Multi-model forecast skill for mid-summer rainfall over southern Africa

Willem A. Landman[a,b]

Asmerom Beraki[c]

[a] Council for Scientific and Industrial Research, Natural Resources and the Environment, Pretoria, South Africa

[b] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa

[c] South African Weather Service, Pretoria, South Africa

Re-submitted to: *International Journal of Climatology*

29 October 2010

Correspondence to: Council for Scientific and Industrial Research, P.O. Box 395, Pretoria, 0001, South Africa; E-mail: WALandman@csir.co.za

Tel: +27-12-841-3395

Fax: +27-12-841-4863

Key words: multi-model, downscaling, seasonal forecasting, ENSO, southern Africa.

ABSTRACT

Southern African December-January-February (DJF) probabilistic rainfall forecast skill is assessed over a 22-year retro-active test period (1980/81 to 2001/02) by considering multi-model ensembles consisting of downscaled forecasts from three of the DEMETER models, the ECMWF, Météo-France and UKMO coupled ocean-atmosphere general circulation models. These models are initialized in such a way that DJF forecasts are produced at an approximate 1-month lead-time, i.e., forecasts made in early November. Multi-model forecasts are obtained by 1) downscaling each model's 850 hPa geopotential height field forecast using canonical correlation analysis (CCA) and then simply averaging the rainfall forecasts, and 2) by combining the three models' 850 hPa forecasts and then downscaling them using CCA. Downscaling is performed onto the 0.5°x0.5° resolution of the CRU rainfall data set south of 10° south over Africa. Forecast verification is performed using the relative operating characteristic (ROC) and the reliability diagram. The performance of the two multi-model combinations approaches are compared with the single model downscaled forecasts and also with each other. It is shown that the multi-model forecasts outperform the single model forecasts, that the two multi-model schemes produce about equally skilful forecasts, and that the forecasts perform better during El Niño and La Niña seasons than during neutral years.

1. Introduction

The scientific basis for the existence of seasonal climate predictability originates from the observation that slowly evolving sea-surface temperature (SST) anomalies influence seasonal-mean weather conditions (Palmer and Anderson 1994). Therefore, estimation of the evolution of SST anomalies, which are often relatively predictable, and subsequently employing them in atmospheric general circulation models (GCMs), potentially provides means of generating forecasts of seasonal-average weather (Graham *et al*. 2000). With the advent of fully coupled ocean-atmosphere models (Stockdale et al., 1998; Saha et al., 2006; Weisheimer *et al*. 2009), evidence that the ocean models participating in fully coupled GCMs can predict the evolution of SSTs to elevated levels of skill has been presented. This notion has been demonstrated conclusively through the DEMETER (Development of a European Multimodel Ensemble system for seasonal to interannual prediction) project (Palmer et al. 2004), and recently the usefulness of these forecasts over the mid-latitudes has been further demonstrated (Coelho *et al*. 2006; Frias *et al*. 2010). In theory coupled models should eventually outperform using GCMs as a second step in a 2-tiered system in which SSTs are first predicted since the former is able to describe the feedback between ocean and atmosphere while the latter assumes that the atmosphere responds to SST but does not in turn affect the oceans (Copsey et al., 2006; Troccoli et al., 2008).

Although GCMs, commonly configured with an effective resolution of 100-300 km, have demonstrated skill at global or even continental scale, they are unable to represent local sub-grid features, subsequently overestimating rainfall over southern Africa (Joubert and Hewitson 1997; Mason and Joubert 1997). Also, the representation of rainfall at mid-to-high latitudes is complex and often not well estimated (Graham *et al*. 2000; Goddard and Mason 2002). Such systematic biases have created the need to downscale GCM simulations over southern Africa. Semi-empirical relationships exist between observed large-scale circulation and rainfall, and assuming that these relationships are valid under future climate conditions and also that the large-scale structure and variability is well characterized by GCMs, mathematical equations can be constructed to predict local precipitation from the forecast large-scale circulation (Landman and Goddard, 2002; Wilby and Wigley 1997). Empirical remapping of GCM fields to regional rainfall has been demonstrated successfully over southern Africa (Bartman *et al*. 2003; Landman and Goddard 2002, 2005; Landman *et al*. 2001; Shongwe *et al*., 2006).

The chaotic inherent variability of the atmosphere requires seasonal climate simulations to be expressed probabilistically. Probabilistic forecasts are made possible through the proper use of GCM ensembles since ensemble forecasting is a feasible method to estimate the probability distribution of atmospheric states (Branković and Palmer 2000). In addition, errors in the initial conditions as well as deficiencies in the parameterizations and systematic or regime-dependent model

errors can be to a large part accounted for through ensemble forecasting (Evans *et al*. 2000). Moreover, there is inevitable growth in differences between forecasts started from very slightly different initial conditions suggesting that there is no single valid solution but rather a range of possible solutions (Tracton and Kalnay 1993). Information contained in the distribution of the ensemble members can subsequently be used to represent forecast probabilities by calculating the percentage of ensemble members that fall within a particular category (e.g. below-normal, near-normal or above-normal). Similarly, forecast probabilities can be produced indicating the percentage of ensemble members in the upper or lower extremes, e.g., 15^{th} percentiles (Mason *et al*. 1999).

There are advantages in combining ensemble members of a number of GCMs into a multi-model ensemble since GCMs differ in their parameterizations and therefore differ in their performance under different conditions (Hagedorn *et al*., 2005). Using a suite of several GCMs not only increases the effective ensemble size; it also leads to probabilistic simulations that are skilful over a greater portion of the region and a greater portion of the time series. Multi-model ensembles are nearly always better than any of the individual models (Dirmeyer *et al*. 2003, Doblas-Reyes *et al*. 2000, 2005, Hagedorn *et al*., 2005; Krishnamurti *et al*. 2000). The benefits from combining ensembles are a result of the inclusion of complementary predictive information since the forecast scheme is able to extract useful information from the results of individual models from local regions where their skill is higher (Krishnamurti *et al*. 2000). In fact, the most striking

benefit obtained from multi-model ensembles is the skill-filtering property in regions or seasons when the performance of the individual models varies widely (Graham *et al*. 2000). Moreover, increased ensemble size leads to further benefits (Brown and Murphy 1996), but the multi-model approach is only beneficial if the individual models produce independent skilful information (Graham *et al*. 2000). A number of ensemble combining algorithms exists. The most simple of these is the unweighted combination of ensembles from different models (Hagedorn *et al*. 2005; Graham *et al*. 2000, Mason and Mimmack 2002; Peng *et al.* 2002; Tippet and Barnston 2008). The improvements of a multi-model over the individual ensemble systems are attributed to the collective information of all the models used in the mean of probabilities algorithm. However, the forecast quality of a simple multi-model ensemble is often difficult to improve on when the available sample size is relatively small (Doblas-Reyes *et al*. 2005).

An association exists between South Africa's summer seasonal rainfall and the equatorial Pacific Ocean. However, the association in the middle to late austral summer season is higher than earlier in the summer rainy season (e.g., Tyson and Preston-Whyte, 2000), and it is also non-linear (Fauchereau *et al*. 2008). Notwithstanding, in the mid-summer months South Africa tends to be anomalously dry during El Niño years and anomalously wet during La Niña years, although wet El Niño seasons and dry La Niña seasons are not uncommon. Indian and Atlantic Ocean SST also have a statistically detectable influence on South African rainfall variability (e.g., Mason, 1995; Reason *et al*., 2006).

Moreover, while the El Niño-Southern Oscillation (ENSO) has a control on rainfall variability over the southern African region, Indian Ocean SST anomalies, sometimes varying independently of ENSO, are important for the skilful simulation of southern African seasonal rainfall variability using atmospheric GCMs (e.g., Washington and Preston, 2006). Since ENSO is the dominant mode of seasonal and interannual climate variability globally, and since ENSO has a strong influence on southern African rainfall, it needs to be investigated to what extent ENSO influences coupled model performance over southern Africa.

The paper consists of three parts: 1) single coupled model downscaled forecast performance during mid austral summer over southern Africa compared with that of multi-models, 2) the comparison between unweighted and weighted combination of forecasts, and 3) multi-model performance during ENSO and during neutral years. For the second part, the unweighted combination involves downscaling and correcting GCM output first before combining, while for the weighted combination weighting is done and then combined before downscaling and correcting.

2. Data, models and methods

2.1. Rainfall data

The season of interest is December-January-February (DJF) when southern Africa is being dominated by influences mainly from the tropics and so is a season of relatively high predictability and ideal for seasonal predictability studies over the region. The University of East Anglia Climatic Research Unit (CRU) global 0.5° x 0.5° monthly data, Version 2.1 (Mitchell and Jones, 2005) are used to construct DJF seasonal averaged rainfall totals for southern Africa south of 10° south for the period 1959/60 to 2001/02. This data set is used for both empirical downscaling and for forecast verification.

2.2. Coupled general circulation models

The atmosphere-ocean models used in this study are from the DEMETER project (Palmer *et al*., 2004) and in particular are the ECMWF, Météo-France and UKMO coupled models. These models were selected since they each have 43 years of available hindcast data, and the longer the record of archived model data the better the chance is to develop robust empirical downscaling equations. Hindcasts had been started from 1 November and nine ensemble members created. Seasonal means are used in the study.

2.3. Model output statistics

Given the low spatial resolution of the coupled models (Palmer *et al*., 2004) there is a need to downscale the global model output to a higher resolution to satisfy end-user needs and to further improve on the forecasts (Landman and Goddard, 2002) through the correction of systematic deficiencies in the global models (Tippet *et al*., 2005). Model output statistics (MOS; Wilks, 2006) equations are developed here because they can compensate for these errors in the model fields directly in the regression equations. The reason why these errors can be overcome is because MOS uses predictor values from the global models in both the development and forecast stages. Notwithstanding, the selection of the appropriate model field require careful consideration: Raw model forecast of rainfall that is a result of, for example, the interaction between atmospheric circulation and topography is poorly resolved, and may therefore not be a good predictor of rainfall observed at ground level. Rainfall fields, even when totalled over a season, are noisy, and normally contain structures on spatial scales well below those resolved by the models. However, variables such as large-scale circulation are more accurately simulated by models than rainfall and should therefore be used instead in a MOS system to predict seasonal rainfall totals (Landman and Goddard, 2002).

The MOS equations are developed by using the canonical correlation analysis (CCA; Barnett and Preisendorfer, 1987) option of the Climate Predictability Tool

(CPT). This tool was developed at the International Research Institute for Climate and Society (IRI; http://iri.columbia.edu). The forecast fields from each GCM used in the MOS are restricted over a domain that covers an area between the Equator and 40°S, and Greenwich to 60°E. Empirical orthogonal function (EOF) analysis is performed on both the predictor (model forecast fields) and predictand sets (CRU data over southern Africa) prior to CCA, and the number of EOF and CCA modes to be retained in the CPT's CCA procedure is determined using cross-validation skill sensitivity tests. Both the models' ensemble mean rainfall and 850 hPa geopotential height fields were separately considered over the available 43-year period (1959/60 – 2001/02) to find out which of the two fields provide the best first estimate for the downscaled forecasts. A 5-year-out cross-validation design was selected and it was found that for both the ECMWF and UKMO models, the height field is the better option, but for the Météo-France model, rainfall was a slightly better performer. Notwithstanding, 850 hPa geopotential heights were selected for all three models for consistency and because of the potential problems mentioned above when rainfall as a downscaling predictor field is used. Considering other model fields such as moisture and geopotential heights at levels other than 850 hPa showed no further benefits over only using the 850 hPa geopotential fields as a single predictor field either.

2.4. Model combination

A number of forecast combining algorithms exists, but only two are considered here. The first is the most simple of all combination schemes and involves unweighted averaging of the forecast probabilities (e.g., Hagedorn *et al*., 2005). For this simple combination approach, the 850 hPa height forecasts from the three coupled models are first separately downscaled to DJF rainfall at the 0.5° x 0.5° CRU resolution and then averaged, and is referred to here as a combination using equal weights (MMeqw). The second approach allows the models to be weighted by combining the 850 hPa geopotential height forecasts fields from the models prior to EOF pre-filtering in the CCA process. Downscaling is then performed as before, but with combined forecast fields (MMcca) as opposed to individual model fields.

2.5. Retro-active forecasts

In order to minimize artificial inflation of forecast skill, the performance of the individual models and the two multi-model systems (MMeqw and MMcca) should be verified over a test period that is independent of the training period and should involve evaluation of predictions compared to their matching observations excluding any information following the forecast year. Such a system mimics a true operational forecasting environment where no prior knowledge of the coming season is available. The individual models and two multi-model systems are first

trained with information from 1959/60 and leading up to and including 1979/80. The seasonal rainfall of the next year (1980/81) is subsequently predicted using the trained models. The various MOS sets of equations are subsequently retrained using information leading up to and including 1980/81 to predict for 1981/82 conditions. This procedure is continued until the 2001/02 DJF rainfall is predicted using MOS systems trained with data from 1959/60 to 2000/01, resulting in 22 years (1980/81 – 2001/02) of independent forecast data. In estimating the skill in predicting DJF rainfall over southern Africa, the observed and predicted fields are separated into three equi-probable categories based on the preceding years' climatology defining above-normal, near-normal and below-normal seasonal rainfall totals.

The distribution of individual ensemble members is intended to be able to indicate forecast uncertainty. However, only a finite ensemble is available (9 members from each coupled model) suggesting that the forecast distribution may be poorly sampled – and so the uncertainty associated with the forecasts has to be estimated. Probabilistic MOS forecasts for each of the 22 retro-active years are obtained here from the error variance of the cross-validated predictions using the ensemble mean (Troccoli *et al*., 2008) for each of the various training periods. The errors in the predictions are assumed to be Gaussian. Cross-validation is performed using a (large) 5-year-out window, which means that 2 years on either side of the predicted year are omitted, in order to minimize the chance of obtaining biased results.

This modelling study also focuses on one of the major sources of predictability over southern Africa, namely the El Niño – Southern Oscillation (ENSO) phenomenon, and how forcing from the equatorial Pacific Ocean influences predictability over the region. The El Niño, La Niña and neutral years considered are those listed by Coelho et al. (2006). Rainfall retro-active forecast skill over the subcontinent is then assessed during El Niño (1982/83, 1986/87, 1987/88, 1990/91, 1991/92, 1992/93, 1994/95 and 1997/98 = 8 seasons), La Niña (1983/84, 1984/85, 1988/89, 1995/96, 1998/99, 1999/00 and 2000/01 = 7 seasons) and neutral (1980/81, 1981/82, 1985/86, 1989/90, 1993/94, 1996/97 and 2001/02 = 7 seasons) events.

2.6. Estimating true forecast performance

For the generation of verification data we adopt an approach that minimizes the inflation of forecast skill by testing the models in an environment that mimics that of an operational centre, i.e. a retro-active forecast setting (Wilks, 2006). However, owing to the limited archived model data set available the MOS equations used for the prediction of the first part of the verification set may not display a robust relationship between the predictor (850 hPa heights) and predictand (rainfall at the surface) throughout the retro-active process, but this problem should become less of an issue as the forecast process progresses

beyond about 30 years of training data. Notwithstanding, here we assume that the relationships remain robust, a notion that will be tested later on in the paper.

Since seasonal climate is inherently probabilistic, seasonal forecasts should be judged probabilistically. The main attributes of interest for probabilistic forecasts are: 1) reliability (is the confidence communicated in the forecast appropriate and are there systematic biases in the forecast probabilities?), 2) resolution (is there any useable information in the forecast?), 3) discrimination (are the forecasts discernibly different given different outcomes?), and 4) sharpness (what is the confidence level that is communicated in the forecast?) (Troccoli *et al*., 2008; Wilks, 2006). The forecast verification measures are the reliability diagram (Hamill 1997; Wilks, 2006) and the relative operating characteristic (ROC; Mason and Graham, 1999; Wilks, 2006). A forecast system is deemed reliable if there is consistency between predicted probabilities of an event such as drought/floods (or below/above-normal rainfall in this paper) and the observed relative frequencies of drought/floods. Reliability diagrams will be used here to assess the reliability and confidence of the forecasts. ROC applied to probabilistic forecasts indicates whether the forecast probability was higher when an event such as drought occurred compared to when it did not occur, and therefore identifies whether a set of forecasts has the attribute of discrimination. Here the area underneath the ROC curve is used as a measure of discrimination in the prediction of below-normal and above-normal DJF rainfall totals.

3. Results

3.1. Deterministic assessment of forecasts

Although the seasonal climate is inherently probabilistic and therefore seasonal forecasts globally are for the most part issued probabilistically, it is often informative to investigate deterministic forecast performance. Figure 1 shows area-averaged (Africa south of 10°S) deterministic cross-validated (5-year-out approach) multi-model DJF rainfall (mm) forecasts over the available 43-year period (1959/60 – 2001/02) compared with the observed. The cross-validation procedure is designed in such a way that the data is “wrapped” around in order to make a 5-year-out approach possible while at the same time producing cross-validated forecasts for the whole period. Forecasts for both MMcca and MMeqw are shown, and El Niño and La Niña seasons are respectively marked with “E” and “L”. The vertical line on the figure divides the time series into two parts: The initial training period for the creation of retro-active forecasts (1959/60 – 1979/80; 21 years) and the retro-active test period (1980/81 – 2001/02; 22 years) for which probabilistic forecasts are generated. The Spearman’s correlation between the area-averaged 22-year forecasts and observations for MMcca and MMeqw are respectively 0.4783 and 0.4873, suggesting about equally skilful area-averaged deterministic forecasts from the two multi-model methods. The Spearman’s correlation is used here since the 1997/98 rainfall predictions are considered outliers (Figure 1). The four driest years during the 22-year test period (1982/83,

1986/87, 1991/92, 1994/95) are associated with El Niño seasons and the four wettest with La Niña seasons (1988/89, 1995/96, 1998/99, 1999/00). For the most part, the forecasts do not capture the size of the observed anomalies for these extreme seasons, but this is often found with linear regression-based downscaling techniques such as the one used here. Notwithstanding, no attempt was made here to inflate the forecasts since variance adjustment of forecasts are generally discouraged (Trocccoli *et al*., 2008).

The length of the training period may have an effect on the robustness or stability of the MOS equations (Doblas-Reyes *et al*., 2005; Wilks, 2006). For stability it is understood that the fitted equations are also applicable to independent data. Since the initial training period (for making the 1980/81 rainfall forecasts) is only 21 years long, investigation into the variation of forecast performance over the various training periods is warranted. Figure 2 shows area-averaged Spearman's correlations (adjusted with the Fisher Z transformation (Wilks, 2006)) for various cross-validation training periods ranging from 12 years to 43 years, using MMcca, and using August-September-October averaged SSTs as predictor in a statistical model (CCA). The SST predictor field is between 170°E to 80°W and 20°N to 20°S in order to capture central and eastern equatorial Pacific SST variability. A 4^{th} order polynomial is fitted to the averaged Spearman's correlations and a gradual improvement in forecast skill can be seen towards a training set consisting of 32 years when MMcca is used, and throughout the whole period when using SSTs as predictor in the statistical model. A skill plateau could have

been attained with the MMcca were it not for the large errors associated with the rainfall prediction of the 1997/98 El Niño season and of the two preceding years. Thereafter a gradual decrease is seen until 43 years are included in the MOS training period. Using the DJF 850 hPa geopotential field predicted at the end of October by the coupled ECHAM4.5-MOM3-DC2 (http://iridl.ldeo.columbia.edu/SOURCES/.IRI/.MP/.RESEARCH/.COUPLED/.GLOBAL/.ECHAM4p5-MOM3-DC2/) as predictor in the same MOS downscaling approach for southern Africa, a similar shape is found in the variation of skill (Figure 2). Here the initial training period is from 1982/83 to 1991/92. It is suggested that the decrease in skill towards the 2001/02 season is therefore not a function of the DEMETER data used here, since a differently configured coupled model produces similar results. Forecast skill using physical models may thus not be constant in time. However, the dominant modes of CCA (Barnett and Preisendorfer, 1987) for the multi-model considered here remain the same (not shown) regardless of the training period used (e.g. Landman and Goddard, 2002), which suggests stability in the selected dominant modes of variability included in the MOS equations, and therefore implies stability in the MOS prediction equations even though forecast skill may not be constant in time.

3.2. Multi-model vs. single model results

By knowing the probability of a predicted category occurring, additional forecast value is obtained (Mason and Graham, 1999), since probabilistic forecasts exhibit

reliability considerably in excess of that achieved by corresponding deterministic forecasts (Murphy, 1998). Probabilistic rainfall forecasts are produced here for three equi-probable categories of above-normal, near-normal and below-normal. Only the verification results for the above- and below-normal categories are presented here since there is little skill to be derived from predicting the near-normal category (Van den Dool and Toth 1991).

A ROC graph is made by plotting the forecast hit rates against the false alarm rates (Wilks, 2006). The area beneath the ROC curve is used as a measure of discrimination here and is referred to as a ROC score. If the area would be ≤0.5 the forecasts have no skill, and for a maximum ROC score of 1.0, perfect discrimination has been obtained. The ROC score can be interpreted here as a probability of the forecast system successfully discriminating respectively above- or below-normal seasons from other seasons.

The ROC graph and its score can be meaningfully applied in seasonal forecasting given the small sample size normally associated with these forecasts (Troccoli *et al*., 2008). Figure 3 shows the area-averaged ROC scores for above- and below-normal DJF rainfall for each of the individual downscaled models (Météo-France – MF; ECMWF and UKMO) and for the two multi-models (MMeqw and MMcca) as calculated over the 22-year test period in a retro-active design. All area-averaged scores are above 0.5, which means that on average there is more than a 50% chance that all the forecast systems have the ability to

successfully discriminate respectively wet and dry seasons from other seasons. Two of the three single models have a greater ability to discriminate the below-normal category as opposed to the above-normal one, but both the multi-models are better able to discriminate the below-normal category. Moreover, the multi-models have higher averaged ROC scores than any of the individual models. In fact, based on the area-averaged scores the multi-models each have at least a 61% chance of discriminating the above-normal category and at least a 63% chance of discriminating the below-normal DJF rainfall. The outperformance by the multi-models over southern Africa confirms what has been found with many other studies that multi-model forecasts usually outscore single model forecasts (e.g. Barnston *et al*., 2003; Doblas-Reyes *et al*., 2005; Hagedorn et al., 2005; Coelho *et al*., 2006; Weigel *et al*., 2008; Wang and Fan, 2009).

The improvement in forecast performance of the multi-models over the single models is further demonstrated in Figure 4 that shows the geographical distribution of ROC score differences between the multi-models and the individual models. Figure 4(a) shows where the multi-model that uses equal weights (MMeqw) outscore each of the individual models, and Figure 4(b) where the weighted forecast combination multi-model (MMcca) outscores them. Shaded areas are where the multi-models outperform the single models. Both sets of maps show that most of southern Africa is associated with positive ROC score differences, thus providing further evidence that the multi-models are outscoring the single models.

The ROC score is sometimes criticized as a measure of forecast performance because of its insensitivity to reliability (Troccoli *et al*., 2008). Figure 5 shows the reliability diagrams for the individual models. In addition to the respective reliability curves for the two categories, their least-squares regression lines are presented on the diagrams. The regression lines are calculated with weighting relative to how frequently forecasts are issued at a given confidence. When these regression lines lie along the diagonal, the forecasts are perfectly reliable. When the regression line lies above the diagonal observed above- or below-normal DJF rainfall tends to occur more frequently than forecast, but when it lies below the diagonal the observed categories respectively tend to occur less frequently than forecast, indicating under- and over-forecasting respectively. The most common slope of the regression line found for seasonal forecasting is one that is shallower than the diagonal line (Troccoli, *et al*., 2008) – the forecasts are said to be over-confident. Histograms are also included in the figures, and they show the frequencies with which forecasts occur in probability intervals of 10%, starting at 5%.

All the forecasts made by the single models for both above- and below normal DJF are over-confident (Figure 5). However, forecasts for below-normal rainfall totals are less over-confident than forecasts for above-normal rainfall for all three single models. Since the single models are over-confident, multi-model ensembles can enhance prediction skill regardless of which combination

approach is used since multi-model combination reduces over-confidence (Weigel *et al*., 2008). Figure 6 shows the reliability diagrams of the two multi-models, and here improved reliability over the single models is in fact seen (the regression lines for both categories tend to be closer to the diagonal). However, for both multi-models the high-probability above-normal forecasts are not reliable, as well as the high-probability below-normal forecasts of the MMeqw model. This result suggests that a simple equal weighting scheme to combine forecasts may not sufficiently reduce over-confidence (Barnston *et al*., 2003) for high-probability forecasts. Difference maps (not shown) of ROC scores (MMcca minus MMeqw) for the two categories show more or less an even split in terms of the areas of positive and negative score differences. This result indicates that both multi-model approaches are not much different in their ability to discriminate events from non-events, and that the MMcca is only slightly better able to produce reliable high-probability below-normal rainfall forecasts. However, such forecasts are often made during El Niño seasons

It has been shown that both the single and multi-models have the ability to discriminate between different observed situations. However, the multi-models outscore the single models, both in terms of discrimination and reliability. Since southern African mid-summer rainfall is influenced by the state of the equatorial Pacific Ocean, there is a need to investigate how skilful a multi-model predicts the two rainfall categories during ENSO and during neutral events separately.

3.3. Multi-model forecast performance during ENSO years

CCA pattern and time series analysis (Barnett and Preisendorfer, 1987) of the multi-model (MMcca) forecast system suggests that the dominant modes of predictor variability (three or four canonical modes that produce the best forecast results over the retro-active forecast period) are partly related to different influences of ENSO on southern African mid-summer rainfall (Fauchereau *et al*., 2008) since the correlations between the Oceanic Niño Index (ONI; www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml) and the three leading canonical temporal scores of the predictor (combined 850 hPa geopotential height fields) are respectively 0.5017 ($p<0.01$), -0.5337 ($p<0.01$) and -0.3023 ($p<0.05$) over the 43-year period. The question may arise then what added benefit there may be in running multi-model systems that consist of physical models that are primarily ENSO driven, over a simple statistical model that uses Pacific Ocean SSTs as predictors and is much cheaper to run. This question is answered by referring back to Figure 2. The gray dashed line is the 4^{th} order polynomial that is fitted to the area-averaged Spearman's correlation obtained by using a simple statistical model (CCA) with central and eastern equatorial Pacific Ocean SST (170° E to 80° W; 20° N to 20° S) as predictor. Although there is convergence in the performance of the forecasting systems towards the end of the cross-validation period, the multi-model outscores the simple model throughout. This result suggests that the coupled models' downscaled forecasts include additional forecast information that cannot be

derived from equatorial Pacific SST alone, which justifies the use of physical forecast models to predict seasonal rainfall variability over southern Africa. Take note that the introduction here of the statistical model was not to set an easy to beat baseline skill level, but to demonstrate that the skill of the GCMs comes from climatological forcings beyond the central and eastern equatorial Pacific Ocean.

The multi-model DJF rainfall forecast performance during the El Niño (8 seasons), La Niña (7 seasons) and neutral (7 seasons) years over the 22-year retro-active period are shown in Figure 7 to 9. The forecasts for the ENSO and non-ENSO years are separately taken from the retro-active forecasts prior to calculating the verification statistics for these years. Since the skill calculations are based on only a few cases (7 or 8) they may be sensitive to sampling errors. ROC calculations are however less sensitive to sampling errors than reliability diagrams (Troccoli *et al*., 2008). Figure 7 presents area-averaged ROC scores and it is shown that on average the multi-model is able to discriminate the above-normal and below-normal rainfall categories during ENSO years, but fails to do so during neutral years (averaged ROC scores are below 0.5 for both categories). Moreover, the multi-model performs best predicting drought during El Niño years and floods during La Niña years, but there is skill in predicting wet El Niño and dry La Niña seasons over southern Africa too. This result is further manifested in the geographical distribution of ROC scores for the above- and below-normal rainfall categories and for ENSO and neutral years as shown in

Figure 8. Large patterns of ROC scores in excess of 0.5 are seen for the El Niño and La Niña cases, but much smaller areas associated with neutral years are found. The multi-model therefore performs poorly during neutral years. The reliability diagrams for rainfall prediction during El Niño and La Niña years are shown in Figure 9. Forecasts are again over-confident, but as is found with the ROC scores there is skill in predicting both drought and wet seasons during El Niño years and predicting wet and drought seasons during La Niña years. The forecasts at least correctly indicate increases and decreases in the probabilities of the wet and dry events.

4. Discussion and conclusions

Southern African mid-summer probabilistic rainfall prediction skill has been assessed by using forecasts from state-of-the-art fully coupled models that are empirically downscaled and combined in order to produce multi-model forecasts. Forecast performance was tested over a retro-active period of 22 years that mimics an operational forecast configuration. Multi-model forecasts outscore single model forecasts and can be used with confidence during El Niño and La Niña seasons. In addition, the two multi-model forecast approaches produce about equally skilful forecasts.

The robustness of the MOS equations was tested and found that although forecast skill may not be constant in time, especially with short training periods,

the dominant modes of variability included in the equations remain similar for a variety of training periods. Regardless of this variation in skill, multi-model performance consistently outscored a simple statistical model that only includes equatorial Pacific Ocean SST variability as predictor. The improved multi-model forecasts are therefore a result of the system's ability to include forecast information in addition to the signal originating from the central and eastern equatorial Pacific Ocean. Both single model downscaled forecasts and multi-model forecasts seems to be able to discriminate between different observed situations such as below-normal and above-normal DJF rainfall seasons, notwithstanding the result that forecasts are overconfident. Prediction of wet or dry conditions during ENSO years is also skilful, but little skill has been found predicting DJF rainfall when the equatorial Pacific Ocean is in a neutral state. Predictions during El Niño seasons are strongly overconfident, but are less so for rainfall predictions during La Niña seasons.

The paper has demonstrated that multi-model systems are able to provide useful operational mid-summer rainfall forecasts over southern Africa, but only during ENSO years. Rainfall forecasts for southern Africa produced by the EUROSIP multi-model, that consists of later versions of the three coupled GCMs discussed here, made near the end of 2009 for the 2009/10 DJF El Niño season show mostly enhanced probabilities for dry conditions to occur. A similar forecast was also issued by other international centres such as the IRI, and also by the South African Weather Service. Moreover, summer rainfall forecasts for 2009/10 issued

to the South African public was made with high confidence, partly based on the result that multi-models can produce reliable drought forecasts and because of the confidence in summer rainfall forecasts during El Niño seasons. However, DJF rainfall over South Africa was anomalously high, especially over the central and western parts of that country (http://www.weathersa.co.za) and so the observed wet 2009/10 austral summer season over the region was largely missed by most forecasting systems. Further model development (e.g. Engelbrecht *et al*., 2007) and modelling studies on how models represent the coupled system over southern Africa are therefore warranted.

Acknowledgements

This work was partly sponsored by the Water Research Commission of South Africa (project K5/1492).

References

Barnett TP, Preisendorfer RW. 1987. Origins and levels of monthly and seasonal forecast skill for United States air temperature determined by canonical correlation analysis. *Monthly Weather Review* **115** : 1825-1850.

Barnston AG, Mason SJ, Goddard L, DeWitt DG, Zebiak SE. 2003. Multimodel ensembling in seasonal climate forecasting at IRI. *Bulletin of the American Meteorological Society*, 1783-1796. DOI: 10.1175/BAMS-84-12-1783.

Bartman AG, Landman WA, Rautenbach CJ deW. 2003. Recalibration of general circulation model output to austral summer rainfall over southern Africa. *International Journal of Climatology* **23**: 1407-1419.

Branković Č, Palmer TN 2000. Seasonal skill and predictability of ECMWF PROVOST ensembles. *Quarterly Journal of the Royal Meteorological Society* **126**: 2035-2067.

Brown BH, Murphy AH. 1996. Improving forecasting performance by combining forecasts: the example of road-surface temperature forecasts. *Meteorological Applications* **3**: 257-265.

Coelho CAS, Stephenson BD, Balmaseda M, Doblas-Reyes FJ, van Oldenborgh GJ. 2006. Toward and integrated seasonal forecasting system for South America. *Journal of Climate* **19**: 3704-3721.

Copsey D, Sutton R, Knight JR. 2006. Recent trends in sea level pressure in the Indian Ocean region. *Geophysical Research Letters* **33**: L19712, doi:10.1029/2006GL027175.

Doblas-Reyes FJ, Déqué M, Piedelieve J-P. 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quarterly Journal of the Royal Meteorological Society* **126**: 2035-2067.

Doblas-Reyes FJ, Hagedorn R, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus* **57A**: 234-252.

Dirmeyer PA, Fennessy MJ, Marx L. 2003. Low skill in dynamical prediction of boreal summer climate: Grounds for looking beyond sea surface temperature. *Journal of Climate* **16**: 995-1002.

Engelbrecht FA, McGregor JL, Rautenbach CJdeW. 2007. On the development of a new nonhydrostatic atmospheric model in South Africa. *South African Journal of Science* **103**: 127-134.

Evans RE, Harrison MSJ, Graham RJ, Mylne KR. 2000. Joint medium-range ensembles from the Met Office and ECMWF systems. *Monthly Weather Review* **128**: 3104-3127.

Fauchereau N, Pohl B, Reason CJC, Rouault M, Richard Y. 2008. Recurrent daily OLR patterns in the southern African/southwest Indian Ocean region,

implications for South African rainfall and teleconnection. Climate Dynamics. DOI:10.1007/s00382-008-0426-2.

Frías MD, Herrera S, Cofiño AS, Gutiérrez JM. 2010. Assessing the skill of precipitation and temperature seasonal forecasts in Spain: Windows of opportunity related to ENSO events. *Journal of Climate* **23**: 209-220.

Goddard L, Mason SJ. 2002. Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Climate Dynamics* **19**: 619-631.

Graham RJ, Evans ADL, Mylne KR, Harrison MSJ, Robertson KB. 2000. An assessment of seasonal predictability using atmospheric general circulation models. *Quarterly Journal of the Royal Meteorological Society* **126** : 2211-2240.

Hagedorn R, Doblas-Reyes FJ, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus* **57A**: 219-232.

Hamill TM. 1997. Reliability diagrams for multicategory probabilistic forecasts. *Weather and Forecasting* **12** : 736-741.

Joubert AM, Hewitson BC. 1997. Simulating present and future climates of southern Africa using general circulation models. *Progress in Physical Geography* **21**: 51-78.

Krishnamurti TN, Kishtawal CM, Zang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S. 2000. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate* **13**: 4196-4216.

Landman WA, Goddard L. 2002. Statistical recalibration of GCM forecasts over southern Africa using model output statistics. *Journal of Climate* **15**: 2038-2055.

Landman WA, Goddard L. 2005. Predicting southern African summer rainfall using a combination of MOS and perfect prognosis. *Geophysical Research Letters* **32**: L15809. DOI: 10.1029/2005GL022910.

Landman WA, Mason SJ, Tyson PD, Tennant WJ. 2001. Retro-active skill of multi-tiered forecasts of summer rainfall over southern Africa. *International Journal of Climatology* **21**: 1-19.

Mason SJ. 1995. Sea-surface temperature – South African rainfall associations, 1910-1989. *International Journal of Climatology* **15**: 119-135.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* **128**: 2145-2166.

Mason SJ, Joubert AM. 1997. Simulated changes in extreme rainfall over southern Africa. *International Journal of Climatology* **17**: 291-301.

Mason SJ, Mimmack GM. 2002. Comparison of some statistical methods of probabilistic forecasting of ENSO. *Journal of Climate* **15**: 8-29.

Mason SJ, Goddard L, Graham NE,. Yulaeva E, Sun L, Arkin PA. 1999. The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bulletin of the American Meteorological Society* **80**: 1853-1873.

Mitchell TD, Jones PD. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology **25**: 693-712. DOI: 10.1002/joc.1181

Murphy AH. 1998. The early history of probability forecasts: Some extensions and clarification. *Weather and Forecasting* **13**: 5-15.

Palmer TN, Anderson DLT. 1994. The prospects of seasonal forecasting – a review paper. *Quarterly Journal of the Royal Meteorological Society* **120**: 755-793.

Palmer TN, Coauthors. 2004. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, DOI: 10.1175/BAMS-85-6-853.

Peng PT, Kumar A, van den Dool H and Barnston AG. 2002. An analysis of multi-model ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research*, **107**.

Reason CJC, Landman W, Tennant W. 2006. Seasonal to decadal prediction of southern African climate and its links with variability of the Atlantic Ocean, *Bulletin of the American Meteorological Society* : DOI:10.1175/BAMS-87-7-941.

Saha S, and Coauthors. 2006. The NCEP climate forecast system. *Journal of Climate* **19**: 3483-3517.

Shongwe ME, Landman WA, Mason SJ. 2006. Performance of recalibration systems for GCM forecasts for southern Africa. *International Journal of Climatology* **26**: 1567-1585.

Stockdale TN, Anderson DLT, Alves JOS, Balmaseda MA. 1998. Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature* **392**: 370-373.

Tippet MK and Barnston AG. 2008. Skill of multimodel ENSO probability forecasts. *Monthly Weather Review*, **136**, 3933-3946.

Tippett MK, Goddard L, Barnston AG. 2005. Statistical-dynamical seasonal forecasts of central-southwest Asian winter precipitation. *Journal of Climate* **18**: 1831-1843.

Tracton MS, Kalnay E. 1993. Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting* **8**: 379-398.

Troccoli A, Harrison M, Anderson DLT, Mason SJ. 2008. *Seasonal Climate: Forecasting and managing risk*. NATO Science Series. Earth and Environmental Sciences Vol 82. Springer.

Tyson PD, Preston-Whyte RA. 2000. *The Weather and Climate of Southern Africa*. Oxford University Press.

Van den Dool HM and Toth Z. 1991. Why do forecasts for near normal often fail? *Weather and Forecasting*, **6**, 76-85.

Wang H, Fan K. 2009. A new scheme for improving the seasonal prediction of summer precipitation anomalies. *Weather and Forecasting* **34**: 548-554. DOI: 10.1175/2008WAF2222171.1.

Washington R, Preston A. 2006. Extreme wet years over southern Africa: Role of the Indian Ocean sea surface temperatures. *Journal of Geophysical Research* **111** : D15104, DOI: 10.1029/2005JD006724.

Weigel AP, Liniger MA, Appenzeller C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society* **134**: 241-260.

Weisheimer A, Dobals-Reyes FJ, Palmer TN, Alessandri A, Arribas A, Déqué M, Keenlyside N, MacVean M, Navarra A, Rogel P. 2009. ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions – Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters* **36**: L21711, doi:10.1029/2009GL040896.

Wilby RL, Wigley TML. 1997.Downsclaing general circulation model output: A review of methods and limitations. *Progress in Physical Geography* **21**: 530-548.

Wilks DS. 2006. *Statistical Methods in the Atmospheric Sciences, 2nd Edition*. Academic Press.

Figure captions

Figure 1. Area-averaged observed (thick line) DJF rainfall (mm) over Africa south of 10° S, versus cross-validation forecasts (thin lines) from the two multi-models described in the text. El Niño (E) and La Niña (L) seasons are also shown. The arrow indicates where the retro-active test period starts. The years on the x-axis refer to the December months of the DJF seasons.

Figure 2. Variation in cross-validation forecast skill predicting DJF rainfall over southern Africa as reflected by area-averaged Spearman's correlation values. The thick black solid line (4th order polynomial) and associated thin black solid line show the MMcca multi-model's performance as a function of cross-validation training period, while the thick black dotted and thin black dotted lines represent the ECHAM4.5-MOM3-DC2 coupled model. The remaining gray lines represent the statistical model that uses equatorial Pacific Ocean SST as predictor. The arrow indicates where the retro-active test period starts.

Figure 3. ROC scores, averaged over the southern African domain, for the above-normal and below-normal rainfall categories. Scores for the single models and for the two multi-models are shown.

Figure 4. ROC score differences between the a) MMeqw multi-model and the single models, and b) MMcca multi-model and single models. Positive ROC score differences are where the multi-models are superior.

Figure 5. Reliability diagrams and frequency histograms for above- and below-normal DJF rainfall forecasts produced by the single models. The thick black curves and black bars of the histogram represent the below-normal rainfall category, while the thick black dotted curves and white bars of the histogram represent the above-normal rainfall category. For perfect reliability the curves should fall on top of the thick black diagonal line. The thin solid and dotted lines are respectively the weighted least-squares regression lines of the above-normal and below-normal reliability curves.

Figure 6. As in Figure 5, but for the two multi-models.

Figure 7. ROC scores, averaged over the southern African domain, for the above-normal and below-normal rainfall categories during El Niño, La Niña and neutral seasons. Scores for the MMcca multi-model are shown.

Figure 8. ROC scores of the MMcca multi-model, for El Niño, La Niña and neutral seasons, and for the above- and below-normal rainfall categories. ROC scores ≥0.5 are shaded.

Figure 9. As in Figure 5, but for rainfall predictions during El Niño and La Niña seasons using the MMcca multi-model.