

CHANGES IN THE PREDICTION OF TEMPERATURE EXTREMES OVER AUSTRALIA FOR THE 21ST CENTURY BASED ON DIFFERENT MEASURES OF MODEL SKILL.

Sarah E. Perkins *, Andy J. Pitman

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

1. INTRODUCTION

Traditionally, future climate studies have focused on changes in the mean (e.g. Allen et al., 2000; Stott and Kettlebrough, 2002; Moise and Hudson, 2008) derived as the difference between the means from future and current projections of climate models. However long-term changes in climate extremes due to global warming may impact on human and biological systems more than changes in the mean (Mearns et al., 1984; Katz and Brown, 1992; Easterling et al., 2000. Examples include human mortality (Trigo et al., 2005) and health (Woodruff et al., 2006), agriculture (Luo et al., 2005) and, particularly in Australia, drought occurrence (Nicholls, 2004).

While there is a relation in changes in temperature extremes to changes in the relative mean (Kharin et al., 2007) the magnitude of change in the extremes cannot be inferred from changes in the mean alone (Shaeffer et al., 2005). When considering the probability density function (PDF) a change in the shape and scale parameters will also influence changes in extreme values (Mearns et al. 1984; Katz and Brown, 1992, Kharin and Zwiers, 2005; Kharin et al. 2007).

Most studies focusing temperature extremes using climate models such as those used in the Intergovernmental Panel of Climate Change (IPCC) fourth assessment report (AR4) are at the global scale. Kharin et al. (2007) found that cold extremes warmed 30-40% faster than warm extremes over numerous SRES emission scenarios. Hegerl et al. (2004) showed that changes in extremes were significantly different from changes in the seasonal mean (i.e. winter for cold extremes, summer for warm extremes) for up to 66% of model grid points. While changes in the mean are well documented over Australia (Moise and Hudson, 2008) few studies have explored the changes in temperature extremes. Alexander and Arblaster (2008) found that warm extreme indices (warm nights, heat wave duration) were projected to increase and cool extreme indices (frost days) were projected to decrease irrespective of the SRES emission scenario used. Pitman and Perkins (2008) studied the change in the annual extremes for minimum temperature (T_{\min}) and maximum temperature (T_{\max}) over Australia for the 21st century.

However, the annual event is unlikely to have a profound impacts on human health, ecosystems or other biophysical systems.

This study explores the change in T_{\min} and T_{\max} extremes with a 20-year return value (i.e. events that occur once every twenty years, on average) over Australia using the AR4 models. This is done using the generalized extreme value (GEV) theory, which has been used to study climate extremes increasingly over the last decade (Zwiers and Kharin, 1998; Kharin and Zwiers, 2000, 2005; Kharin et al., 2005; Kharin et al., 2007).

Before using models to project future climate, current simulations are evaluated against the observed climate. Yet within the climate modeling community, there is no agreed "best" method, with various measures of model skill available Watterson (1996), Taylor (2001), Knutti et al. (2006), Piani et al. (2005), Shukla et al. (2006), Perkins et al. (2007) and Whetton et al. (2008).

While comparing observed and modeled means is common in model evaluation, it is clearly limited if the model is then used to project changes in rare events; a model that simulates the mean well may not simulate the tails of a distribution equally well. This implies a need for a non-means based assessment of model capacity that uses data with a temporal resolution coincident with the time scales of the specific extreme in question. This study also explores the influence different methods of model validation on multi-model ensemble projections. While recent studies looking at future changes in climate using the AR4 models have employed an evaluation method, all models were still included in the ensemble used for future projections, regardless of its evaluation performance (e.g. Moise and Hudson, 2008; Alexander and Arblaster, 2008). First, we assess the models' ability to simulate the observed 20th century mean. Second a skill-score based on Perkins et al. (2007) is used. Finally a revised 'tail-skill' which focuses on the extremes of the PDF is used. Each measure of skill is calculated at a regional scale across Australia and models that score relatively poorly are omitted from the ensemble used to project future changes in the 20-year return value. Section 2 explains the data used (observed and modeled) and methods, section 3 the results, section 4 a discussion and section 5 concluding remarks.

2. METHODS

*Corresponding author address: Sarah Perkins, Climate Change Research Centre, Mathews Building, the University of New South Wales, Sydney, Australia, 2052
s.perkins@student.unsw.edu.au

2.1 Modeled and Observed Data

All model data was downloaded from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory in the USA (http://www-pcmdi.llnl.gov/about_ipcc.php).

Daily data for T_{\min} and T_{\max} for the Climate of the 20th Century (1981-2100) and A2 emission scenario (2081-2100, hereafter 2100) was utilised for all models that had data common for all experiments (6 models for T_{\max} , 9 models for T_{\min}). See Perkins et al. (2009) for all models used, their respective resolutions and the number of independent realizations for each variable. Models with multiple realizations were concatenated to form a single sample to avoid selective sampling of any one realization, though this did not affect the results.

Daily observed T_{\max} and T_{\min} were obtained for 1178 stations from the Australian Bureau of Meteorology for 1981-2100. Their spatial distribution, homogeneity issues etc are discussed in Perkins et al. (2007) and are shown in Figure 1. Randomly removing 10% of stations or 10% of data at each station negligibly affects the resulting PDFs (Perkins et al., 2007).

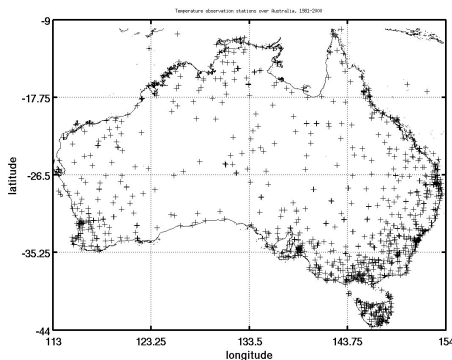


Figure 1 – Spatial distribution of observed temperature stations over Australia containing data for some time period between 1981-2000.

2.2 GEV method

Samples for the GEV are taken as the annual maxima for T_{\max} or annual minima for T_{\min} from the original dataset (in this study the sample size is 20). The GEV distribution is then fitted to the data by using either the method of maximum likelihood or the method of L-moments to estimate the distribution parameters. The method of L-moments was used to estimate distribution parameters (Von Storch and Zwiers, 1999); although this assumes stationarity of annual extremes, the more flexible method of maximum likelihood is less efficient for short samples (Kharin et al., 2007).

The GEV distribution has three parameters, location ξ , scale α , and shape k . The GEV distribution comprises three distributional families, distinguished by the tail

parameter, k , which is estimated from the sample data. When $k=0$, the GEV reduces to the Gumbel distribution; when $k<0$ the Fréchet distribution; when $k>0$ the Weibull distribution is obtained. Zwiers and Kharin (1998), Kharin and Zwiers (2000) and Kharin et al. (2005) outline in depth the GEV theory.

Once the GEV is fitted, a cumulative density function (CDF, $F(X)$) is produced, which is inverted to estimate the return period over a given time period. For maxima extremes $F(X) = 1 - 1/T$, for minima extremes $F(X) = 1/T$ (Wehner, 2004). In this study $T=20$, therefore maxima $F(X) = 0.95$ and minima $F(X) = 0.05$.

Return values were estimated both continentally and regionally. At the continental scale the GEV was fitted to each model grid box at its native resolution for both current and future scenarios and the 20-year return values were estimated. Each model was then interpolated to a common $2^\circ \times 2^\circ$ resolution and placed in the appropriate ensemble/s defined by the different measures of skill (see section 2.3). At the regional scale, for regions 2, 3 and 10 defined by Perkins et al. (2007) model samples were calculated by taking the annual maxima for each grid box within the region and concatenating to form a model-specific single sample. The sample size in this case is dependant on the models' native resolution. 20-year return values were then estimated and placed in the appropriate ensemble. In order to quantify in-sample uncertainty, 1000 non-parametric bootstrap samples were generated for each model grid box at its native resolution. Return levels were calculated for each sample to provide 90% bootstrap confidence intervals and estimates of standard errors at the regional scale.

2.3 Model Evaluation

All 12 regions defined by Perkins et al. (2007) were used in this study for each evaluation method. All models resolved multiple climate model grid squares for each region. All measures of skill were calculated over 1981-2000 for T_{\min} and T_{\max} separately. For each of the three validation methods outlined below, two ensembles were created, one consisting of the top models based on the validation method and one consisting of the bottom models. We also include an all-model ensemble to demonstrate that the weaker models influence this ensemble.

The first validation method is the absolute difference between the annual mean of a given model and the observed for each variable and region. The second validation method is the skill-score developed by Perkins et al. (2007). This calculates the cumulative minimum value of two distributions of each binned value, thereby measuring the common area between two PDFs. If a model simulates the observed conditions perfectly, the skill-score will equal one, which is the total sum of the binned values in a given PDF:

$$S_{score} = \sum_1^n \min(Z_m, Z_o)$$

where n is the number of bins used to calculate the PDF for a given region, Z_m is the proportion of values in a given bin from the model and Z_o is the proportion of values in a given bin from the observed data. Perkins et al. (2007) explore the robustness of the skill score against data quality issues.

The third validation method concentrates on the tail of the PDF (left tail for T_{min} , right tail for T_{max}). The “tail-skill” focuses on the top (bottom) 5% for T_{max} (T_{min}), based on the observed PDF. It is the weighted sum of absolute differences between the model and observed PDF proportions:

$$Tail_{skill} = \sum_1^i W_n |Z_o - Z_n|$$

The weighting is based on the number of bins in the observed tail. The weight for bin i is $i*10/n$ for $i=1, \dots, n$. The weighting is capped at this amount, and if there is a difference between the observed and modelled tails beyond where the observed tail stops, the weighting is the same as that of the last bin with an observed frequency. All bins below (above) the observed 5% limit are weighted zero. In this method a perfect skill equals zero, where there is no difference between the observed and modelled tail. Poor skill scores equal or exceed 1.0, which occurs when the model tail is much larger than the observed, i.e. the model is over estimating the magnitude of the extreme values. Figure 2 illustrates how the measures of skill differ from one another when considering two PDFs.

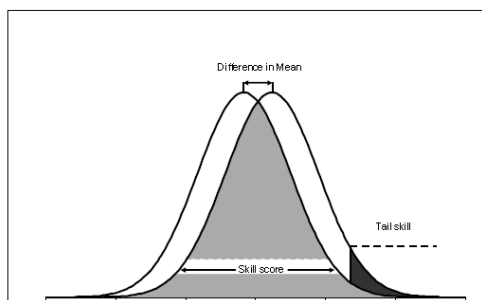


Figure 2 – Schematic diagram showing how each measure if skill differs in terms of the parts of the PDF being compared.

Models were ranked from highest to lowest for each region using each skill measure. For T_{max} the top three and bottom three models were selected and for T_{min} the top four and bottom four were selected to form the two ensembles based on the measure of skill. The difference in sample size was due to the number of AR4 models with available data. Continental results are presented as maps showing the all-model ensemble 20-year return value, and the difference between the skill

ensemble and the all-model ensemble projections. This is shown for the all-model ensemble and the top skill ensembles. Regional analysis is presented in ‘stock plots’, which show the regionally calculated minimum, mean and maximum for each ensemble based on skill and return value for regions 2, 3 and 10 defined by Perkins et al. (2007). Both the better and poorer ensembles are shown here to demonstrate the influence the poorer models have over the whole model ensembles.

3. RESULTS

Below are the regional and continental results for A2 2100. The same analysis was also carried out on the A2 scenario for 2050 and the B1 scenario for both 2050 and 2100 and similar results of a lower magnitude were found. The A2 scenario was chosen because analysis by Raupach et al. (2007) suggests that the B1 scenario is now unrealistic.

3.1 Continental Projections

Figure 3a shows the projected T_{max} 20-year return value across Australia for the all-model ensemble. Much of inland and far western Australia can expect return values of 50-52°C by 2100. Southern and eastern coastal regions can expect slightly cooler return values of up to 46°C, warming to 48°C when heading inland, particularly in the south.

Figure 3b compares the projection of the 20-year return value from the PDF-based ensemble to that of the all-model ensemble. It is clear to see that the PDF-based ensemble produces cooler return values over the majority of the continent. Much of the east has return values up to 5°C cooler in the skill ensemble, and values up to 2°C cooler in the north and south. Projections in the centre of the continent are similar to the all-model ensemble. Figure 3c shows the difference between the all-model ensemble and the tail tail-based skill ensemble. The tail-based skill ensemble is even cooler than the PDF-based ensemble, being up to 5°C cooler than the all model ensemble for much of the east, north and small patches in the west. Projections in the central south are up to 2°C cooler in the tail-based skill ensemble compared to the all-model ensemble. Figure 3d compares the mean-based ensemble to the all-model ensemble. Cooler return values seen in figures 3b and 3c are more isolated along the eastern seaboard and the north and south. The central south and west have return values similar to the all-model ensemble, however there is a patch in the central north 2-5°C warmer than the all-model ensemble.

Figure 4 is the same as figure 3 but for T_{min} . In figure 4a, the all-model ensemble projects return values of up to 20°C in the north. Return values become cooler head southwards, reaching -1°C in the far southeast. Figure 4b shows the difference between the all-model ensemble and the PDF-based

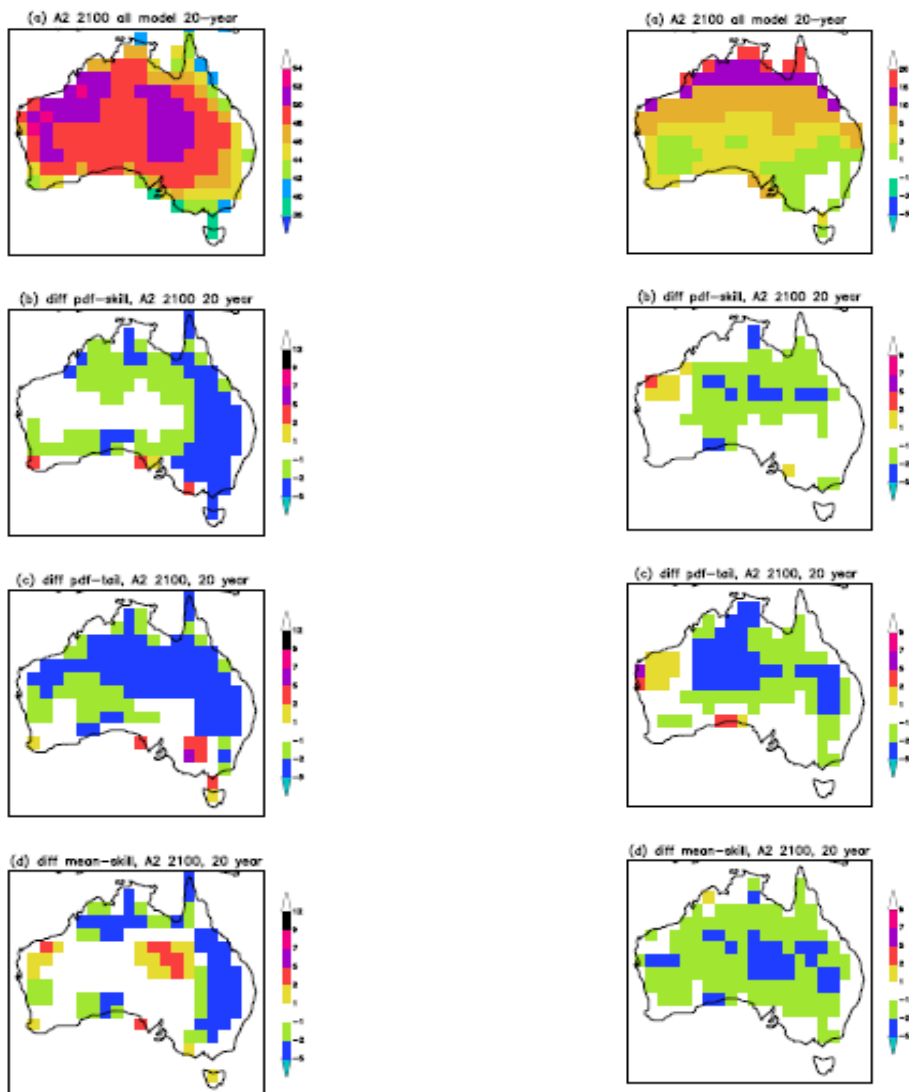


Figure 3 – T_{\max} : a) all-model ensemble projection for the 20-year return value for A2 2100; b) difference in projections between the PDF-based ensemble and all-model ensemble; c) difference in projections between the tail skill-based ensemble and the all-model ensemble; d) difference between the mean-based ensemble and the all-model ensemble.

Figure 4 – Same as figure 3 but for T_{\min} .

ensemble. The results are patchy, with the northeast and central south 3-5°C cooler, and much of the rest of the continent with similar projections between the ensembles.

There is a small patch in the far west that has projections 2°C warmer in the PDF-based ensemble. Similar results are seen in the tail-based skill ensemble (figure 4c) although the cooler patches extend further southwards. The mean-based ensemble projections (figure 4d) are at least 2°C cooler than the all-model ensemble for almost all of Australia. There are isolated patch of projections 5°C cooler in the central east and west of the continent.

3.2 Regional Results

Exploring why the ensembles give different patterns of changes requires a focus on specific regions. Here we focus on three regions used by Perkins et al. (2007) – a temperate region (including Sydney, Region 2), a sub-tropical region (including Brisbane

Region 3) and a tropical region (Region 10) noting that all other regions show similar behaviour. Figure 5 shows T_{\max} at 20-year return levels, including a 90% bootstrap confidence interval (1000 samples). The all-model ensemble for each region (first bar) shows a large range of projected temperatures. In each region, and irrespective of skill-score used, the projected T_{\max} is *always* lower in the stronger models than the weaker models – and the 90% confidence levels for the two ensembles do not overlap in Regions 2 and 3.

This suggests that the projected 20-year interval temperatures from the weak models are statistically significantly higher than those projected by the strong models, irrespective of whether “strong” is defined using the mean, PDF or tail skill measure. Figure 6 shows for T_{\min} that the samples are not consistently significantly different at a 90% confidence level, but there is a systematic difference in that the weaker models always

simulate larger amounts of increase in T_{MIN} than the stronger models.

4. CONCLUSIONS

Studying the projection of temperature extremes is important to understand and to mitigate the impact on many human, biophysical and social systems. This study has shown that by 2100 warmer 20-year return values can be expected for both T_{max} and T_{min} (although not shown similar results were found for 2050). However, this study has also found that selecting models based on their level of skill impacts the amount of warming. This impact is generally cooler over all measures of skill for both variables, however some stronger models project warmer 20-year T_{max} return values in central Australia and warmer 20-year T_{min} values in the far west by 2100.

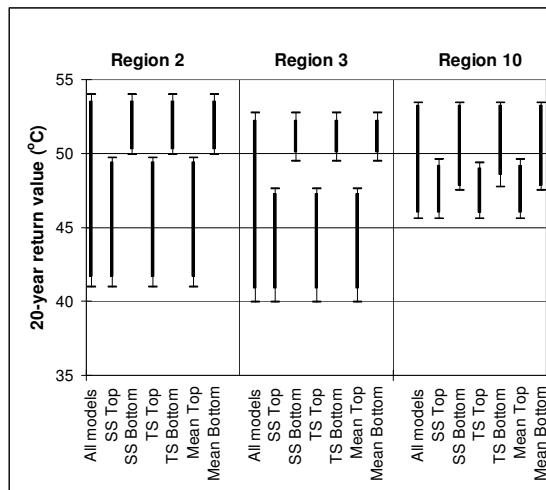


Figure 5 – Regional stock plots showing the range and the 90% bootstrapped confidence interval in the T_{max} 20-year return value for all ensembles over regions 2, 3 and 10 defined by Perkins et al. (2007).

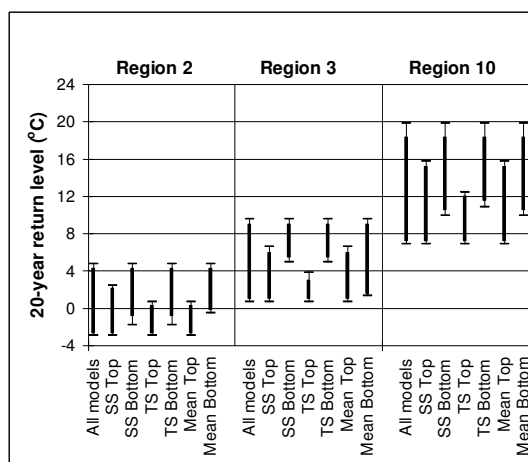


Figure 6 – Same as figure 5 but for T_{min} .

Our results clearly show at a regional level that models considered weak based on their level of skill simulate statistically significantly more warming in extreme temperatures than the models considered strong, and influence all-model ensemble projections towards warmer return values for T_{max} and T_{min} . We have demonstrated this for three different climatic types – but this result is true in all regions of Australia. At least, over Australia, the use of an all-model ensemble should be avoided. Our results do not advocate the use of a single best model based on a given measure of skill, but to create an ensemble with models shown to be demonstrably better than others. While the use of an all-model ensemble is better than using any given single model (Randall et al., 2007) at a global scale, the use of an all-model ensemble at regional scale clearly biases the projections of extreme temperature over Australia.

5. References

Allen M R, P A Stott, J F B Mitchell, R Schnur, and T L Delworth, (2000): Quantifying the uncertainty in forecasts of anthropogenic change. *Nature*, **407**, 617-620.

Alexander L V and Arblaster J M, (2008): Assessing trends in observed and modelled climate extremes over Australia in relation to future projections. *Int. J. Climatol.*, DOI: 10.1002/joc.1730.

Easterling D R, G A Meehl, C Parmesan, S A Changnon, T R Karl, and L O Mearns LO, (2000): Climate Extremes: Observations, Modeling, and Impacts. *Science*, **289**: 2068-2074

Hegerl G C, F W Zwiers, P A Stott, and V V Kharin, (2004): Detectability of Anthropogenic Changes in Annual Temperature and Precipitation Extremes. *J. Climate*, **17**, 3683-3700.

Katz R and B Brown, (1992): extreme events in a changing climate: variability is more important than averages. *Clim. Chg*, **21**, 289-302.

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

Kharin V, F Zwiers, and X Zhang, (2005): Intercomparison of near surface temperature and precipitation extremes in AMIP-2 simulations. *J. Climate*, **18**, 5201-5223.

Kharin V V, F W Zwiers, X Zhang and G C Hegerl, (2007): Changes in temperature and Precipitation Extremes in the IPCC Ensemble of Global Couple Model Simulations. *J. Climate*, **20**, 1419-1444.

- Knutti, R, G A Meehl, M R Allen, and D A Stainforth, (2006): Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate*, **19**, 4224–4233.
- Luo Q, R N Jones, M Williams, B Bryan and W Bellotti, (2005): Probabilistic distributions of regional climate change and their application in risk analysis of wheat production. *Clim. Res.*, **29**, 41-52.
- Mearns LO, R Katz, and S Schneider, (1984): Extreme high-temperature events: Changes in the probabilities with changes in mean temperature *J. Appl. Meteorol.*, **23**, 1601-1613.
- Meehl G, F Zwiers, J Evans, T Knutson, L Mearns, and P Whetton, (2000): Trends in extreme weather and climate events: Issues related to modelling extremes in projections of future climate change. *Bull. Amer. Meteorol. Soc.* **8**, 427-436.
- Moise A F and D A Hudson, (2008): Probabilistic predictions of climate change for Australia and southern Africa using reliability ensemble average of IPCC CMIP3 model simulations. *J. Geophysical Research*, **113**, DOI: 10.1029/2007JD009250.
- Nicholls N, (2004): The changing nature of Australian droughts. *Clim. Chg*, **63**, 323-336.
- Perkins S E, A J Pitman, N J Holbrook and J McAneney, (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.
- Perkins S E, A J Pitman and S A Sisson, (2009): Smaller projected increases in 20-year temperature returns over Australia in skill-selected climate models. *Geophys. Res. Lett*, submitted.
- Piani C, D J Frame, D A Stainforth, and M R Allen, (2005): Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett*, **32**, L23825.
- Randall D, R A Wood, S Bony, R Colman, T Fichefet, J Fyfe, V Kattsov, A J Pitman, J Shukla, J Srinivasan, R J Stouffer, A Sumi, and K Taylor, (2007): Climate models and their evaluation, in *Climate Change 2007: The Scientific Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Solomon Set al. (eds.). CUP, NY, USA.
- Raupach, M.R., G. Marland, P. Ciais, C. Le Quééré, J.G. Canadell, G. Klepper, and C.B. Field, 2007, Global and regional drivers of accelerating CO₂ emissions, *PNAS*, 104, 10288-10293, doi:10.1073/pnas.0700609105.
- Schaeffer M, F M Selten and J D Opsteegh, (2005): Shifts in means are not a proxy for changes in extreme winter temperatures in climate projections. *Clim. Dyn.*, **25**, 51-63.
- Shukla J, T DelSole, M Fennessy, J Kinter, and D Paolino, (2006): Climate model fidelity and projections of climate change. *Geophys. Res. Lett.* **33**, L07702, doi:10.1029/2005GL025579.
- Stott P A and J A Kettleborough, (2002): Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, **416**, 723-726.
- Taylor K E, (2001): Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106 (D7)**, 7183–7192.
- Trigo R, R Garia-Herrera, J Diaz, I Trigo, and M Valente, (2005): How exceptional was the early August 2003 heatwave in France? *Geophys. Res. Lett.* **32**,1071-1074.
- Von Storch H and F W Zwiers, (1999): *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.
- Watterson I G, (1996): Non-dimensional measures of climate model performance. *Geophys. Res. Lett*, **31**, L24123, DOI:10.1029/2004GL021276.
- Wehner, M F, (2004): Predicted Twenty-first-Century Changes in Seasonal extreme Precipitation Events in the Parallel Climate Model. *J. Climate*, **17**, 4281-4290.
- Whetton P, I Macadam, J Bathols, and J O'Grady, (2007): Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse patterns of climate models. *Geophys. Res. Lett*, **34**, L14701, doi:10.1029/2007GL030025.
- Zwiers F and V Kharin, (1998): Changes in the extremes of climate simulated by CCC GCM2 under CO₂ doubling. *J.Climate*, **11**, 2200-2222.