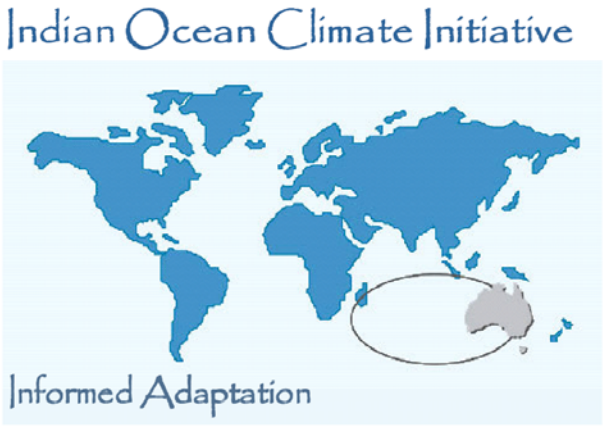

Attachment 3

Indian Ocean Climate Initiative Stage 2:

Unabridged Reports of Phase 1 Theme 3 Activity

July 2003 – Dec 2004

December 2004

Contents

Project 3.1: Improved procedures for selecting atmospheric predictors for statistical downscaling schemes. 3
Project 3.2: Improved rainfall amount simulations generated by the statistical downscaling model. 3
Project 3.3: Processed atmospheric fields from hindcasts produced by CAR and IRI for Climate Prediction (USA). 6
Project 3.4: Analysis of synoptic events that lead to rainfall extremes. 13
Project 4.1: Development of a hybrid model of factors influencing SWWA rainfall. 15
Project 4.2: Customised nonlinear data mining tools. 23

Project 3.1: Improved procedures for selecting atmospheric predictors for statistical downscaling schemes.

This project aims to improve the statistical downscaling results by looking for better links between the large scale atmospheric fields and daily rainfall at the point scale.

Selecting predictors in statistical procedures in general requires that the model fit be compared with and without each predictor. If there are many potential predictors then this can be computationally intensive, particularly in complex models. Therefore selection procedures for downscaling used to date are ad hoc and rely heavily on professional judgement. Thus two different users of the downscaling model could well yield different results. In addition, some features are hard to capture unless predictor selection is data-driven, such as northwest cloud band activity and interaction with frontal systems. More automatic, data-driven predictor selection techniques are required.

A variety of formal and informal approaches are currently being used for predictor selection, given the very large number of potential predictors in any application. A method known as boosting, drawn from the machine learning literature, has been used with some success in IOCI-sponsored work to investigate rainfall regime changes in southwest WA. Boosting has not been applied to statistical downscaling (SD) before, and requires an extension so that multiple rainfall stations can be used simultaneously to select optimal predictors. The technical requirements of the enhanced procedure have been identified and broken down into research tasks that are currently being implemented. The predictors selected by these boosting methods will be compared to those previously selected for the SWWA SD models. This will include re-fitting the SD models using any new predictor combinations and assessing their performance.

Collaboration with Sergey Schreider, University of California, Irvine (UCI) and Andrew Robertson, International Research Institute for Climate Prediction (IRI), is also leading to new methodologies for selecting combinations of atmospheric predictors. UCI and IRI are developing machine learning algorithms and have commenced applying them to Australian datasets.

Additionally, IOCI is benefiting from research undertaken as part of a GRDC funded project "*Climate change, wheat yield and cropping risks in Western Australia*" that has investigated a larger number of candidate predictors from the NCEP/NCAR Reanalysis data set for three geographic sub-regions covering the wheatbelt of SWWA. This has led to better performing SD models with new predictor sets tuned to the sub-regions. Previously dew-point temperature depression (DTD) at the 850 hPa level was used as a predictor in the SWWA SD model, whereas the new SD models perform better when using DTD from the 700 hPa level. This could indicate that moisture at the 700 hPa level better represents north-west cloudbands, for example.

Project 3.2: Improved rainfall amount simulations generated by the statistical downscaling model.

This project aims to develop techniques which can improve the rainfall amount simulations.

Previous investigations (statistically downscaling CAR Mk2 seasonal forecasts for the late 1990s and early 2000s) found there was some predictability associated with the probability of winter (May-October) rainfall occurrence but little associated with rainfall amount. Improved methodologies for generating simulations of multi-site, daily rainfall amounts conditional on atmospheric predictors have been developed and tested. In addition, a modified algorithm for resampling the daily rainfall amounts has also been developed and tested. This uses the gamma distribution, whereas the previous rainfall amounts model used the empirical distribution of the fitting data. Performance comparison has thus been undertaken for three versions of the rainfall amount module:

i. empirical distribution conditional on rainfall occurrence at neighbouring stations only (the original model);

ii. empirical distribution conditional on rainfall occurrence at neighbouring stations and atmospheric predictors (the extended model);

iii. gamma distribution conditional on rainfall occurrence at neighbouring stations and atmospheric predictors (the further extended model).

Although improvement in performance is not consistent across all stations, in many cases better reproduction of inter annual variability and daily probability density functions is evident for the model version using the gamma distribution. An example plot is shown (Figure 3.1). Note this assessment is for simulations driven with observed atmospheric predictors. This assessment has not been undertaken for seasonal forecasts due to the current limitations of the COCA2 hindcasts as outlined in P3.3 below.

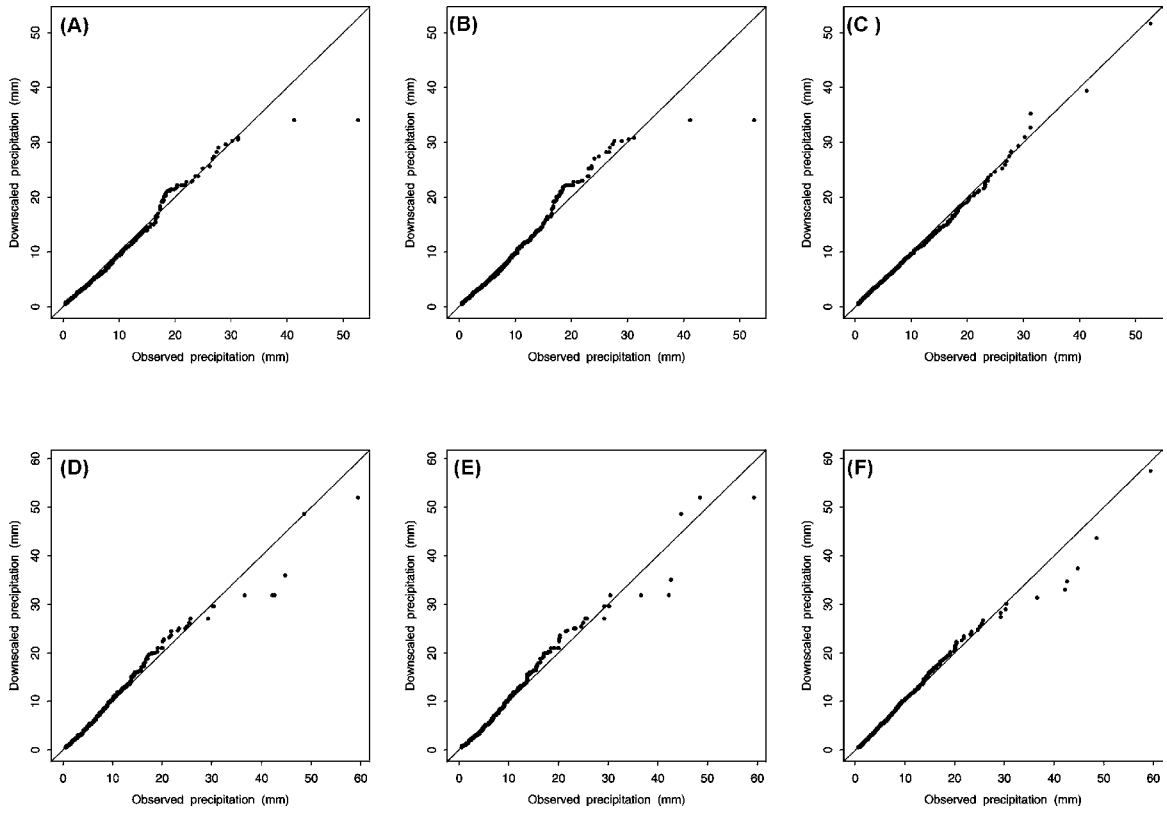

Figure 3.1. Daily rainfall amount reproduction for validation period using the three versions of the rainfall amount module (i to iii, left to right) for Kellerberrin (A to C) and Merredin (D to F).

Also, in collaboration with IRI and UCI, a modification to the current SD amounts methodology is being tested. This involves a version of the NHMM SD model that uses a mixture model, consisting

of a delta function (essentially a spike function) to model dry days (zero rainfall) and a mixture of exponentials to describe rainfall amounts on wet days. The model code has been obtained and testing on SWWA data is on-going.

Project 3.3: Processed atmospheric fields from hindcasts produced by CAR and IRI for Climate Prediction (USA).

This project aims to demonstrate the potential skill of dynamical seasonal forecasting schemes by downscaling the simulated atmospheric fields.

A hindcast is a retrospective forecast by a model based on a priori information. If it is applied to independent data, the model skill is a true measure of forecast skill.

The CSIRO seasonal prediction model is based on the CSIRO Mk3 coupled climate model and is referred to as COCA2. This model superseded the previous seasonal prediction COCA1 which was based on the CSIRO Mk2 climate model. COCA2 has been used to produce seasonal hindcasts for the 1980-2003 period as part of the CSIRO Water for a Healthy Country Flagship. The hindcasts data set begins in 1980 since, prior to this time, both observed SSTs and observed winds, which are required to initialize the prediction model, are regarded as less reliable. For each year, the model was initialized at 4 specific dates -January 1, April 1, July 1 and October 1 - and each time run forward for 12 months. This resulted in 24 separate 12-month predictions for January to December, April to March, July to June and October to September - in all, a total of 96 individual predictions. A 6-month lead prediction refers to the result for June from a January 1 start, the result for August from an April 1 start etc. (In practice, real-time predictions from the 1st of each month can only be completed mid-way through the month because of the delay in accessing the observational data and the time required to run the model. Technically speaking, real-time lead times will be one month less than those referred to here).

Here we report on the results of an analysis into the feasibility of predicting seasonal rainfall for SWWA region. In the first section we analyse the hindcast results involving monthly mean rainfall, mean sea level pressure (MSLP) and Southern Oscillation Index (SOI). In the second section, we consider statistical downscaling using a nonhomogeneous hidden Markov model which has been shown to be successful at reproducing the characteristics of multi-site, daily gauge precipitation (Charles et al., 2004)

Monthly mean results.

We begin by analysing the results of the April 1 hindcasts since these offer the possibility of winter rainfall predictions. We can refer to these as 3.5-month lead time hindcasts since this describes the time interval between the date of the hindcast and the mid-point of the target period (JJA).

Figure 3.2 compares the observed winter (JJA) rainfall totals over the SWWA region (see IOCI, 2002) with those hindcasted by COCA2 over a box area covering most of this region (116^{o} to 188^{o}E, 35^{o} to 32^{o}S). Also shown are the NCEP rainfall totals for the same box area. The COCA2 values underestimate the observations by about a factor of 2 while the NCEP values underestimate the observations by about a factor of 3. This is partly the result of averaging over slightly different regions

but is mainly due to the relatively coarse resolution of the COCA2 and NCEP models. It has been shown that the use of finer horizontal resolution provides a better representation of topography which contributes to higher rainfall amounts. Despite this bias, and the (partly related) reduced variability, the NCEP values are highly correlated (r =+0.88) over the 24-year period. It should be remembered that, while the NCEP values are model generated values, they are driven by daily observations of pressure, winds, moisture etc. and so are expected to closely represent the observations. In the case of the COCA2 values, the only drivers are predicted SST values and the correlation with the observations is only +0.19 indicating that, for this region at this time of year, there is essentially no predictability associated with the model rainfall values at a 3.5-month lead time.

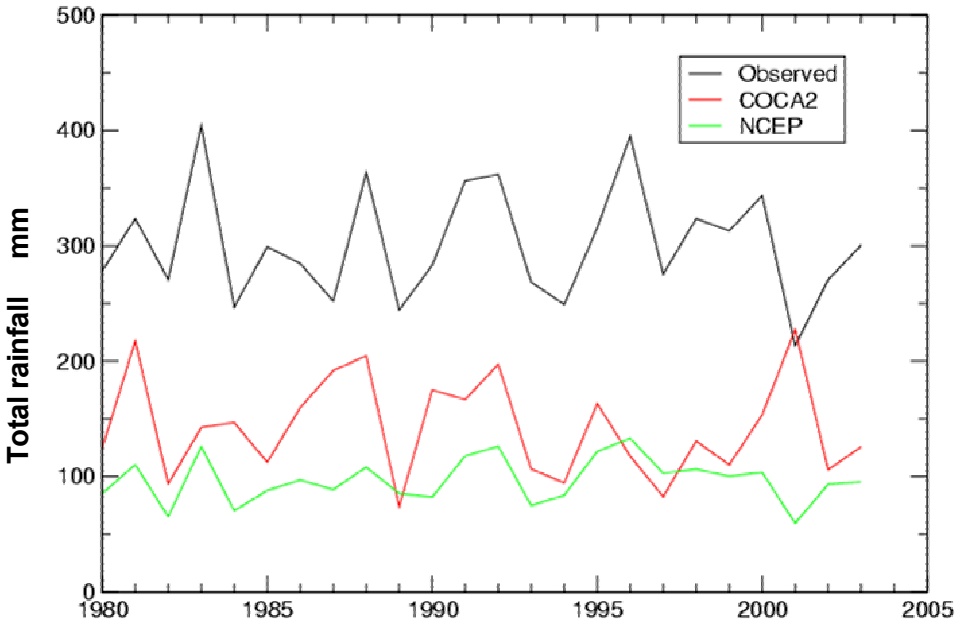

Figure 3.2: SWWA rainfall for winter (June to August) from observations, as hindcasted from COCA2 April 1 starts and from NCEP reanalysis data. The observed values are based on the Bureau of Meteorology 0.25° gridded data set averaged over the region as defined in IOCI (2002). The COCA2 and NCEP values are box-average values over the region 35°S to 32°S and 116°E to 118°E.

Previous studies have documented the strong inverse relationship between Perth mean sea level pressure (MSLP) and SWWA winter rainfall (Allan and Haylock, 1993; Smith et al. 2000; IOCIP 2002). In addition, statistical downscaling of large scale atmospheric fields also identified the north-south MSLP gradient and dew-point temperature depression at 850hPa as important factors which affect point-scale daily rainfall (Charles et. al, 2004). Charles et al. (2004) found that COCA1 statistically downscaled seasonal forecasts showed some predictability associated with the probability of rainfall occurrence, but little associated with rainfall amount. An alternative assessment of potential predictability is to focus on the skill of COCA2 in simulating mean winter MSLP over SWWA from the April 1 start dates. Figure 3.3. compares the observed and hindcasted JJA MSLP time series over the period 1980 to 2003. Apart from a bias of about +4 hPa, the COCA2 values show very little skill since the correlation coefficient between the two series is only -0.19. This in turn implies that little predictability would be associated with downscaling of the COCA2 MSLP fields over SWWA. This is confirmed by the fact that the COCA2 MSLP values are positively, rather than negatively, correlated (r=+0.20) with the observed rainfall (see Table 3.1).

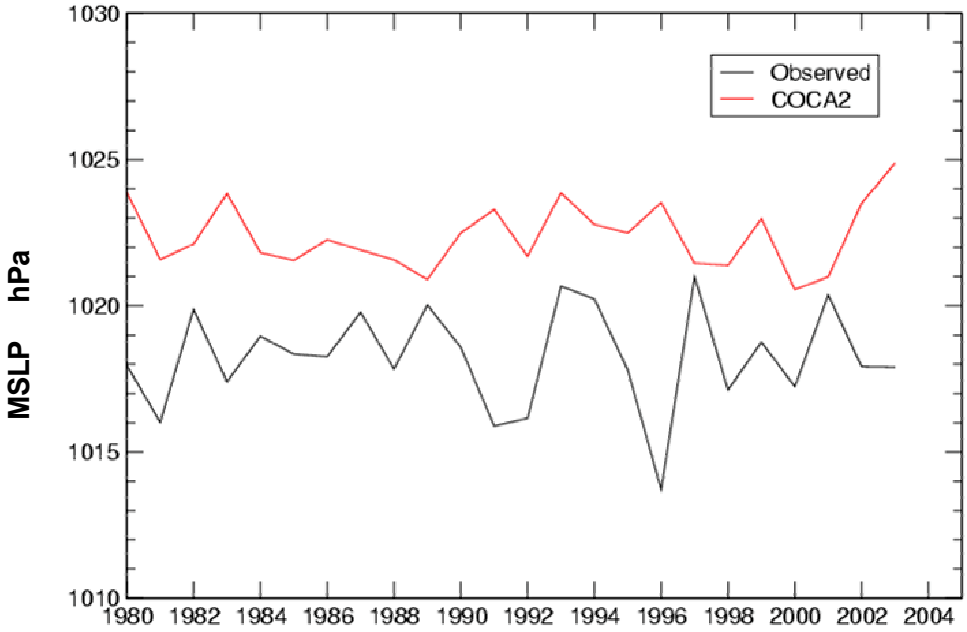

Figure 3.3: Mean sea level pressure (MSLP) for winter (June to August) from observations and as hindcasted from COCA2 April 1 starts.

Table 3.1: Correlations of various predictors with observed winter rainfall.

Observed winter (Jun-Aug) rainfall		
Index	**(1980-2003)**	**(1948-2003)**
NCEP rainfall	+0.88	+0.63
NCEP MSLP	-0.80	-0.82
NCEP RH850	+0.21	+0.40
SOI (obs)	+0.43	+0.34
COCA2 rainfall	+0.19	
COCA2 MSLP	+0.20	
COCA2 SOI	-0.13	
Observed late winter (Jul-Oct) rainfall		
Index	**(1980-2003)**	**(1948-2003)**
NCEP rainfall	+0.75	+0.59
NCEP MSLP	-0.80	-0.78
NCEP RH850	+0.16	+0.41
SOI (obs)	+0.39	+0.56
COCA2 rainfall	+0.18	
COCA2 MSLP	-0.14	
COCA2 SOI	+0.35	

While there is little evidence of predictability associated with the April 1 start dates, it could be expected that predictions beginning on July 1 might be more skilful, if only because they begin outside the autumn period which corresponds to the ENSO predictability barrier. Table 3.1 also shows the results from an analysis of the July 1 start date results for the late winter target season July to October. While the relationship between rainfall and the NCEP-derived time series is similar to before, the COCA2 rainfall predictions for late winter exhibit a similarly low correlation (r=+0.18) as for peak winter. However, there is some improvement in the correlations between MSLP and rainfall (r =-0.14 compared to +0.20) and between the model SOI and rainfall (r=+0.35 compared to -0.13).

Figure 3.4 compares observed rainfall anomalies with those derived from the model SOI values. This degree of correlation is comparable with that between the observed SOI and rainfall (+0.39, Table 1). Over the longer period 1948 to 2003, the observed SOI correlates much more strongly (+0.56) with the rainfall. This raises the possibility that the model SOI and rainfall relationship may also be stronger in the long term and that the 1980-2003 period may have been a relatively difficult period for predictions.

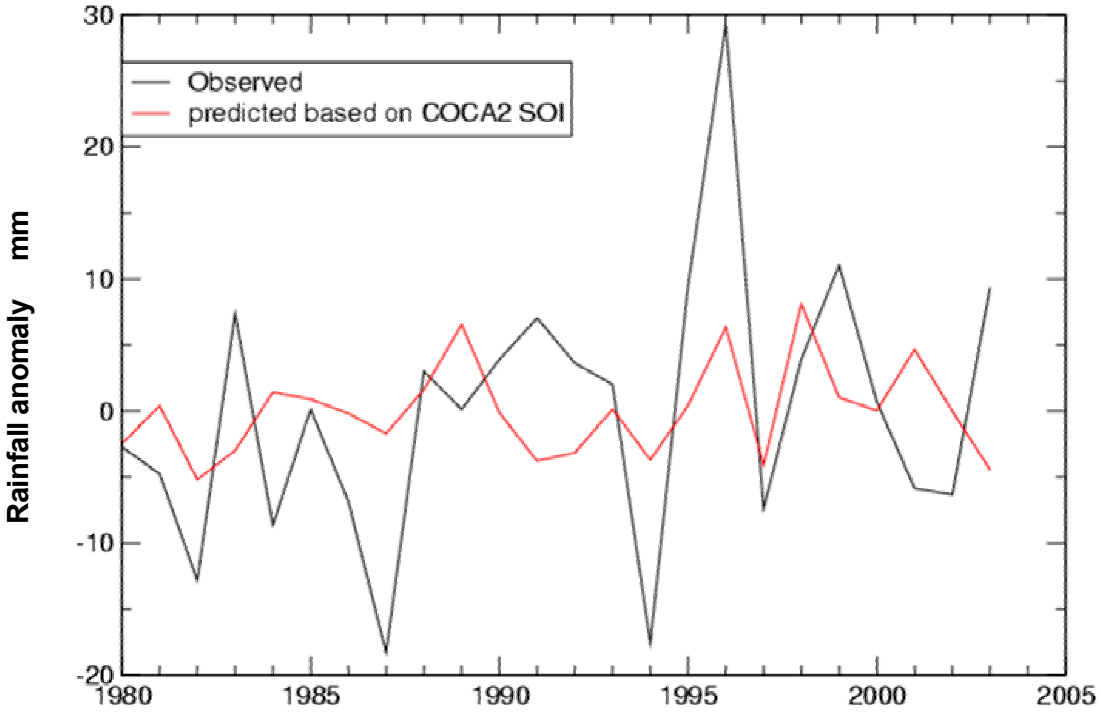

Figure 3.4: July to October rainfall anomalies from observations and as derived from the hindcasted SOI values from the July 1 start dates.

Downscaling of daily data

The statistical downscaling (SD) technique (see 3.1 and 3.2) , involves taking hindcast daily atmospheric variables data to produce hindcast daily, multi site rainfall series for the 30 stations across SWWA. MSLP is one of the SD model predictors and, as shown above, the COCA2 reproduction of winter MSLP variability over the region is poor. This is confirmed when comparing the MSLP from NCEP (i.e. observed), COCA2 and the previous generation COCA1 data (Figure 3.5 and Figure 3.6).

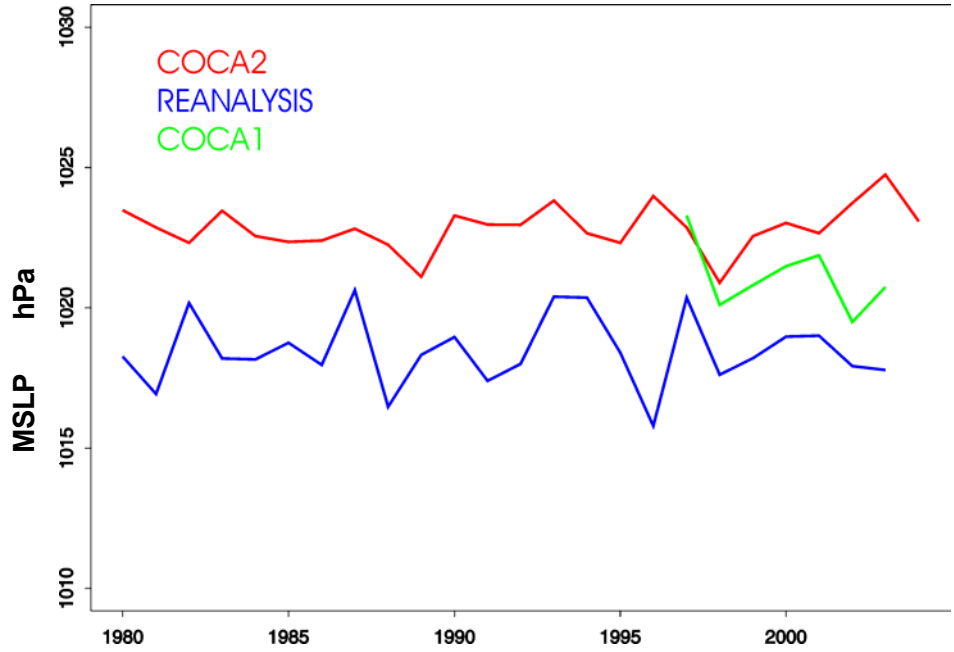

Figure 3.5: May to October MSLP over SWWA for the region used as a predictor for statistical downscaling model

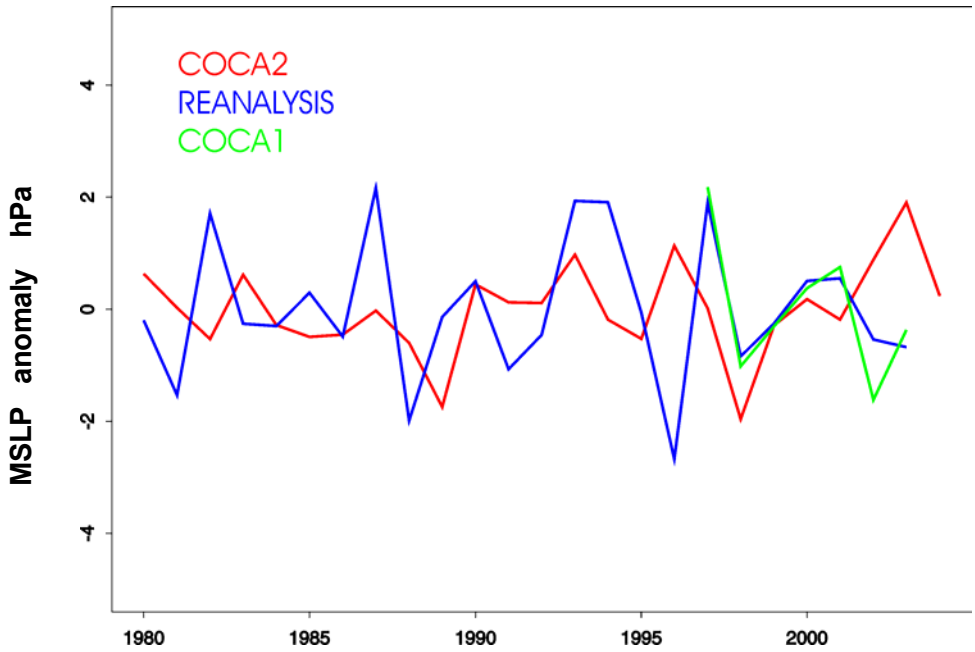

Figure 3.6: May to October centred MSLP over SWWA, as used in the statistical downscaling model

The COCA2 data appear less skilful than COCA1and, not surprisingly, the SD rainfall series obtained from driving the COCA2 hindcasts is correspondingly poor (Figure 3.7). However, this observation relates only to a sample of 7 years (1996 to 2002) and is possible that there is no real difference between the two models. Certainly we would expect the more recent version of the climate model (Mk3) with 18 levels in the vertical to perform better than the older version (Mk2) with only 9 levels. Experiments with the even higher resolution climate model CCAM (see Section 2.3) indicate that increased horizontal resolution can improve the mean winter rainfall over SWWA. However, they also indicate that SSTs alone are insufficient to reproduce interannual variability and that some knowledge of the changes to the large scale circulation (via the observed upper level winds) can dramatically improve the CCAM simulation of winter rainfall variability. Downscaling of the output from this type of experiment is expected to yield results consistent with those obtained using the raw NCEP data.

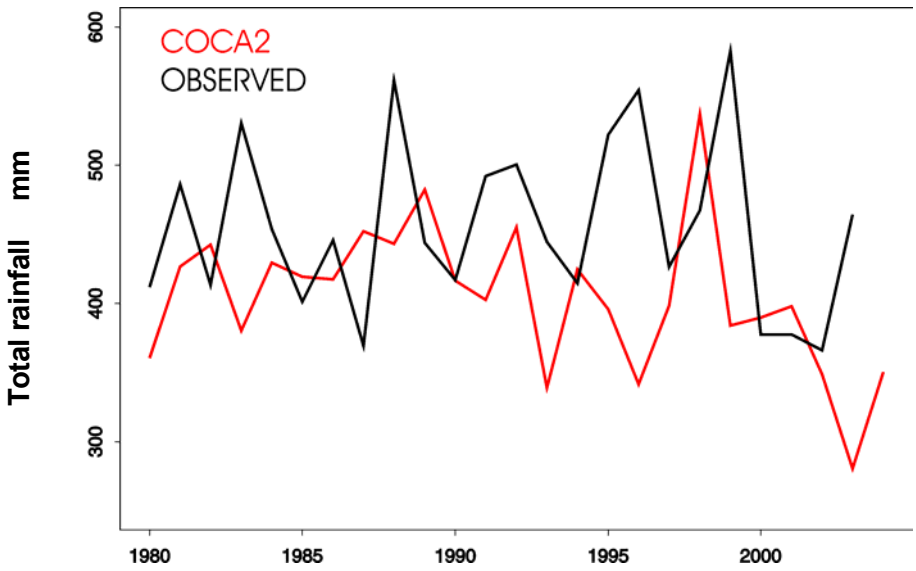

Figure 3.7: May to October observed and COCA2 statistically downscaled rainfall averaged over the 30 SWWA stations.

In summary the COCA2 hindcasts have revealed little skill for SWWA. However, these results are very premature. It is not clear whether the deficiencies result from inadequacies in the model or inadequacies in the SST and wind stress forcing. Furthermore, these results refer to the skill of just a single seasonal prediction model and it remains unclear as to how other models may perform. Consequently, it is planned to analyse the existing hindcast data sets from several other models. The DEMETER set comprises the results from several European seasonal prediction models and is available for analysis. It has been decided that accessing this data set may be more efficient than accessing the IRI data set.

Project 3.4: Analysis of synoptic events that lead to rainfall extremes.

This project aims to seek greater capacity to develop outlooks for extreme events

This project is work carried out under the CSIRO Water for a Healthy Country Flagship Program and may be of interest to State partners. It is arguable that the greatest potential impacts on the Australian environment, economy and society will be felt through changes in extreme weather events. The development of mitigation and adaptation strategies requires an understanding of potential impacts at relatively fine scales in the landscape. We also seek greater capacity to develop outlooks for extreme events.

This work is a first step in building models for predicting rainfall extremes in space and time, particularly under enhanced greenhouse effect scenarios. An extensive data audit has been completed to provide a relatively clean data set. The initial study region is focused on the Avon catchment, shown in Figure 3.8 below.

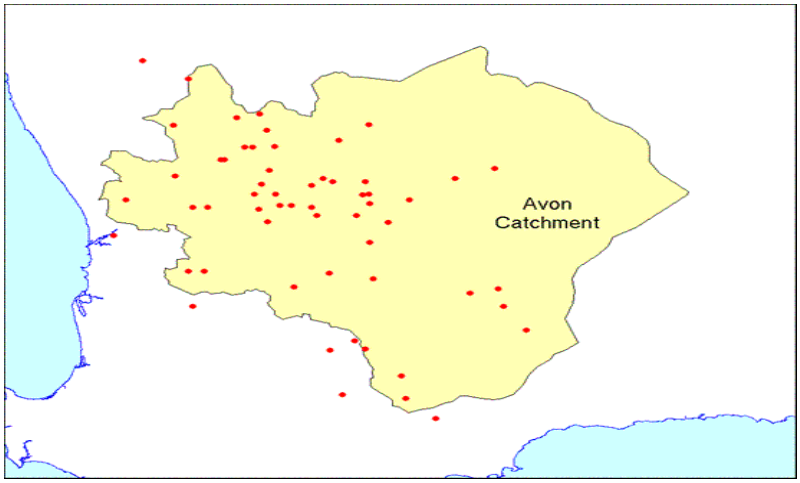

Figure 3.8: Study region, with rainfall measurement stations shown as red dots.

Methods from the statistical modelling of extreme events have been used to calculate thresholds beyond which a rainfall event is considered to be extreme; these thresholds have been mapped in Figure 3.9. A clear trend in rainfall is present, with north-west to south-east bands present. It would therefore be sub-optimal to make predictions station-by-station. Instead we have focused on building models for the spatial patterns, which are allowed to vary through time.

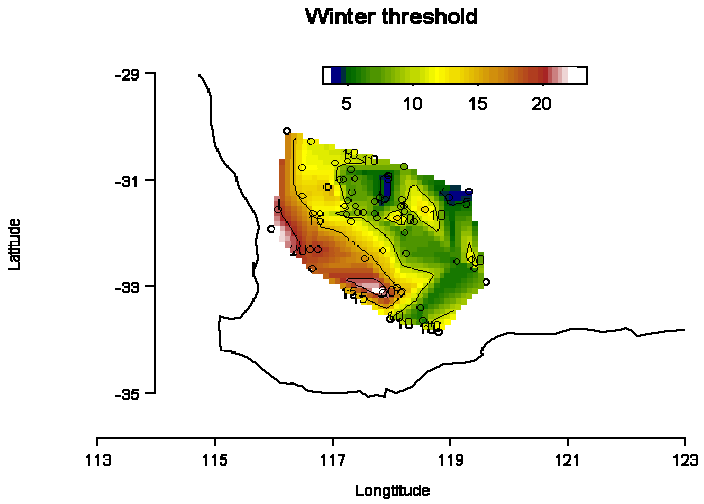

Figure 3.9: Thresholds defining extreme winter rainfall events, mapped across the Avon catchment.

A proto-type model has been developed that incorporates spatial relationships between rainfall stations, and this model has been used to explore trends in winter and summer maxima in the period 1950 to 2003. The modelling has allowed the team to take 'time slices' through the spatial patterns, and a set of these are shown below (Figure 3.10) for 1950, 1975 and 2000.

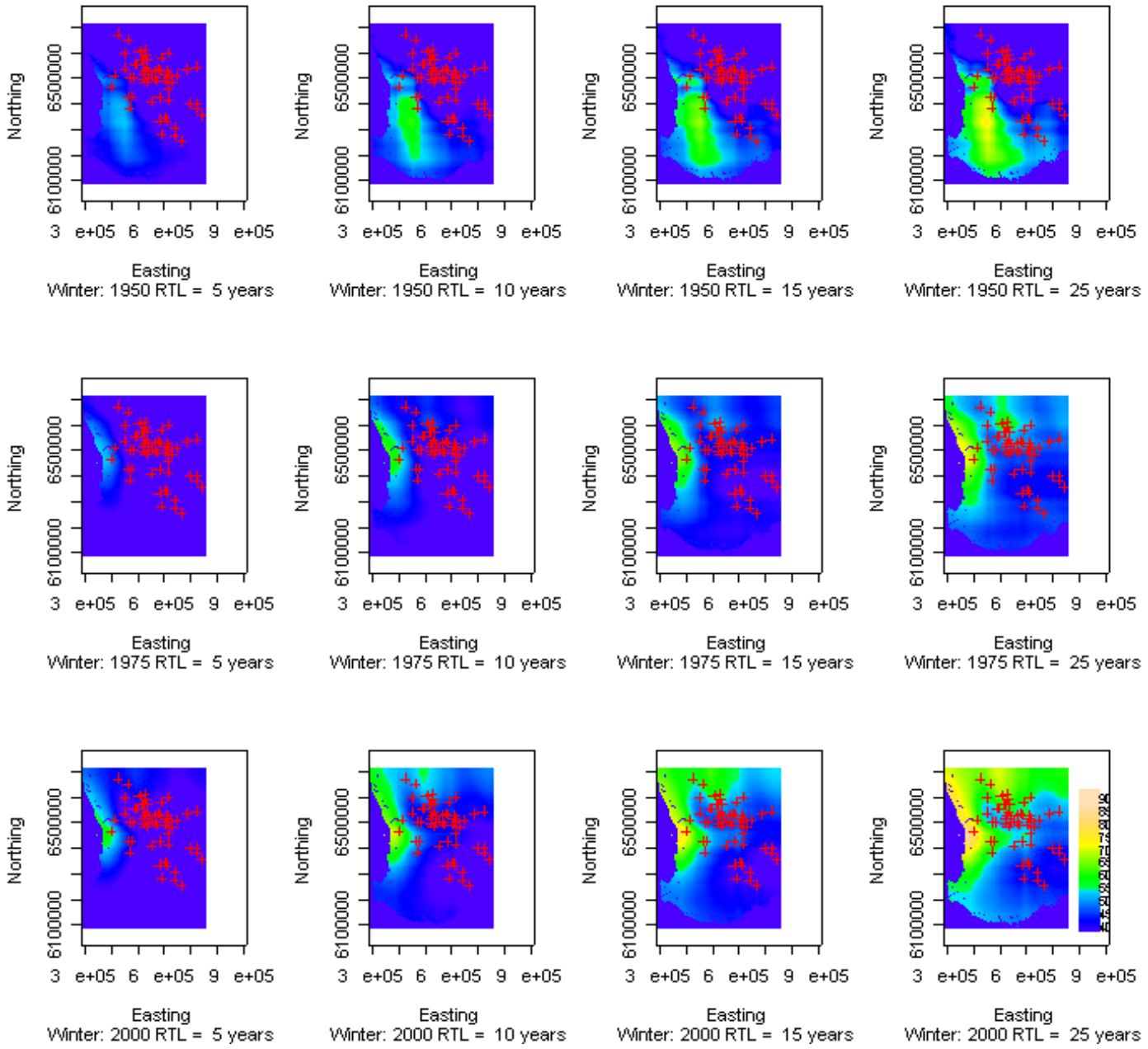

Figure 3.10: Time slices through the Avon catchment using the prototype spatial model for winter rainfall return periods.

The left-most column is for a return period of 5 years, and we see an increase in extreme rainfall to the west and north of the region, and a decrease to the south. This should be treated with caution however as this is beyond the domain of the rainfall stations. To the south-east there appears to have been a mild decrease. For a return period of 25 years we see a similar increase to the north, and a mild decrease to the south-east. An interesting feature is an expanding strip of increased extreme rainfall to

the northwest. There is a reasonable case for believing that this might be due to interactions with northwest cloud bands. There does therefore appear to be evidence for an increase already in rainfall extremes to the north, but perhaps with a more recent decline to the south. There is a statistically significant link to the Antarctic Oscillation, but more detailed work is required to determine the nature of this. The extreme events will next be explored for evidence of common synoptic patterns.

Project 4.1: Development of a hybrid model of factors influencing SWWA rainfall.

This project investigates methodologies to essentially couple together physical and statistical models to provide more effective forecasts

The coupling may range from driving a statistical model using physical predictors, to a fully integrated physical-statistical model. A key attractive feature of this approach to modelling is that uncertainty is fundamental, and forecasts are naturally derived as probability distributions. A challenge is that the development of these models is truly multi-disciplinary, but to date an ENSO forecasting scheme has been published, as well as some other geophysical applications.

There are two main approaches to seasonal forecasting in the literature. Perhaps the most common approach uses observations of the climate system as potential predictors of future climate in a statistical model. Predictors derived from physical considerations are sometimes incorporated. Another common approach is to use regional climate model, using observation to drive the model. Each approach has advantages and disadvantages. Statistical forecasting schemes are especially strong in their handling of uncertainty, whilst physical models provide a means to understand and accurately predict nonlinear behaviour.

We are investigating an entirely new approach in which a statistical model for the observations is coupled with one or more physical models. This is depicted in Figure 3.11 below.

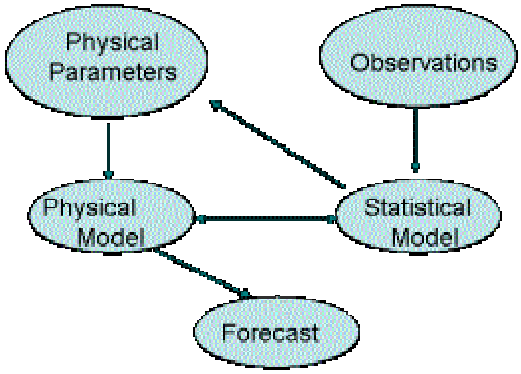

Figure 3.11: A schematic of a hybrid physical-statistical forecasting scheme.

We use the physical model to describe any known physics, whilst the statistical model captures any uncertainties and can be used to explore for relationships not captured by the physical model. The approach used updates knowledge of the physical parameters in the light of the observations, and generates forecasts as probability distributions. A seasonal forecasting framework using hybrid models has been developed by Campbell (2004b), and is available from the members' area of the IOCI web page.

In addition to the most general class of hybrid model, we have also investigated models we term 'feed-forward'. In so-called feed-forward models a statistical model for rainfall receives inputs from observations and climate models to produce forecasts. This allows both 'now' and lagged variables to be incorporated into the forecast model, with 'now' variables supplied by a climate model. A schematic of a feed-forward model is shown in Figure 3.12 below.

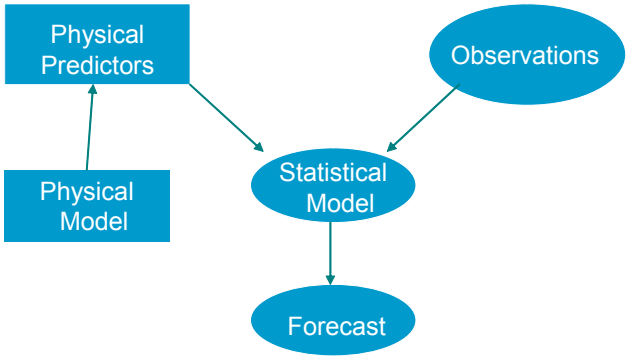

Figure 3.12: Schematic of a feed-forward hybrid physical-statistical forecasting scheme.

The most general class are may be termed 'feed-back' models, where the physical model is essentially calibrated against the available data using feedback from a statistical model for the observations. The forecasts produced integrate both physical and statistical elements.

Some results for feed-forward hybrid models

In general, coarse-scale climate models tend to underestimate rainfall amount and variability which can affect the reliability of any simulated changes – be they due to interannual variability or climate change. Statistical techniques can improve this situation, particularly when it can be shown that rainfall is strongly dependent on atmospheric variables such as MSLP, relative humidity, winds etc. For example, the SD approach (using the nonhomogeneous hidden Markov model) has been shown to be successful at reproducing the characteristics of multi-site, daily gauge precipitation over SWWA during winter (Charles et al., 2004). In these situations, model values for these variables can be used to derive rainfall estimates which are superior to the raw model rainfall estimates themselves.

The question we address here is whether feed-forward models using monthly mean variables can be used to improve rainfall estimates on the large scale. We focus on the use of monthly mean variables from both the NCEP reanalysis data and also the CCAM output from the run in which it is forced by observed upper level winds. Statistical regression models are firstly developed using the NCEP atmospheric data and the observed rainfall data from the training period 1948 to 1988. The models are then validated using data from the independent period 1989 to 2000.

Identifying important predictors

We began by considering the following variables as potential descriptors of SWWA JJA rainfall (R): mean sea level pressure (MSLP), precipitable water (PW), 850mb pressure level relative humidity (RH), 850mb pressure level specific humidity (SH), surface zonal wind (U), 200mb pressure level zonal wind (U200) and surface meridional wind (V). All values were calculated as seasonal means over the SWWA region. In addition, we also considered the Antarctic oscillation index (AOI) which is a measure of the MSLP gradient between mid- and high- latitudes. We then applied a statistical technique described by Friedman (2001) to select the most important variables. The ranking (in order of importance) of these variables is shown in Figure 3.13. The four most influential variables are, MSLP, RH, U200 and U. The percentages refer to the relative importance of each variable and it can be seen that MSLP (73%) dominates followed by RH (15%). There is effectively no significant relationship between rainfall and other variables on the seasonal time scale.

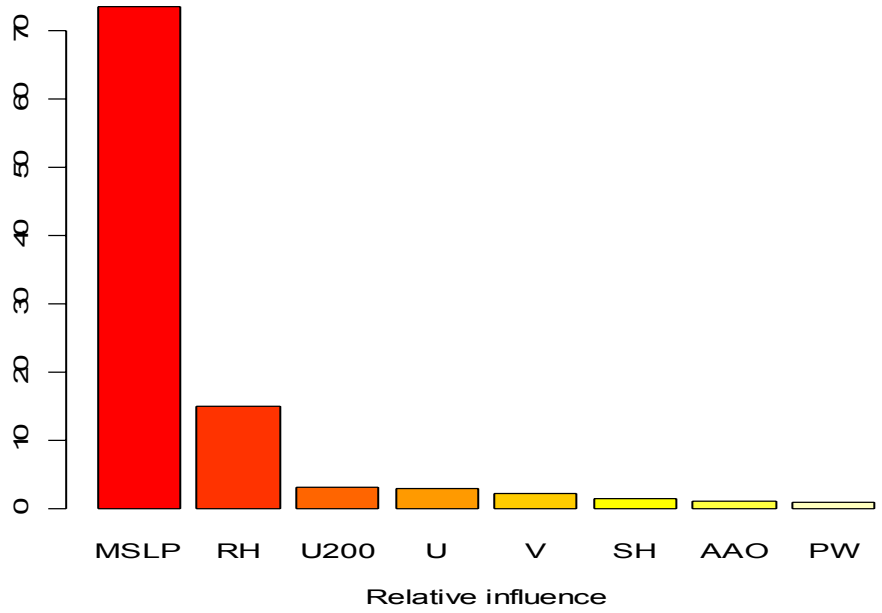

Figure 3.13: Relative influence of 8 atmospheric variables to SWWA JJA rainfall. In order of relative importance: mean sea level pressure (MSLP, 73%), relative humidity (RH, 15%), 200mb pressure level zonal wind (U200, 3%), surface zonal wind (U, 3%), surface meridional wind (V, 2%), 850 mb specific humidity SH (1%), Antarctic Oscillation Index AOI (1%), and precipitable water PW (1%).

The relationship between rainfall and both MSLP and RH is demonstrated by the scatter plots in Figure 3.14 and includes simple fits to the data using either simple linear regression or nonparametric piece-wise linear regression.

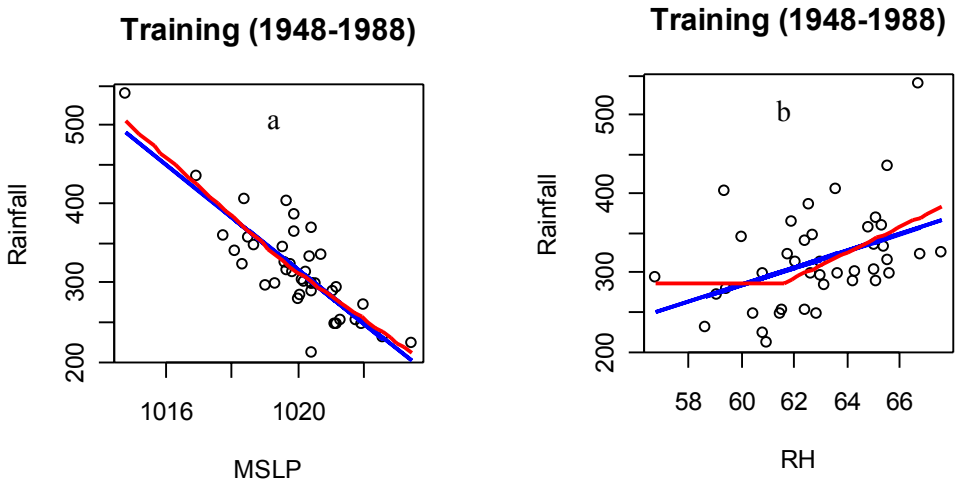

Figure 3.14: Scatter plot of SWWA JJA rainfall and atmospheric variables: (a) MSLP (1948 to 1988) and (b) RH (1948 to 1988). The blue lines represent a simple linear fit to the data, the red lines represent a non-linear fit to the data.

The statistical models

While it is apparent that MSLP and RH are the two most important variables, we nevertheless use multiple linear regression to derive a statistical model that also includes the next two most important variables (U200 and U). The resultant model based on the training period is

$$R_t = 34630 - 33.66\,\mathrm{MSLP}_t - 0.071\,\mathrm{RH}_t - 5.22\,\mathrm{U}_t + 0.8475\,\mathrm{U200}_t$$

In addition, we also derive a non-linear model using the technique described by Friedman (1991). The resultant model is

$$R_t = 326.72 - 30.74\left[\mathrm{MSLP}_t - 1019.6\right]_+ + 4.43\left[\mathrm{RH}_t - 61.58\right]_+ + 2.126\left[-\left(\mathrm{MSLP}_t - 1019.6\right)\right]_+ \left[\mathrm{U200}_t - 34.55\right]_+$$

where $[\square]_+ = \max(0, \square)$ and is a technique for handling threshold values, either side of which relationships between observed rainfall and predictors of rainfall can change.

In addition to applying the linear and non-linear models to the NCEP atmospheric variables, we have also applied them to the same variables as output by the CCAM model. In this case we have used the output from the CCAM experiment in which the upper level winds were nudged towards observations (see Section 2.3). This was shown to result in a good representation by CCAM of the observed rainfall and it could be expected that the associated atmospheric variables as simulated by CCAM would be similarly related to the rainfall as are the NCEP data. Figure 3.15 illustrates the relationship between the various data sets investigated in this section. In all, there are 7 time series to compare comprising the NCEP raw rainfall, the CCAM raw rainfall, the statistically derived rainfall derived from both the NCEP and CCAM atmospheric data using either the linear (LM) or non-linear (NLM) models, and finally the observed rainfall.

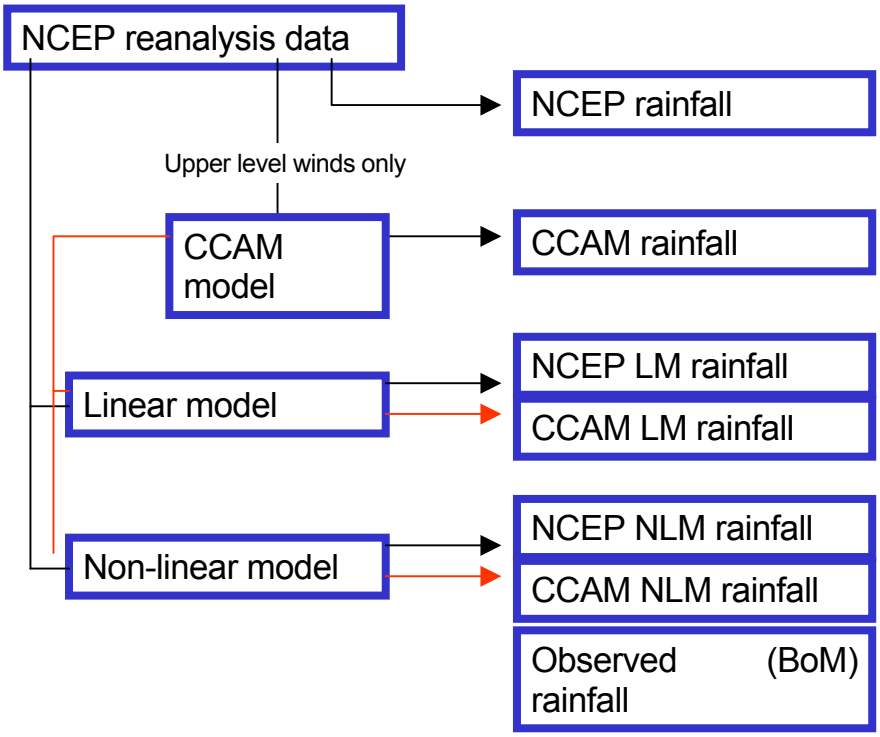

Figure 3.15. Sources of the seven rainfall time series analysed in this section (see text).

Figure 3.16 compares the seven rainfall time series over the training period 1948 to 1988. Again, the NCEP raw rainfall represents a serious underestimate but is highly correlated with the observations. The CCAM rainfall represents only a slight underestimate but is also highly correlated. The statistically derived rainfall time series, either linear or non-linear, all provide close fits to the observations.

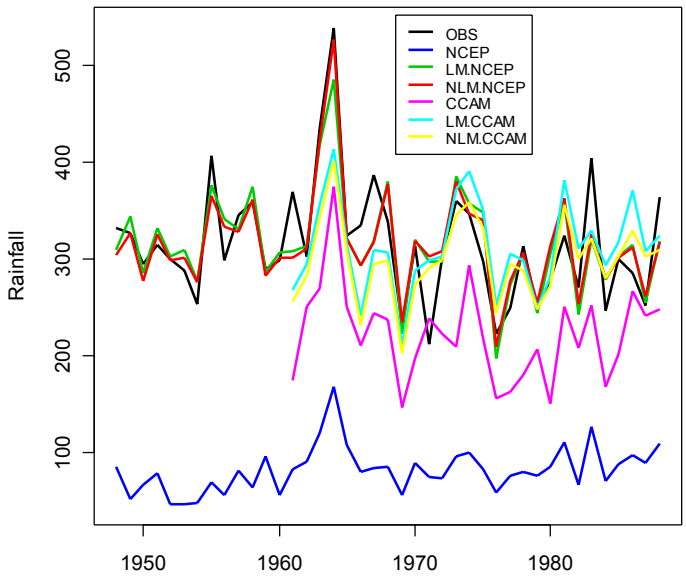

Figure 3.16: Time series of SWWA JJA rainfall over the training period 1948 to 1988 as observed (OBS), according to the NCEP and CCAM models, and as statistically derived from linear (LM) and non-linear (NLM) models using the NCEP and CCAM output atmospheric variables.

Figure 3.17 compares the seven rainfall time series over the validation period 1989 to 2000. Here we see a similar pattern, with the statistically derived time series all providing much improved fits to the observations compared to the raw NCEP or raw CCAM rainfall. This result suggests that the statistical models are robust and that the level of skill is evident over the training period is not artificial.

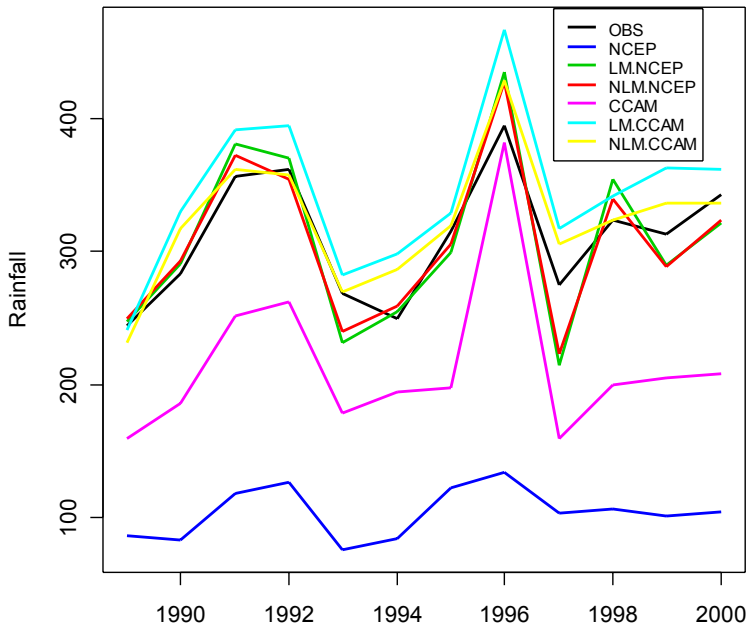

Figure 3.17: As for Figure 3.16except for the validation period 1989 to 2000.

The degree of improvement afforded by the statistically derived rainfall over the raw product is demonstrated by Figure 3.18. It represents a measure of skill by plotting the correlation coefficient versus the root mean square error. Optimum skill corresponds to points which fall to the top left-hand corner of the graph, and vice-versa. The insert records the individual values for both RMSE and r.

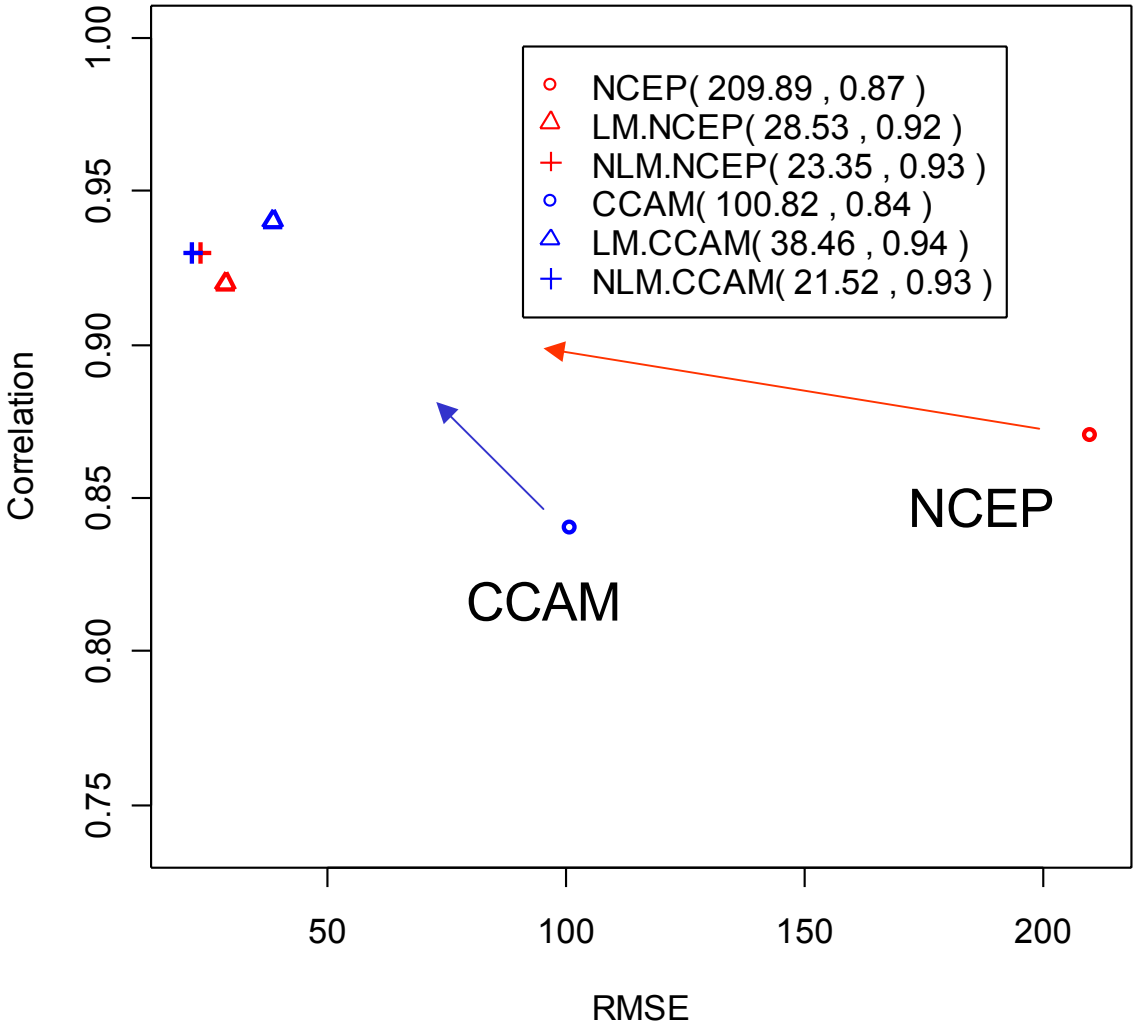

Figure 3.18: Skill associated with the various rainfall time series over the validation period 1989 to 2000. The arrows indicate the improvement in skill achieved by going from the raw (model) rainfall outputs to the statistically derived rainfall.

It can be seen that the raw NCEP rainfall product suffers from an RMSE of 210 mm representing the magnitude of the underestimate, but is still highly correlated (r=+0.87). The statistically derived rainfall products using NCEP outputs are characterized by much smaller RMSE values yet higher correlation values. The same pattern is seen when comparing the raw CCAM and statistically derived products using CCAM outputs. The arrows indicate the improvement in skill in going from the raw to the statistically derived products. The RMSE values lie between 20 and 40 mm while the correlation values are close to +0.93. In both cases, there is little to distinguish between the final products or the type of statistical approach adopted. If anything, the non-linear method appears to yield slightly lower RMSE values.

In summary, we have used both multiple linear and a simplified multivariate regression spline techniques to describe the relationship between the observed SWWA JJA rainfall and atmospheric variables – the most important being MSLP and RH. These were tested by partitioning the data sets into a training period 1948 to 1988 and an independent validation period 1989 to 2000. There is little to distinguish between the two techniques except that the non-linear technique may yield slightly lower RMSE values. Both techniques yield rainfall estimates which are superior to the raw NCEP and CCAM products. On this basis, it appears that winter rainfall for SWWA can be better estimated by using MSLP and RH within a simple statistical model rather than relying on the raw rainfall values generated by climate models. This result further confirms the application of statistical techniques

(including the statistical downscaling technique described in Section 2.2) as a means of better interpreting the results of climate change simulations. It also indicates that skilful rainfall predictions can be achieved by skilfully predicting MSLP and RH changes. Whether these are possible is still unclear.

Some results for feed-back hybrid models

Feed-back hybrid models are quite different. Rather than having a physical climate model supply 'data' to a statistical rainfall model, the physical and statistical models interact. Two types of model are therefore required: (1) statistical model for the observations to be used and (2) a physical model (conceptual or more detailed) of the physical processes involved. Note that to forecast rainfall we don't require a physical model of rainfall- the physical model would describe processes known to influence rainfall. The statistical model would make the link to rainfall data. A limiting factor in preventing implementation of a feed-back hybrid model has been the relatively poor state of knowledge of processes influencing SWWA rainfall. Work to date has collected together the available literature and developed a candidate framework for seasonal forecasting (Campbell, 2004b). The strength in these methods falls into 3 categories: (1) Constraining statistical forecasts by known physics; (2) Integrating observations and physics and (3) Direct production of probability distributions for forecasts and parameters.

As an example of the probability outputs that are available, consider the conceptual model of Suarez and Schopf (1988) for the ENSO phenomenon. ENSO may be thought of as a periodic disturbance in the ocean-atmosphere circulation in the Pacific Ocean. The model incorporates a negative, delayed feedback mechanism of given amplitude to explain the long time scale of ENSO (typically 2-4 years). The hybrid model would provide information on the unknown amplitude as a probability distribution, for example as shown in Figure 3.19. This figure shows the probability distribution for the amplitude prior-to and posterior-to incorporating available data. It turns out then to be very easy to produce forecasts as a probability distribution. For more details see Campbell (2004a).

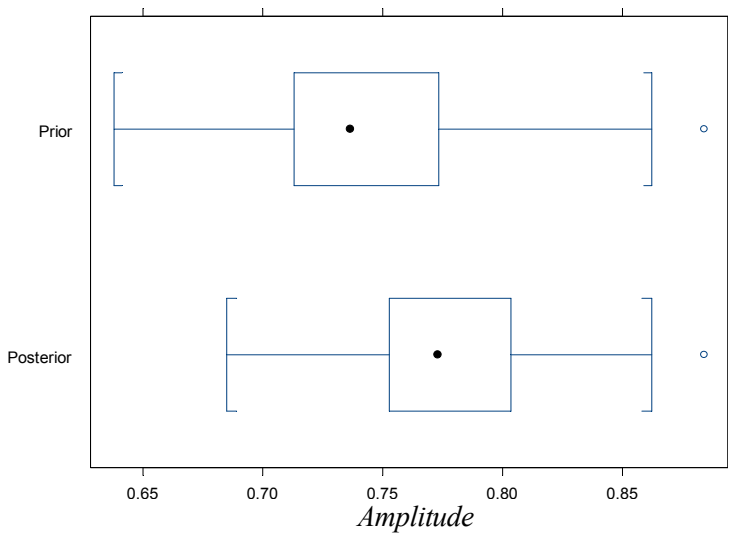

Figure 3. 19. Boxplots of prior and posterior samples for amplitude. The central box bounds the lower and upper quartiles, with the median (half-way point) shown as a black circle. The lines extend to the maximum and minimum values, except for outlying values which are flagged using clear circles.

Project 4.2: Customised nonlinear data mining tools.

This project seeks to make these tools available in a form suitable for climate applications

A key concern in climate forecasting is the identification of predictors of future climate, which is essentially a statistical modelling exercise. There are many widely available approaches to implement

standard statistical methods. However, climate applications are complicated in that the processes of interest are often nonlinear, and typically comprise multiple interacting components. There are tools available in the statistics and machine learning literature to address this problem, but they are not well known within the climate community. In addition, some tools have been developed as part of IOCI. This project seeks to make these available in a form suitable for climate applications.

Tools have been developed to incorporate physical thresholds into statistical modelling, and we may split these into techniques for identifying potential predictors with and without interactions. Methods that do not use interactions were developed and applied in IOCI Stage I, and these tools will be available later this year for web download. A range of tools have subsequently been developed and/or applied for handling interactions, mainly using a technique known as regression splines, and also for selecting predictors from large sets of candidate predictors. A problem with many existing tools is that they are hard to use, and really need to be used in combination. For example, it is natural to use boosting for predictor selection in combination with regression splines for model fitting, so we are developing a new version which combines both. We are also preparing a new tool that can handle multiple predictands, such as collections of rainfall stations.

REFERENCES

Allan, R.J. and Haylock, M.R., Circulation features associated with the winter rainfall decrease in Southwestern Australia, Journal of Climate, 6, 1356-1367, 1993.

Campbell, E. P. (2004a). Application of Bayesian Hierarchical Modelling to a Delayed Action Oscillator Model for the El Niño-Southern Oscillation. CSIRO Mathematical and Information Sciences, Perth, Technical Report 2004/139, 20 pp.

Campbell, E. P. (2004b). An introduction to physical-statistical modelling using Bayesian methods. CSIRO Mathematical and Information Sciences, Perth, Western Australia, Technical Report 2004/49, 18 pp.

Charles, S.P., Bates, B.C., Smith, I.N. and Hughes, J.P. , 2004, Statistical downscaling of daily precipitation from observed and modelled atmospheric fields, *Hydrological Processes*, **18**, 1373-1394.

Friedman, J.H., 2001: Greedy function approximation: A gradient boosting machine. The *Annals of Statistics*, 29(4).

Friedman, J.H., 1991: Multivariate adaptive regression splines (with discussion). The *Annals of Statistics*, 19, 1-141.

IOCI (2002). Climate variability and change in south west Western Australia. http://www.ioci.org.au/publications/pdf/IOCI_TechnicalReport02.pdf. ISBN 1-920687-03-3.

Ridgeway, G., 1999: The state of boosting. *Computing Science and Statistics*, 31, 172-181.

Smith, I.N., McIntosh, P., Ansell, T.J., Reason, C.J.C., McInnes, K. , 2000, Southwest Western Australia winter rainfall and its association with Indian Ocean climate variability, *International Journal of Climatology*, **20**, 1913-1930.