal wind features (HLM and BI) with considerable success. In some other tests, ACCESS1.3 ranks behind the Reference model, while Mk3.6 consistently ranks lowest.

Conclusions

The three Australian models submitted to CMIP5, ACCESS1.0, ACCESS1.3 and Mk3.6 simulate the present climate (specifically over 1975–2004) with considerable realism. The global and Australian means of temperature and precipitation are close to those observed, although Mk3.6 has a cool bias and ACCESS1.3 has a moist bias, in both domains.

Based on 'M' skill scores of the simulated seasonal climatologies the ACCESS models perform better than Mk3.6 in most variables from a suite of 13, particularly those associated with atmospheric circulation. On average, ACCESS1.0 performs slightly better than ACCESS1.3. However, substantial deviations from observations, relative to the spatial variation in the fields (as measured by M), remain in each model, particularly for some physical fields (including rainfall and cloud).

A combined skill score, for temperature, precipitation and pressure, was determined for each of 25 CMIP5 models, over both the globe and Australia. The two ACCESS models and the two MOHC submissions rank in the top eight in both cases. ACCESS1.0 ranks highest for Australia, while the MPI-ESM-LR model topped the global test.

A suite of assessments of features of Australian climate developed for the CAPTIVATE project has also been applied to our three models. These provide further evidence that the ACCESS models, particularly ACCESS1.0, has matched the performance of the reference HadGEM2-AO model. ACCESS1.3 performs better in the simulation of cloud and of variability in the Pacific-Indian Dipole index of equatorial SSTs, and the associated teleconnection to regional Australian rainfall. Mk3.6 also performs well in this, although with excessive variability in the western Pacific.

How these assessments can be used in the development of projections for Australian climate in the 21st century remains to be seen. The differences in the simulation of quantities associated with physical processes and feedbacks (latent heating, radiation, and cloud) means that differences in the response to forcing can be expected. Preliminary assessment of the climate change simulations in CMIP5, in particularly for the PID pattern and Australian rainfall, indicates that projections will continue to include a range of possibilities.

Acknowledgments

We thank all the modelling groups contributing to CMIP5. The CAWCR Earth Systems Modelling Program's Model Evaluation Team, led by Lawrie Rikus, has given ongoing support to this work. Janice Bathols and colleagues provided the CMIP5 model averages. Arnold Sullivan provided the CMIP5 histograms. Extensive comments from Ian Smith led to major improvements to the paper. In addition to the models named in Table 2, the other 14 CMIP5 models used were FGOALS (g2, s2), GFDL (ESM2G, ESM2M), GISS (E2-H, E2-R), IPSL-CM5A (LR, MR), INM-CM4, MIROC (5, ESM, ESM-CHEM), MRI-CGCM3, and NorESM1-M. This work has been undertaken as part of the Australian Climate Change Science Program, funded jointly by the Department of Climate Change and Energy Efficiency, the Bureau of Meteorology and CSIRO. This work was supported by the NCI National Facility at the ANU.

References

Adler, R.F., Huffman, G.J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S. Bolvin, D., Gruber, A., Susskind, J., Arkin, P. and Nelkin, E. J. 2003. The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol., 4,* 1147–67.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., Hirst, T., Kowalczyk, E., Golebiewski, M., Sullivan, A., Yan, H., Hanna, N., Franklin, C., Sun, Z., Vohralik, P., Watterson, I., Zhou, X., Fiedler, R., Collier, M., Ma, Y., Noonan, J., Stevens, L., Uhe, P., Zhu, H., Hill, R., Harris, C., Griffies, S. and Puri, K. 2013. The ACCESS Coupled Model: Description, Control Climate and Preliminary Validation, *Aust. Met. Oceanogr. J.*, 63, 41–64.

CSIRO and Bureau of Meteorology. 2007. *Climate Change in Australia*. Technical report, 148 pp. CSIRO, Melbourne.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc., 137,* 553–97.

Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A., White, A.A. and Wood, N. 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorol. Soc., 131,* 1759–82.

Gordon, H., O'Farrell, S., Collier, M., Dix, M., Rotstayn, L., Kowalczyk, E., Hirst, T. and Watterson, I. 2010. The CSIRO Mk3.5 Climate Model. *CAWCR Technical Report No. 021.* 61 pp. ISBN: 978-1-921605-666

HadGEM2 Development Team: Martin, G.M., Bellouin, N., Collins, W.J., Culverwell, I.D., Halloran, P R., Hardiman, S.C., Hinton, T.J., Jones, C.D., McDonald, R.E., McLaren, A.J., O'Connor, F.M., Roberts, M.J., Rodriguez, J.M., Woodward, S., Best, M.J., Brooks, M.E., Brown, A.R., Butchart, N., Dearden, C., Derbyshire, S.H., Dharssi, I., Doutriaux-Boucher, M., Edwards, J. M., Falloon, P.D., Gedney, N., Gray, L.J., Hewitt, H.T., Hobson, M., Huddleston, M.R., Hughes, J., Ineson, S., Ingram, W.J., James, P.M., Johns, T.C., Johnson, C.E., Jones, A., Jones, C.P., Joshi, M.M., Keen, A.B., Liddicoat, S., Lock, A.P., Maidens, A.V., Manners, J.C., Milton, S.F., Rae, J. G.L., Ridley, J.K., Sellar, A., Senior, C.A., Totterdell, I.J., Verhoef, A., Vidale, P.L. and Wiltshire, A. 2011. The HadGEM2 family of Met Office Unified Model climate configurations, *Geosci. Model Dev., 4,* 723–57, doi:10.5194/gmd-4-723-2011.

Jones, P.D., New, M., Parker, D.E., Martin, S. and Rigor, I. G. 1999. Surface air temperature and its variations over the last 150 years. *Rev. Geophys., 37,* 173–99.

New, M., Hulme, M. and Jones, P.D. 1999. Representing twentieth century space-time climate variability. part 1: development of a 1961-90 mean monthly terrestrial climatology. *J. Clim., 12,* 829–56.

Pincus, R., Batstone, C.P., Hofmann, R.J.P., Taylor, K.E. and Glecker, P.J. 2008. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res., 113,* D14209, doi:10.1029/2007JD009334.

Rashid, H.A., Sullivan, A., Hirst, A.C., Bi, D. and Zhou, X. 2013. Evaluation of El Niño–Southern Oscillation in ACCESS coupled model simulations for CMIP5. *Aust. Met. Oceanogr. J. 63,* 161–180.

Rayner, N.A., Parker, D.E., Horton, E.B., Folland, C.K., Alexander, L.V., Rowell, D.P., Kent, E.C., Kaplan, A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. 108,* D14, 4407 doi:10.1029/2002JD002670.

Risbey, J.S., Pook, M.J., McIntosh, P.C., Wheeler, M.C. and Hendon, H.H. 2009. On the Remote Drivers of Rainfall Variability in Australia. *Mon. Weather Rev., 137,* 3233–53

Rotstayn, L.D., Jeffrey, S.J., Collier, M.A., Dravitzki, S.M., Hirst, A.C., Syktus, J I. and Wong,K.K. 2012. Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations. *Atmos. Chem. Phys., 12,* 6377–404, doi:10.519/acp-12-6377-2012.

Scaife A.A., Copsey, D., Kendon, L., Vidale, P.-L., Martin, G.M., Milton, S., Moufouma-Okia. W., Roberts, M.J., Palmer,Walters, D.N. and Williams, K. 2011. Final Report of the CAPTIVATE model development project. Deliverable D3.2.7 (Internal Report), Met. Office Hadley Centre.

Smith, I., Syktus, J., Rotstayn, L. and Jeffrey, S. 2013. A brief assessment of Australian CMIP5 models based on rainfall and ENSO metrics. Submitted to *Aust. Met. Oceanogr. J.*

Taylor, K.E., Stouffer, R.J., Meehl, G.A. 2012. An Overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc., 93,* 485–98, doi:10.1175/BAMS-D-11-00094.1.

Watterson, I.G. 2008. Calculation of probability density functions for temperature and precipitation change under global warming. *J. Geophys. Res., 113,* D12106, doi:10.1029/2007JD009254.

Watterson, I.G. 2012. Understanding and partitioning future climates for Australian regions from CMIP3 using ocean warming indices. *Climatic Change, 111,* 903–22, doi:10.1007/s10584-011-0166-x

Watterson, I.G. and Dix, M.R. 2005. Effective sensitivity and heat capacity in the response of climate models to greenhouse gas and aerosol forcings. *Q. J. R. Meteorol. Soc., 131,* 259-79.

Watterson, I.G. and Whetton, P.H. 2011. Distributions of decadal means of temperature and precipitation change under global warming. *J. Geophys. Res., 116,* D07101, doi:10.1029/2010JD014502

Watterson, I.G., Hirst, A.C. and Sullivan, A. 2011. Evaluation of simulated Australian climate for the Hadley Centre's CAPTIVATE project. *CAWCR Research Letters, 7,* 22–30

Appendix: Observational results

The Bureau of Meteorology Australian rainfall series for 1900–2010 are used (with 111 yearly values, including summer, which includes January–February of 2011). For reference, the mean for each region (see the earlier 'Variability assessment' subsection for details) is given in Table 6, for the annual case. The standard deviation (SD) ranges from 16 per cent (SW) to 24 per cent (MDB) as a percentage of the mean in the annual case (Table 6). The peak values among the four seasons feature in Table 7. Except for the SE and SW, with a peak in winter, there is generally more rain in summer. Variability as a percentage peaks in different seasons, depending on the region. It tends to be largest in the driest season. The average of the SD values, over all 35 cases, is given in Table 4.

For SST, the HadISST data (1° grid) over 1950–2010 are used (plus Dec 1949). This provides 61 yearly values, for both the SD and the correlation with rain. The average of the SD values (five cases) is 0.35 K (Table 4). For the wind indices, reanalysis data over 1957–2010 are used. Over 1989–2010, ERA-Interim is used, and prior to that ERA-40. In the dozen years of overlap between these two reanalysis sets, the values are closely related; nevertheless, any bias in the time-mean (including that due to the different grids) in that period is added to the earlier ERA-40 data. Interestingly, the average SD for HLM is close to that of the two BIs combined (Table 4).

The observational results for teleconnections, calculated from coincident periods of the data, are of particular interest. The correlation between each regional series and the annual series of each index is shown in Table 6. The HLM wind index is positively correlated, except for the SW. This might be expected from previous results for the (all-longitude) SAM (e.g. Risbey et al. 2009), given that the drying associated with higher SAM is mostly seen in the far south. Enhanced easterlies further north tend to raise rainfall over much of the continent. The largest magnitude seasonal correlations are given in Table 7. The SW value for HLM is more strongly negative in winter, while the positive cases are stronger in summer or spring. The average magnitude of the correlation, given in Table 5, is relatively small.

Correlations for the BIs are mostly negative (Tables 6 and 7). They tend to be larger in the southern regions, for the index most adjacent (in longitude), and also in winter. The peaks of –0.7 seem typical of values for comparable BIs in other studies (e.g. Risbey et al. 2009). Substantial positive values occur in summer for the north. Conceivably, these are 'coincidental' with both the northern rainfall and the wind anomalies (well to the south) being part of an ENSO teleconnection.

For all the regions there is a negative correlation with the PID index. Except for SW it is substantial, both in the annual case and the seasons. (An exception is in summer, when the east–west gradient here seems less important than the variation in overall equatorial temperatures associated with ENSO.) In each case, the peak seasonal value is in spring, which also has the largest SD of PID. The average of the *r* values, in Table 5, is notably large, indicating the overall importance of this relationship.

Table 6. Observed annual statistics for the seven rainfall regions. Shown are the mean and SD of rainfall (Ppn), and the correlation of annual values with those of the four driver indices. These are HLM (a sectoral SAM), EBI (eastern blocking index), WBI (western blocking index), PID (SST index, Pacific-Indian Dipole).

Table 6. Observed annual statistics for the seven rainfall regions. Shown are the mean and SD of rainfall, and the correlation of annual values with those of the four driver indices. These are HLM (a sectoral SAM), EBI eastern blocking index, WBI western blocking index, PID (SST index, Pacific-Indian Dipole).

	Ppn Mean mm/d	*Ppn SD (per cent)*	*HLM r × 100*	*EBI r × 100*	*WBI r × 100*	*PID r × 100*
All	1.24	17	16	–9	11	–55
East	1.67	21	18	–22	–4	–63
Nth	1.41	20	10	5	25	–45
Sth	1.05	17	23	–33	–17	–64
SE	1.72	17	7	–51	–33	–55
SW	1.86	16	–19	–21	–39	–28
MDB	1.29	24	26	–30	–15	–63

Table 7. Observed seasonal statistics for the seven rainfall regions and their drivers (as in Table 6). The season with the largest magnitude result is indicated, along with the result. Code 'S' summer (DJF), 'A' autumn (MAM), 'W' winter (JJA), 'P' spring (SON).

	Ppn Mean mm/d	*Ppn SD (per cent)*	*HLM r × 100*	*EBI r × 100*	*WBI r × 100*	*PID r × 100*
All	S 2.31	P 37	P 26	S 47	S 47	P –76
East	S 2.92	P 39	P 26	W –34	S 32	P –69
Nth	S 3.33	W 57	W 21	S 50	S 51	P –77
Sth	S 1.14	A 33	S 43	W –48	W –54	P –67
SE	W 2.06	S 33	S 37	W –68	W –67	P –58
SW	W 3.62	S 67	W –36	P –37	W –61	P –24
MDB	S 1.56	A 48	S 37	W–41	W –39	P –61