South Eastern Australian Climate initiative

Final report for Project 2.2.5P

Document and assess methods for generating inputs to hydrological models and extend delivery of projections across Victoria

Principal Investigator: Ian Smith
Co-Investigator: Francis Chiew

CSIRO Land and Water
Ph: 02 6246 5751
seaci@csiro.au
http://www.seaci.org

Australian Government
Department of Climate Change
Bureau of Meteorology

MURRAY–DARLING BASIN AUTHORITY

State Government Victoria
Department of Sustainability and Environment

National Research FLAGSHIPS
Water for a Healthy Country
CSIRO

Abstract

There is a need to better identify methods for selecting GCM results for use in regional impacts studies. It builds on an approach adopted by several published assessment studies, the paper by Smith and Chandler (2009) and discussions held at the "GCM Selection for Regional Studies Workshop" held at the Bureau of Meteorology, 23-24 October, 2008.

Project Objectives

The objective of this project is to document and assess methods for generating inputs to hydrological models and extend delivery of projections across Victoria.

Significant research highlights, breakthroughs, and snapshots

- This project has summarized a number of GCM assessments which compare features of the simulated climate from 22 CMIP3 GCMs with available observations. The result is a ranking of the GCMs which indicates some consistently well performed GCMs and some consistently poorly performed GCMs. When we consider the projected changes in annual rainfall for three Australian regions from all the GCMs, we find little evidence that rankings provide much discrimination for Northern Australia and south-eastern Australia (at least for the current GCMs). However, it is apparent that the better performing GCMs indicate a drier future for south-west Western Australia than the remainder.

- On a practical level, it is recommended that researchers:

    - Consider excluding the results from the poorer performing GCMs. There is enough evidence to indicate that up to nine GCMs could be excluded on the basis of this study. There is very little reason to include the results from at least one (GISSEH).

    - Plot projected changes as a function of GCM rankings (as done in Figure 1) in order to detect evidence of clustering. This can potentially lead to less uncertain results (i.e. a smaller range).

Progress on meeting objectives

1. Review the work on downscaling techniques and write an overview report that describes the positives and negatives of the various techniques and recommends which techniques to use in which circumstances.

This was not completed during the duration of the project. Reviews are already available elsewhere, and modelling experiments to assess and compare the techniques are planned for Phase 2 of SEACI.

2. Write a guideline on how best to get future climate data to drive hydrological models based on current climate change science and hydroclimate.

A report has been prepared (see attachment) which provides a guide to GCM selection for use in regional climate change studies. This has been delivered to two CSIRO sustainable yields projects (McVicar et al., 2009; Post et al., 2009).

3. Provide the necessary climate data as inputs into runoff models across the MDB and southern Victoria to provide a suite of updated rainfall, temperature, evaporation, and runoff

projections at 5 km resolution at annual and seasonal timescales. Provide mapped outputs at the 5 km resolution and also present results as averages across appropriate spatial units determined in consultation with water resources managers (e.g. averages for upper, middle and lower river basins for Victoria and the Catchment Management Authorities).

Due to time limitations, no new projections were developed during the duration of this project. 5 km climate and runoff projections for the Murray-Darling Basin and Victoria are already available in www.climatechangeinaustralia.gov and from Phase 1 of SEACI. Improved projections with more research will come from Phase 2 of SEACI.

Summary of links to other projects

Recommendations for GCM selection have been provided to the North Australia Sustainable Yields Project (McVicar et al., 2009) and the Tasmanian Sustainable Yields Project (Post et al., 2009). The review of existing methods is a precursor to the improved methods to be developed under Phase 2 of SEACI.

Publications arising from this project

Chiew FHS, Kirono DGC, Kent D and Vaze J (2009) Assessment of rainfall simulations from global climate models and implications for climate change impact on runoff studies. MODSIM 2009 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australian and New Zealand, Cairns, July 2009

Acknowledgement

This work was funded by the South Eastern Australia Climate Initiative.

GCM selection for regional studies - Ian Smith

1. Introduction

This report is based on a need to better identify methods for selecting GCM results for use in regional impacts studies. It builds on an approach adopted by several published assessment studies, the paper by Smith and Chandler (2009) and discussions held at the "GCM Selection for Regional Studies Workshop" held at the Bureau of Meteorology, 23-24 October, 2008.

In addressing the problem of GCM selection, it seems logical to assume that a necessary condition for utilizing climate change results from a GCM is that it provides a credible representation of the present day climate. Recent studies have indicated that some GCMs suffer relatively large errors associated with simulations of present day climate and that these raise doubts about the quality of their results for climate change, particularly at regional scales. Whether these particular results should be included in any sample used to generate climate change projections is a decision for others. However, there is a definite need for some form of overall assessment which can assist researchers in their selection of GCMs. The aim here is to provide just such an assessment so that researchers can chose to reject or weighted model results.

2. Not all available GCM results are equal

Despite the claim that global climate models are based on the equations of motion and physics of transport and exchange of heat, momentum and water, this does not guarantee reliability. The CMIP3 data set contains the results from over 20 different GCMs which were used to provide climate change simulations for the IPCC Fourth Assessment Report (AR4) (IPCC, 2007). These results supersede those generated for the Third Assessment Report (AR3) (IPCC, 2005). While the more recent GCMs represent an overall improvement over the AR3 GCMs, it is also true that the CMIP3 sample contains some results from GCMs which can be identified as inferior to some of those of the earlier AR3 GCMs. Therefore, the idea that the CMIP3 sample represents the optimum sample for producing climate change projections, while desirable, is not strictly true. However, the climate change community generally assumes this to be the case since they invariably ignore the results from the AR3 models in favour of more recent and, it is assumed, better performing models.

Major differences between GCMs include:

- Horizontal resolution e.g. 400 km versus 125 km;
- Physics e.g. interactive sea ice versus prescribed sea ice;
- Parameterizations e.g. convection;
- Flux adjustments e.g. MRI-CGCM3.2, INMCM3.0, CGCM3.1(T47), and CGCM3.1(T63) ;
- Corrupted outputs e.g. the BCC-CM1 results (Jun et al., 2008); and
- Model flaws, which are hard to detect but are likely to be present in one form or another.

The effects of the above factors can often be detected when comparing the GCM results with observations of present day climate. In general, the GCMs provide reasonably good (and improving) simulations of global-scale average quantities (such as temperature, MSLP and rainfall). However, it must be recognized that the GCMs are developed and tuned to do exactly this. Tuning for global-scale averages does not guarantee the results at regional scales, nor for all fields. It is apparent that, when comparing the results for multiple fields at large scales, or single fields at regional scales, that the there exist some relatively poorly performing GCMs. The differences with observations can be quite large in some cases, sometimes sufficiently large to cast doubt about their appropriateness for inclusion in any climate change projections.

This is also recognized by the fact that methods have been adopted which weight some model results in favour of others depending on their level of "skill". The issue of weighting models based on performance is a controversial topic, with some arguing for no weighting on the grounds that there is no consensus concerning suitable metrics (Stainforth et al., 2007; Tebaldi and Knutti, 2007) or that, for any given metric, quite often it makes little sense to calculate a weighting factor for a poorly performing model that is greater than that of a very poorly performing model. Raisanen (2007) and Stainforth et al. (2007) recommend rejecting model results when it can be demonstrated that they suffer from large biases with respect to current climate.

Here we summarize the results from a range of GCM model assessments from various authors and indicate which GCMs can be described as consistent underperformers, irrespective of variable, region or scale of interest. It is argued that consistent failure across these assessments indicates potentially serious flaws which can be regarded as sufficient to render these GCMs inappropriate for climate change projections. Furthermore, we also provide values for various rankings and metrics which can be used to detect evidence of clustering, or model agreement.

3. Methods of assessment

Quantifying model performance can be difficult because of the range of metrics, variables, spatial scales and temporal scales of interest. It is fair to say that any measure of performance can be subjective, simply because it will tend to reflect the priorities of the person conducting the assessment. When different studies yield different measures of performance, this can be a problem when deciding on how to interpret a range of results in a different context. On the other hand, there is evidence that some models consistently perform poorly, irrespective of the type of assessment.

One method for combining the results of various models is referred to as the reliability ensemble average (REA) approach (Giorgi and Mearns, 2002, 2003) and allows for weightings which reflect model performance and which also penalizes outliers, or results that appear to be very different from the sample mean. This last step has been severely criticized (e.g. Raisanen, 2007) and the method has been refined over time in order to remove this criterion (Tebaldi and Knutti, 2007).

Reichler and Kim (2007) demonstrated that, at the global scale, the errors associated with simulations of current climate, on average, tend to reduce with subsequent generations of models. However, it is also true that the errors associated with some models from previous generations are less than the errors associated with some models of subsequent generations and that the errors associated with some models can easily be described as unacceptable. They assessed the biases of 21 models using annual average values of 14 atmospheric and oceanic variables across the globe as a guide. Their results clearly demonstrated the fact that the errors of the poorer performing models can be up to twice those of the better performing models (and, that one model in particular stands out as very much poorer than the rest).

Whetton et al. (2007) demonstrated that model performance, based on the simulation of the current climate, is relevant to the performance of simulations of the future climate. The similarity between different models was quantified by comparing simulated regional and global patterns of seasonal average temperature, mean sea level pressure (MSLP) and precipitation. Using the results from 17 models, they found that, for most extra-tropical regions of the globe, models with similar patterns for the current climate tended to yield similar change patterns of change. They found powerful cross-variable connections (e.g. current climate precipitation was the best variable for discriminating temperature change) and that comparing global patterns of current climate can be as useful for discriminating regional patterns of change as comparing regional patterns of current climate. These features can include simple long-term averages at the grid scale, spatial patterns at the continental scale, the annual cycle (based on average monthly values), interannual variability (e.g. El Nino Southern Oscillation – ENSO – events) where this is important and, finally, although not considered

here, recent long-term trends which may, or may not, be the result of greenhouse gas forcing of the global climate.

Several studies have been published that involve the development of projections for Australia and all adopt different methods. Suppiah et al. (2007) (referred to hereafter as S7) assessed the performance of 23 models with respect to how well they reproduced patterns of seasonal average temperature, mean sea level pressure (MSLP) and rainfall over the Australian continent. They reduced the sample to 15 by rejecting those models which frequently failed to meet certain root mean square error (RMSE) and spatial correlation thresholds across the four seasons. They then generated climate change projections based on the average and range of this 15-member sample without discriminating between the results.

Watterson (2008) described a method for generating projections using PDFs that allow for the weighting of model results via the M-statistic of Watterson (1996), determined from simulated and observed patterns of seasonal average temperature, MSLP and rainfall over the Australian continent. An M value between one (perfect match) and zero (no-skill) was obtained for each of 23 models for each of the three variables for each of the four seasons. The average of the 12 M skill scores for each model ranged from between about 0.3 to 0.7.

Perkins et al. (2007) assessed 14 models based on their ability to simulate daily rainfall and daily minimum and maximum temperatures for 12 regions of Australia. They focussed on the ability of the models to reproduce the frequency distribution functions of the three variables. They noted that some of the models exhibited considerable skill and that it was possible to identify relatively poorly performing models. Maximo et al. (2007) took the same approach but focussed on a single region of Australia known as the Murray-Darling Basin (MDB). They assessed 17 models and showing that some were clearly flawed, with only four recommended for use in impact assessments over this region. Pitman (private communication) used the results of the Perkins et al. (2007) assessment to exclude poorly performing models and obtained less uncertain projections of changes to daily temperatures and rainfall, even though the mean changes were not substantially different to those previously. Note that the daily data required for the assessment was only available for 16 of the 23 available models. Charles (2007) also assessed model performance over the MDB but focussed on the ability of the models to simulate both daily MSLP patterns and the seasonal cycle of monthly average MSLP. Of the 11 models and two other models assessed, four were clearly superior to the others. Smith and Chandler (2009) adopted a similar approach to Suppiah et al.(2007) but restricted their assessment to rainfall only, using RMSE and spatial correlation thresholds to stratify the performance of 22 GCMs (see Appendix A).

Finally, if we are interested in the effects of co-varying long-term changes in the atmosphere and ocean, then it makes sense to also assess models on their ability to simulate the ENSO phenomenon, which involves co-varying changes in the atmosphere and ocean over several years. Van Oldenborgh et al. (2005) considered how well models were able to realistically simulate a number of important ENSO features. Of the 19 models that were assessed, only six were judged to be acceptable. The GISS-AOM and GISS-ER models (for example) were found to exhibit no ENSO variability at all while other models had either too short an ENSO period, too small or too large sea surface temperature (SST) amplitude variations or variability concentrated in the wrong part of the Pacific.

3. Summary of GCM assessments

The GCM data set assessed here comprises results from 22 models which are available from the PCMDI web site: http://www-pcmdi.llnl.gov/ipcc/info_for_analysts.php. Officially known as the WCRP CMIP3 multi-model dataset or CMIP3 models for brevity, these are listed, in Appendix A. We have omitted the BCC-CM1 model results because of previously identified problems with these data (Jun et al., 2008). Table 2 summarizes the performance of the models based on various criteria adopted in a number of studies.

Table 2. Summary of model assessments:

	A Aus	B Aus	C Aus	D Aus	E ENSO	F North Pacific	G MDB	H MDB	I GLOBE	J NH	K SH
BCCR-BCM2.0	5	5	590	Yes		No	No		No		No
CCSM3	0	2	677	No	No	Yes	No		Yes	7	Yes
CGCM3.1(T47)	1	8	518	No	No	Yes	Yes	No	Yes	10	
CGCM3.1(T63)	1	10	478			Yes	No		Yes		Yes
CNRM-CM3	0	4	542		No	No		No	No	3	No
CSIRO-Mk3.0	1	7	601	Yes	No	No	Yes	No	No	14	Yes
ECHAM5/MPI	0	1	700	Yes	Yes	No	No	No	Yes	1	Yes
ECHO-G	0	4	632	Yes	No	Yes	Yes	No			
FGOALS-G1.0	2	2	639	No	No	No	Yes		No	15	No
GFDL-CM2.0	0	2	671	Yes	Yes	Yes	No	Yes	Yes	5	No
GFDL-CM2.1	0	2	672	Yes	Yes	Yes	No	Yes	Yes	2	Yes
GISS-AOM	1	8	564	No	No	No	Yes		No		
GISS-EH	5	14	304		No	No			No		No
GISS-ER	0	8	515	Yes	No	No	No	No	No	12	No
INM-CM3.0	1	7	627		No	No		Yes	No	8	No
IPSL-CM4	2	14	505	No	No	No	Yes		No	11	No
MIROC3.2(hires)	0	7	608		Yes	Yes	Yes		Yes		Yes
MIROC3.2(medres)	2	7	608	Yes	Yes	Yes	Yes	No	Yes	4	No
MRI-CGCM2.3.2	1	3	601	No	No	Yes	Yes	Yes	Yes	13	Yes
PCM	3	11	506		No	No			No	10	Yes
UKMO-HadCM3	0	6	608		Yes	Yes			Yes	6	Yes
UKMO-HadGEM1	0	2	674		No	No			Yes		Yes
GROUPINGS		F1			F2		F3		F4		

F1,F2,F3 and F4 represent the average failure rate within each group of columns.

Column A: Number of rainfall criteria failed (Smith and Chandler, 2009)

Column A reflects the performance of the GCMs in terms of their ability to capture key features of Australian seasonal rainfall only, according to the criteria adopted by Smith and Chandler (2009). The values represent the number of demerit points (or failures) based on RMS error and spatial correlation thresholds (see Appendix A). There are two thresholds for each season and only 10 models (see Table 2) pass both criteria in all four seasons.

Column B: Demerit points based on criteria for rainfall, temperature and MSLP (Suppiah et al., 2007)

Column B represents the assessment by Suppiah et al. (2007) and also represents demerit points (in this case the maximum is 24, comprising two metrics, three variables and four seasons). The best performing GCM in this assessment is ECHAM5 (one demerit point) while the worst are GISSEH and IPSL (14 demerit points).

Column C: M-statistic representing goodness of fit at simulating rainfall, temperature and MSLP over Australia (Watterson, 2008)

Column C shows the skill scores (or "M-statistic") calculated by Watterson (2008) and represents how well each GCM captures features of the rainfall, temperature and MSLP fields over Australia in each of the four seasons. In this case the best performing GCM is again ECHAM5 (700) and the worst GISSEH (304).

Column D: Satisfied criteria for daily rainfall over Australia (Perkins and Pitman, 2008)

Column D indicates which GCMs satisfactorily captured features of the daily temperature and daily rainfall probability distributions over 12 Australian regions according to the criteria of Perkins and Pitman (2008). Note that only 14 GCMs sets of GCM results were available for this assessment.

Column E: Satisfied ENSO criteria (Min et al., 2005; van Oldenborough et al., 2005)

Column E represents an assessment based on simulations of the ENSO phenomenon. According to the assessment of Van Oldenborgh et al. (2005) the following four models do not provide credible representations of ENSO: CCSM3, CNRM-CM3, UKMO_HADGEM1 and GISS-ER. The ECHO-G model was not included in the van Oldenborogh (2005) assessment but an assessment by Min et al. (2005) indicates that this model also has difficulty reproducing the ENSO phenomenon. In particular, the amplitude of the simulated SST anomalies is almost twice that observed while the frequency spectrum is dominated by a two-year peak compared to the observed three-seven year peak, possibly as a result of relatively poor horizontal resolution (400km) (Guilyardi et al., 2004). The ENSO performance criteria may appear to be severe but, in the case of Australia, it is difficult to argue for the inclusion of model results for several decades into the future when it has been judged that the model appears incapable of adequately simulating important changes to the climate system that occur on a time scale of just a few years.

Column F: Satisfied criteria for SST variability (Overland and Wang, 2007)

Column F indicates which GCMs satisfied criteria for North Pacific SST variability according to Overland and Wang (2008).

Column G: Satisfied criteria for daily rainfall over MDB region (Maximo et al., et al., 2008)

Column G represents the results of a similar assessment carried out by Maximo et al. (2008), but this time restricted to a single region over south-eastern Australia.

Column H Satisfied criteria for MSLP over MDB region (Charles et al., 2007)

Column H represents the results of the assessment by Charles et al. (2007), for the same region, but focusing on daily and seasonal MSLP patterns.

Column I: Below median errors for 14 variables (Reichler and Kim, 2008)

Column I represents an assessment of GCM performance over the northern hemisphere in terms of the number of below median rankings for temperature, MSLP and rainfall. Column I summarizes above and below median performing GCMs according to the assessment by Reichler and Kim (2008) in which they assessed 14 different variables at the global scale.

Column J: Below median rankings for temperature, MSLP and precipitation over NH (Walsh et al., 2007)

Column J represents an assessment of temperature, MSLP and precipitation over NH (Walsh et al., 2007).

Column K: Below median rankings for 4 variables over Antarctica and the globe (Connolley et al., 2007)

Column K represents an assessment of four variables over Antarctica and the globe (Connolley et al., 2007).

Although Table 2 provides a comprehensive overview of GCM performance against a variety of metrics, variables and spatial scales, not all GCMs were included in all the assessments.

Because the individual assessments are not completely independent, there is some degree of overlap in the results which could skew any overall assessment based on simple averaging, particularly as most of them are based on results over the Australian continent. If we group the assessments into those over Australia (Columns A, C, D and E) those over the relatively small MDB region (F and G), those over the Pacific Ocean (B and I) and those over the globe and/or hemispheres (H, J and K), we can partly account for some of this interdependence. For each geographic region we firstly calculate the average failure rate, and then form the average of these four values to provide an overall failure rate, shown in Table 3. Here the results have been ranked and the GCMs grouped into those with failure rates greater or equal to 60 per cent (highlighted in red), those with failure rates greater than 30 per cent and less than 60 per cent (highlighted in yellow) and those with failure rates less than or equal to 30 per cent (highlighted in blue). While this process is relatively crude, a similar result is achieved by simply calculating the simple average failure rate across all the columns of Table 2.

Table 3. GCM rankings based a weighted average of the failure rates in Table 2. A subset of 15 GCMs has been highlighted. (These GCMs provided daily data to the CMIP3 data base and have been used in impact studies over selected regions as part of the Sustainable Yields programs.)

GCM ID	Weighted failure rate (%) (Average of F1 F2 F3 and F4 from Table 2)
UKMO-HadCM3	0
MIROC3.2(hires)	8
GFDL-CM2.1	13
GFDL-CM2.0	20
MIROC3.2(medres)	25
ECHO-G	33
UKMO-HadGEM1	33
ECHAM5/MPI	38
MRI-CGCM2.3.2	40
CCSM3	44
CGCM3.1(T63)	50
GISS-AOM	58
INM-CM3.0	59
CGCM3.1(T47)	63
FGOALS-G1.0	63
CSIRO-Mk3	73
CNRM-CM3	75
IPSL-CM4	75
BCCR-BCM2.0	88
GISS-ER	88
PCM	89
GISS-EH	100

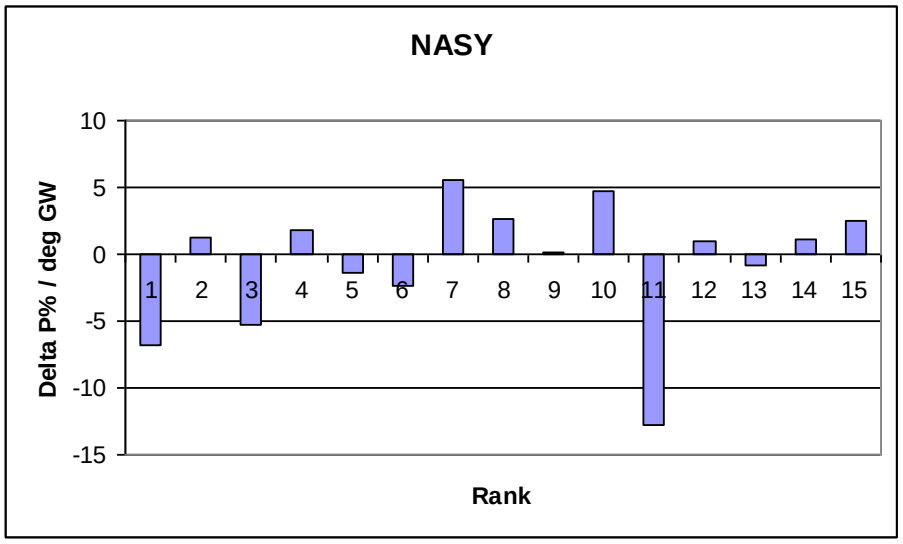
Studies suggest that the uncertainty associated with derived climate change projections can potentially be reduced by identifying and omitting the poorer performing GCMs. This effect can be partly detected by plotting the projected changes from each GCM as a function of some measure of model performance. Walsh et al. (2007) show that there is a tendency for models with smaller errors to simulate a larger greenhouse warming over the Arctic. Since several models have substantially smaller systematic errors than the other models, the differences in warming imply that the choice of a subset of models may offer a viable approach to narrowing the uncertainty and obtaining more robust estimates of future climate change in regions such as Alaska, Greenland and the broader Arctic.

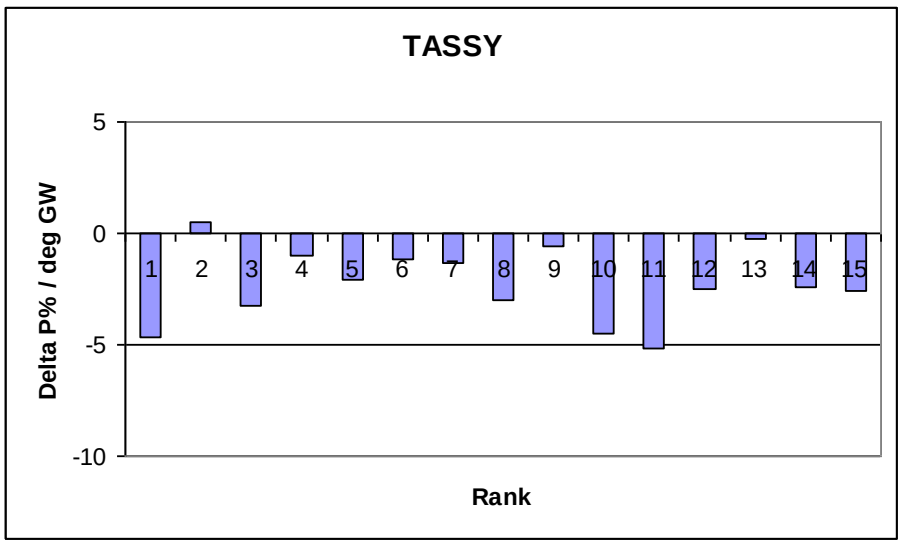
This approach can also be adopted for other regions and variables of interest. Evidence of clustering of projected changes amongst the better performing models may indicate that the poorer performing models may be a source of uncertainty, i.e. if the better performing models tend to agree, then this can be construed as evidence that GCM performance is relevant to the variable and region of interest. Otherwise there is little point in attempting to distinguish between the different results and the uncertainty remains.

We have analyzed the projected percentage changes in annual rainfall over four regions from each of 15 GCMs (for which daily data were supplied to the CMIP3 data base). These 15 GCMs are highlighted in Table 1 but do not include the top three ranked GCMs. Figure 1 provides an example of how the

model rankings can be utilized. In the case of Tasmania (TAS), the projected changes, from this sample of GCMs, are unaffected by the GCM rankings. However, there is a relatively small range of values with a mean value of close to -2 per cent.

In the case of northern Australia (NA), the mean change is close to zero, although there is a relatively large range of projected changes (-13 per cent to +6 per cent). There is an indication that the better performing GCMs yield lower values. In the case of south-west Western Australia (SW), there is a much more pronounced difference between the higher and lower ranked GCMs. The range is still relatively large (-14 per cent to +5 per cent) and the mean is close to -5 per cent. However, clustering amongst the better performing models suggests this may underestimate the drying.





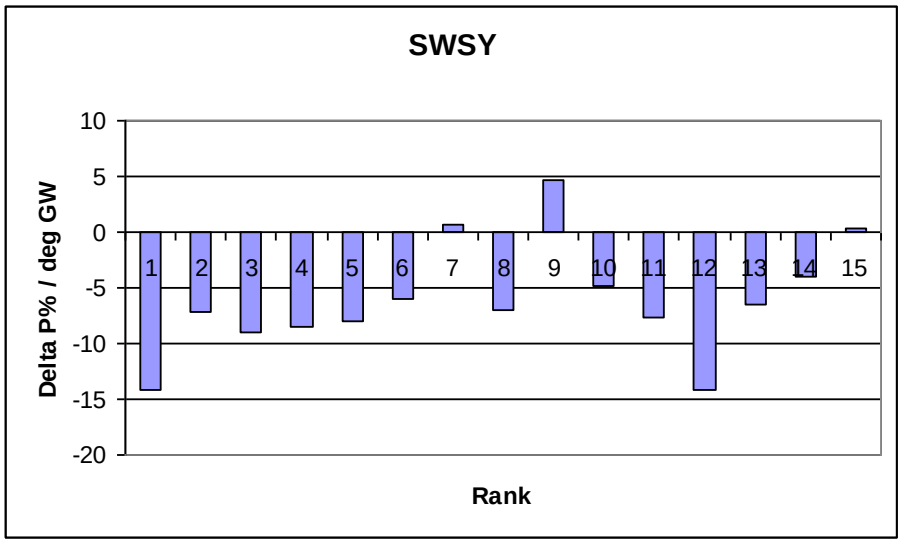


Figure 1. Projected percentage change in annual rainfall as a function of GCM ranking (best=1, worst=15, see Table 3): (a) Northern Australia, (b) Tasmania and (c) South-west Western Australia.

Finally, Figure 2 shows the effect on the projected changes to both MAM and JJA rainfall over the MDB region. The top five (out of the 15) models project decreases in MAM rainfall, while the top four project decreases in JJA rainfall. While the effect is not strong, it does suggest a stronger consensus towards decreases for this region. Note that Smith and Chandler (2009) detected a stronger effect when they analyzed the projections from all 22 GCMs.

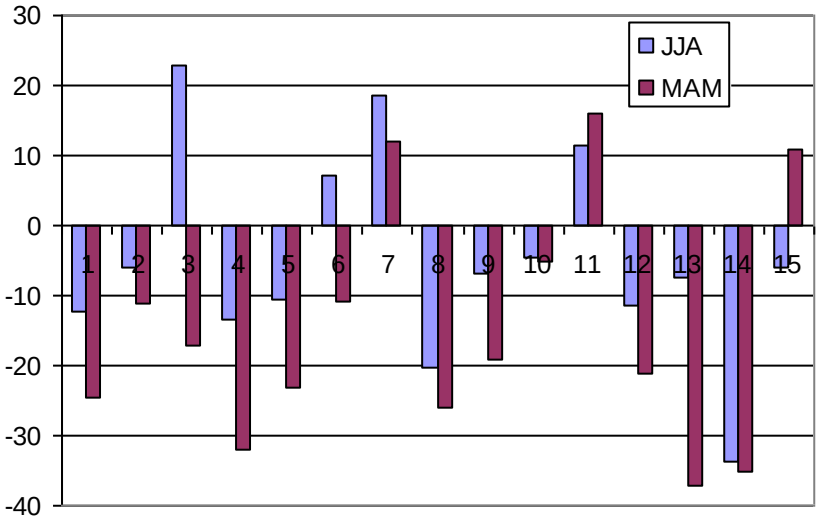


Figure 2. As for Figure 1, except for MAM and JJA seasonal rainfall for the MDB region.

6. Discussion

Is it possible that we are interpreting relative performance as an indication of absolute performance? i.e. are we unfairly penalizing GCMs for being ranked lowly when, in fact, all the GCMs may be perfectly adequate? While this simple methodology can be criticized, we argue that the relative ranking of the GCMs is unlikely to be significantly altered by the employment of more sophisticated methods.

Finally, one argument against penalizing available model results, based on subjective measures of performance, is that, we cannot really know at this point in time which models provide the best results (since we must wait for the passage of time to make this assessment). Therefore it is folly to penalize models which may, in fact, contain correct climate change signals. The answer to this argument is, simply, that at this point in time we are more interested in a practical question (that end-users typically ask of scientists):"Given a wide range of climate model results, which do you regard as being the most reliable and why?" In fact, we are not concerned so much with being proved "right" or "wrong" with regard to climate change projections at the end of the 21st century, as with providing expert advice that is both transparent, and can be acted on now. If the methodology used to sort the available information and generate advice is sound, then the important point is that due diligence is followed. There is no point in contemplating the fact that some (unspecified) apparently poor model results have been penalized may contain correct climate change signals since this would imply some particular unrecognized skill. If this cannot be recognized now, using fairly simple and logical criteria, then it is pointless arguing that all model results need to be treated equally otherwise, in an extreme sense, there would be no point in excluding the results of a random number generator.

7. Conclusions

We have summarized a number of GCM assessments which compare features of the simulated climate from 22 CMIP3 GCMs with available observations. The result is a ranking of the GCMs (Table 3) which indicates some consistently well performed GCMs and some consistently poorly performed GCMs. When we consider the projected changes in annual rainfall for three Australian regions from a sample of 15 GCMs, we find little evidence that rankings are important for northern Australia and Tasmania. However, it is apparent that the better performing GCMs indicate a drier future for south-west Western Australia than the remainder. There is also evidence that the better performing GCMs favour decreases for the MDB for the MAM and JJA seasons.

On a practical level, it is recommended that researchers:

(1) consider excluding the results from the poorer performing GCMs. There is enough evidence to indicate that up to nine GCMs could be excluded on the basis of this study. There is very little reason to include the results from at least one (GISSEH).

(2) plot projected changes as a function of GCM rankings (as done in Figures 1 and 2) in order to detect evidence of clustering. This can potentially lead to less uncertain results (i.e. a smaller range).

Acknowledgements

We thank Dewi Kirono and Janice Bathols for assistance in extracting data for the three Australian regions and the following people for very useful discussions: Richard Cresswell, Stephen Charles, Ian Macadam, Penny Whetton, Ramasamay Suppiah, Ian Watterson, Murray Peel, Andy Pitman. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making

available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. This research is supported by the South East Australian Climate Initiative (SEACI) and the Australian Climate Change Science Program (ACCSP).

References

Charles, S.P., Bari, M., Kitsios, A., and B.C. Bates, (2007) Effect of GCM bias on downscaled precipitation and runoff projections for the Serpentine Catchment, Western Australia. Int. J. Climatol., 27: 1673-1690.

William M. Connolley, W.M. and T. J. Bracegirdle (2007) An Antarctic assessment of IPCC AR4 coupled models, Geophys. Res. Let., 34, L22505, doi:10.1029/2007GL031648.

CSIRO and Bureau of Meteorology (2007), Climate Change in Australia. Technical Report, 140pp.

Giorgi, F. and L.O. Mearns (2002) Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method. J.Clim., 15(10), 1141-1158.

Giorgi, F. and L.O. Mearns (2003) Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. Geophys. Res. Lett., 30(12), 1629, doi:10.1029/2003GL017130.

Guilyardi, E., Gualdi, S., Slingo, J., Navarra, A., Delecluse, P. and coauthors. (2004) Representing El Nino in coupled ocean–atmosphere GCMs: the dominant role of the atmosphere component. J. Clim. 17, 4623–4629.

Jun, M., R. Knutti and D. W. Nychka (2008) Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?, Journal of the American Statistical Association Applications and Case Studies, (in press).

Maximo, C.C., McAvaney, B.J., Pitman, A.J., Perkins, S.E. (2007) Ranking the AR4 climate models over the Murray-Darling Basin using simulated maximum temperature, minimum temperature and precipitation. In. J. Climatol., in press. DOI: 10.1002/joc.1612

McVicar, T., Cresswell R. et al. (2009) "Climate Change 2009: Faster Change and More Serious Risks". *NASY Climate Report* (in preparation).

Min, S-K., S. Legutke, A. Hense, W-T Kwon (2005) Internal variability in a 1000-yr control simulation with the coupled climate model ECHO-G - II. El Niño Southern Oscillation and North Atlantic Oscillation. Tellus A, 57 (4), 622–640.

van Oldenborgh, G.J., Philip, Y.S., Collins, M. (2005) El-Nino in a changing climate: a multi-model study. Ocean Science, 1, 81-95.

Overland, J.E. and M. Wang (2007). Future Climate of the North Pacific Ocean. Eos 88(16), 178-182.

Pennell C. and T. Reichler (2009) On the Effective Number of Climate Models. Geophys. Res. Let.,(submitted).

Perkins, S.E., A.J. Pitman, N.J. Holbrook and J. McAveney (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions, J. Climate, 20, 4356-4376.

Post DA, Chiew FHS, Teng J, Vaze J, Yang A, Mpelasoka F, Smith I, Katzfey J, Marston F, Marvanek S, Kirono D, Nguyen K, Kent D, Donohue R, Li L and McVicar T. (2009) Climate scenarios for Tasmania. *Tasmania Sustainable Yields Project. A report to the Australian Government from the CSIRO* Tasmania Sustainable Yields Project. CSIRO, Australia.

Raisanen, J. (2007) How reliable are climate models? Tellus (2007), 59A, 2–29.

Reichler, T. and J. Kim (2008) How Well Do Coupled Models Simulate Today's Climate? Bull. Amer. Met. Soc., 89(3), 303-311.

Smith, I.N and E.Chandler (2009) Refining Rainfall projections for the Murray Darling Basin of South-East Australia: The Effect of Sampling Model Results Based on Performance. Climatic Change (in press).

Stainforth, D,A., M. R. Allen, E. R. Tredger, and L. A. Smith (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Phil. Trans. R. Soc. A, 365, 2145–2161

Suppiah, R., K.J. Hennessy, P.H. Whetton, K. McInnes, I. Macadam, J. Bathols, J. Ricketts and C.M. Page (2007) Australian climate change projections derived from simulations performed for the IPCC 4th Assessment Report. Aust. Met. Mag. 56, 131-152.

Tebaldi, C. and R. Knutti (2007) The use of the multi-model ensemble in probabilisticclimate projections. Phil. Trans. R. Soc. A , 365, 2053–2075.

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

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.

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

Appendix A.

The models used in this study. For details see:
http://www-pcmdi.llnl.gov/ipcc/info_for_analysts.php

CMIP3 ID	Approximate horizontal resolution (km)	Flux Adjusted ?
BCCR-BCM2.0	200	
CCSM3	150	
CGCM3.1(T47)	400	Yes
CGCM3.1(T63)	200	
CNRM-CM3	200	
CSIRO-Mk3.0	200	
ECHAM5/MPI-OM	200	
ECHO-G	400	
FGOALS-g1.0	300	
GFDL-CM2.0	300	
GFDL-CM2.1	300	
GISS-AOM	300	
GISS-EH	400	
GISS-ER	400	
INM-CM3.0	400	Yes
IPSL-CM4	300	
MIROC3.2(hires)	125	Yes
MIROC3.2(medres)	300	
MRI-CGCM2.3.2	300	Yes
PCM	300	
UKMO-HadCM3	300	
UKMO-HadGEM1	125	

Appendix B

Originating Group(s)	Country	CMIP3 I.D.
Bjerknes Centre for Climate Research	Norway	BCCR-BCM2.0
National Center for Atmospheric Research	USA	CCSM3
Canadian Centre for Climate Modelling & Analysis	Canada	CGCM3.1(T47)
Canadian Centre for Climate Modelling & Analysis	Canada	CGCM3.1(T63)
Météo-France / Centre National de Recherches Météorologiques	France	CNRM-CM3
CSIRO Atmospheric Research	Australia	CSIRO-Mk3.0
Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-OM
Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model and Data group.	Germany / Korea	ECHO-G
LASG / Institute of Atmospheric Physics	China	FGOALS-g1.0
US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.0
US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.1
NASA / Goddard Institute for Space Studies	USA	GISS-AOM
NASA / Goddard Institute for Space Studies	USA	GISS-EH
NASA / Goddard Institute for Space Studies	USA	GISS-ER
Institute for Numerical Mathematics	Russia	INM-CM3.0
Institut Pierre Simon Laplace	France	IPSL-CM4
Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC3.2(hires)
Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC3.2(medres)
Meteorological Research Institute	Japan	MRI-CGCM2.3.2
National Center for Atmospheric Research	USA	PCM
Hadley Centre for Climate Prediction and Research / Met Office	UK	UKMO-HadCM3
Hadley Centre for Climate Prediction and Research / Met Office	UK	UKMO-HadGEM1