NONLINEAR STATISTICAL METHODS FOR CLIMATE FORECASTING

Edward P. Campbell[1], Bryson C. Bates[2] and Stephen P. Charles[2]

CSIRO

[1]CSIRO Mathematical and Information Sciences

[2]CSIRO Land and Water

Phase 1 report to the

Indian Ocean Climate Initiative

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

During the first year of the Indian Ocean Climate Initiative (IOCI), CSIRO Mathematical and Information Sciences (CMIS) and CSIRO Land and Water (CLW) have examined the use of statistical methods for climate forecasting in south-west Western Australia (SWA). We have progressed according to the following steps:

1. Survey the climate forecasting literature to assess the current state of knowledge from a statistical perspective;

2. Identify the main issues that need to be addressed using statistical methods, and assess the suitability of methods currently in use;

3. Identify a research program that has the potential to deliver enhanced statistical forecasting tools at inter-seasonal, inter-annual and decadal time scales.

Our progress against each of these themes has been as follows.

1. We have conducted an extensive review of the climate forecasting literature, and it is intended that this will be submitted to a peer-reviewed journal. The objectives of the review were: to gain an appreciation of the physical processes involved, particularly their impact on statistical modelling and analysis; to assess the current use of statistical methods; and identify potentially beneficial opportunities both for statistical research and enhanced application of contemporary statistical methods.

2. Perhaps the most pervasive method in climatology is that of empirical orthogonal function (EOF) analysis. This method assumes stationarity in time explicitly, but also implicitly assumes that the geophysical field of interest is spatially homogeneous. We discuss some more modern approaches to spatial-temporal modelling in this report. The criterion for extracting orthogonal functions is also purely statistical, and the patterns that emerge are usually predictable *a priori*. A great deal of work has been done on the use of oblique rather than orthogonal rotations in an attempt to obtain greater physical insight. That is, we no longer insist on the empirical functions being orthogonal by changing the optimisation criterion used to extract them. We suggest projection pursuit as a potentially more useful general framework. In this approach we

use physical knowledge to define an optimisation criterion, which is then optimised to derive low-dimensional, interpretable descriptors of complex, high-dimensional data.

Considerable use is made of regression and correlation methods, where linear relationships are typically assumed. This assumption is of particular concern since many potentially predictive relationships will never be discovered. For example, a quadratic relationship between two variables has a correlation of zero. However, if we correctly identify the quadratic relationship then we naturally find a high correlation between fitted and observed values. An emphasis on statistical model building using appropriate tools is required, and we suggest a number of these.

Statistical methods are typically of most value in data rich/knowledge poor scenarios or where uncertainty and noise are present. Climate forecasting is a combination of both of these, in varying degrees depending on the particular application. Therefore there is a clear need for appropriate statistical methods. Our review of the climate forecasting literature yielded two key themes: climate processes are inherently non-stationary and nonlinear. The statistical methods applied are typically linear in nature, and require assumptions of stationarity to be meaningful. Viewed in this light it is unlikely that such methods will reveal more than superficial physical insights and predictability.

3. We have sought to develop a statistical framework that has at its foundation a nonlinear dynamical statistical model, which we call a statistical-dynamical model. Results of preliminary model fitting to monthly rainfall data are very encouraging, with fundamental properties of the rainfall series reproduced with some degree of success.

Future work will involve:

1. Completion of the rainfall and sea surface temperature (SST) exploratory data analysis commenced in Phase 1 of IOCI;

2. Development of the nonlinear modelling approach within a Bayesian statistical framework. This effort will be focused initially on rainfall forecasting;

3. Apply a range of contemporary statistical methods to identify potential ocean and atmospheric rainfall predictors for the nonlinear model;

4.   Develop an approach to predictor selection in nonlinear models using the reversible jump Markov chain Monte Carlo methodology;

5.   Apply the nonlinear modelling approach to the downscaled 1000-year CSIRO9 GCM run.

Outcomes from this work will include:

1.   A Bayesian statistical framework that uses probability distributions to represent uncertainty about model parameters. This will provide a probabilistic risk assessment approach to climate forecasting. An example output would be a predictive probability distribution for winter rainfall some months ahead;

2.   Physical insights into the nonlinear mechanisms that generate rainfall. In particular, the modelling results will provide some information on the factors that influence the switching of rainfall between different regimes;

3.   A statistical approach to identifying important climate predictor variables, applied to rainfall in the first instance; and

4.   Statistical insights into the nonlinear mechanisms producing GCM output, which may be helpful when interpreting observational data.

Our proposed research linkages are:

1.   *Bureau of Meteorology (Research Centre and Perth Regional Office)* – Identification of potential rainfall predictors; physical interpretation of nonlinear modelling results; forecasting skill comparison of nonlinear models and existing approaches.

2.   *CSIRO Atmospheric Research* – Application of nonlinear modelling approach to GCM output; physical interpretation of data-based nonlinear modelling results.

3.   *CSIRO Land and Water* – Downscaling of GCM output; physical interpretation of nonlinear modelling results.

4.   *CSIRO Marine Research* - Identification of potential rainfall predictors; physical interpretation of nonlinear modelling results.

ACKNOWLEDGMENTS

1. INTRODUCTION

The work conducted in Phase 1 has explored the connections between physically and statistically based modelling of climate processes. The only assumption we have made is that there is a need for both, in the light of our current understanding of the mechanisms of climate variability. This is in keeping with developments in the nonlinear time series community towards a fusion of statistical and dynamical systems methodologies. Such an approach recognises that both deterministic and stochastic elements have a role to play in climate modelling, and that these elements may interact in potentially complex ways. Perhaps the strongest exposition of this trend was provided by Tong (1990).

The remainder of this report is structured as follows. In section 2 we provide a brief review of the climate forecasting literature from a statistical perspective, draw some conclusions and suggest some potentially useful methods that have thus far received little or no exposure. In section 3 we summarise our data analyses so far, which have been used to develop a nonlinear statistical framework that is described in section 4. We provide some conclusions and directions for future work in section 5. A glossary of terms and a list of acronyms are supplied as appendices.

2.   A REVIEW OF THE CLIMATE FORECASTING LITERATURE

2.1   Survey of Statistical Methods

2.1.1   Correlation and Regression Analysis

Extensive use has been made of correlation and regression methods in attempting to establish evidence of "teleconnections" (Nicholls, 1991), which are climate events that are related whilst being far removed in space and/or time. Teleconnections are a feature of the Southern Oscillation because of the vast spatial scale on which it takes place. Nicholls (1991) notes a characteristic of the El-Niño Southern-Oscillation (ENSO) "… droughts in India, North China, Australia and parts of Africa and the Americas tend to occur approximately simultaneously …". Lagged correlations have been used to detect time-lagged climate effects (e.g. Drosdowsky, 1993c). Linear regression methods have been used by many authors to model the relationship between the Southern Oscillation Index (SOI- normalised difference between sea-surface air pressure at Darwin and Tahiti) and a variety of phenomena of interest in economic terms. Note that temporal correlation is often ignored in such analyses, neglecting potentially important information in the data.

Nicholls (1986) used lagged correlations to show a statistically significant correlation between Australian sorghum yield and Darwin pressure, an indicator of the Southern Oscillation. Nicholls found that stronger relationships exist between yield and Darwin pressure *trends*, rather than absolute or scaled readings. After the removal of trends, due to improved technology and introduction of new cultivars for example, a linear regression model relating yield to the January-March to June-August trend explained about 50% of the yield variation. Given that the crop is planted between October and February this is a potentially very useful management tool. Indeed, a slightly improved model incorporating the trend up to October was rejected since a prediction at the time of planting was considered less useful from a management perspective.

Rimmington & Nicholls (1993) examined the use of the SOI to forecast wheat yields in Australia. As found by Nicholls (1986), trends in the Southern Oscillation, as measured by SOI, had more predictive power than the observed value. Rimmington & Nicholls (1993) note that "A large proportion of the year-to-year variation of Australian wheat yield is due to variation in the available soil moisture which is determined by the balance of rainfall and

evapotranspirative losses." They go on to state that "Much of the inter-annual variation in rainfall over the Australia (*sic*) wheat-belt is related to the … (ENSO) phenomenon … Variations in temperature, wind and therefore evapotranspiration may also be related to ENSO events." However, the pre-existing soil moisture profile is not used in model building. The skill level, as measured by R^2, peaks at 36% for Queensland and falls as low as 6% for South Australia. This suggests that whilst SOI alone accounts for a significant share of the variation in Australia wheat yield, a significant proportion is left unexplained.

An alternative to linear regression is provided by Russell *et al*. (1993), who use a methodology known as Alternating Conditional Expectations (ACE). This method identifies nonlinear transformations of the data that allow a linear regression model to be applied to the transformed data. Russell *et al*. (1993) note that variables besides SOI are potentially of use in building a predictor of rainfall, noting in particular seas-surface temperatures and that ACE can be applied when multiple predictors are available. They only apply ACE to SOI data however. The level of complexity of including sea-surface temperature (SST) is not great, as is claimed in this paper- at least if suitable potential predictors can be extracted from the SST data. We return to this issue in section 2.2.1.

A number of interesting observations on the use of correlation and regression methods were made by Drosdowsky & Williams (1991), and are worth quoting here:

> '… these anomalies (geopotential height) are not symmetric, being more intense during the positive (SOI) phase. This may be due to the *nonlinear* nature of the latent heat forcing over northern Australia and Indonesia. During ENSO (negative phase) events convection is reduced, but not entirely absent over this region, while the increases in precipitation during the positive phase may be many times above the mean.

> … Locally, however, there are significant "nonlinear" deviations. The most significant deviations of these occur during summer in the geopotential height field over the Tasman sea, …

> … The effects of these nonlinearities on the evolution of extreme events need to be studied further, as do the implications for linear-regression forecasting schemes using the SOI as a predictor.'

This acknowledges a clear need to model inherently nonlinear phenomena.

The correlation coefficient is ubiquitous in climate forecasting research and is often used to search for relationships between quantities of interest. It must be remembered that the correlation coefficient is a useful measure only of linear relationships. For instance, it is quite possible for a correlation of 0 to be found when a perfectly deterministic relationship exists between two variables. In light of the above quote, the correlation coefficient is not a useful screening tool in seeking general relationships between meteorological variables. We should instead apply methods capable of detecting general relationships, rather than assuming a linear relationship *a priori*. There are a number of possibilities in contemporary statistical practice, and we describe some of these in section 2.2.1.

2.1.2 Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOFs) feature at least as often as the correlation coefficient in climate forecasting research. Without delving into the theory too deeply, the reason for this is quite straightforward. In general, physical fields, such as SST, exhibit complex behaviour. EOFs characterise the field as being a weighted sum of components that are mutually independent (orthogonal) of one another. Each of these is known as an empirical orthogonal function (EOF). In principle this greatly aids interpretation since each component can be interpreted without reference to the others.

There is a straightforward theory for calculating EOFs, based on the Karhunen-Loève expansion (Freiberger & Grenander, 1965). This technique also provides a method for ranking the EOFs in terms of their explanatory importance. An important assumption though is that the meteorological field under study is stationary in time, which may be considered unrealistic.

Statistically speaking, the EOFs are likely to follow a predictable pattern. The first (most important) EOF will tend to be a large scale average, whilst the second EOF will tend to describe large scale contrasts (Drosdowsky, 1993a). This is one example of the domain shape dependence of this technique, described in some detail by Richman (1986). The technique is therefore of limited climatological usefulness, which has led to rotation of EOFs being proposed as an approach to increasing interpretability. Jolliffe (1989) suggested that EOFs having approximately equal explanatory importance be rotated, but there is no single

rotation method available, as is clear from the extensive review of Richman (1986). These issues are discussed in more detail in section 2.1.3 below.

EOFs, and their rotated versions, are found by optimising explained variance, an essentially statistical criterion, to capture variability in a small number of components. A potentially more informative approach is known as projection pursuit (Jones & Sibson, 1987), which requires a climatologically interesting optimisation criterion to be defined. This criterion replaces the variance criterion used in EOF analysis. Given a sensible choice, the resulting components are, by definition, 'interesting'. We discuss this approach in section 2.2.3.

There are many applications of EOFs published in the meteorology and climatology literature. Cohen & Jones (1969) used the Karhunen-Loève expansion to examine a regression model defined on a continuous random field. The example they use to illustrate the methodology involves a response variable being the temperature at National Airport, Washington, D.C. and the random field is the 700 mb height observed on a grid of points. Data were observed every 12 hours for 2 years. Obled & Creutin (1986) cast the method in terms of optimal interpolation, and provide a case study on interpolation of rainfall fields. Nicholls (1987) discusses two published applications of EOFs to study the nature of Southern Oscillation teleconnections. This paper also noted the use of extended or complex EOF analysis that incorporates temporal as well as spatial relationship. Barnett (1983) described the theory in some detail, which is identical to the method termed 'complex PCA' by Horel (1984). A further recent example was provided by Burkhardt & James (1998), who used an extended EOF (EEOF) analysis to measure the intensity of a storm track.

2.1.3   Principal Component Analysis

Principal Component Analysis (PCA) arises indirectly in the literature as a method of estimating EOFs (see section 2.1.2), and directly as an exploratory statistical tool. We discussed EOFs above, and concentrate here on its use in exploratory analysis.

The data in meteorological applications typically comprise 3 dimensions: Stations (location), Time and the variables observed so the data matrix may be thought of as a cube. There are two distinct modes of analysis depending on whether we consider the observations to be stations for a fixed time (S-mode), or if we consider each individual time to be a variable and each station an observation (T-mode), thus revealing temporal associations. This is discussed by Drosdowsky (1993a), who also cites a number of such analyses in the literature.

Drosdowsky also discusses the rotation of principal components, given that principal components tend to provide a first component with "generally large loadings", with structures of interest in the second component that are often predictable, such as dipoles. This restricts the usefulness of the method, as described in section 2.1.2.

Nicholls (1989) used PCA to simplify the pattern of Australian rainfall, and then examined correlations of the first 2 principal components with sea-surface temperature (SST). A comparison of Varimax and Promax rotations was made, with little difference found in this case. Smith (1994) examined the ability of PCA to predict Australian winter rainfall using Indian Ocean SSTs, employing principal components regression to find relationships between SST and rainfall principle components. Paterson *et al*. (1978) used PCA in a very different application to classify regions of the south-west of Western Australia so that experimental locations could be chosen that represented the range of experimental locations in the state.

An alternative to PCA is Complex Principal Component Analysis (CPC), a good introduction to which is provided by Horel (1984), and is also known as 'Complex EOF analysis' (or Extended EOF, EEOF), seemingly interchangeably. A key advantage of CPC over real PCA is its ability to detect travelling waves, where PCA can only detect standing oscillations. By changing the optimisation criterion used to derive principal components we can generalise the technique to nonlinear principal component analysis. This method has been applied in climatology by Monahan (1998).

2.1.4 Singular Value Decomposition

Singular value decomposition (SVD) has often been viewed as competing with canonical correlation analysis (CCA - see section 2.1.6). It is used in climatology to examine covariance relationships between two physical fields, such as SST and barometric height anomaly. This is done by finding linear combinations of the two fields that have maximal covariance. Both CCA and SVD therefore assume that we wish to detect a linear relationship between the two fields. Cherry (1996) compares the two methods and makes the important point that because CCA is based on correlation, the re-scaling involved might obscure some effects. He argues that both methods should be used so that the impact of the re-scaling in CCA can be made.

There are many applications of SVD in the literature, such as Wallace *et al*. (1992) who compare the method with PCA and then use SVD to calculate canonical correlation vectors.

They also use a technique known as combined PCA (CPCA), described in more detail by Bretherton *et al*. (1992), in which the data from two fields are combined. Bretherton *et al*. (*op. cit.*) also describe a technique known as second field PCA (SFPCA), in which principal component amplitudes (loadings) are correlated with the second field.

2.1.5 Cluster Analysis

Cluster analysis is a technique for grouping multidimensional observations. Drosdowsky (1993a) used cluster analysis in an attempt to regionalise Australian rainfall anomalies, comparing it to a rotated principal component analysis. The results showed that for a small number of clusters the regional structure was somewhat different using the two techniques, whilst for a larger number of clusters the regional structure was essentially the same. The likely explanation for this is that the two techniques are driven by different optimisation criteria. Cluster analysis was also employed by Wolter (1987) in an exploratory data analysis mode. Conventional cluster analysis is essentially limited to this role, but there are developments in the literature addressing model-based clustering techniques that have the potential to be more useful. See, for example, Banfield & Raftery (1993) and Jolliffe (1998).

2.1.6 Canonical Correlation Analysis

Canonical Correlation Analysis is a technique for exploring relationships between a collection of potential explanatory variables (**X**) and a set of response variables (**Y**). Nicholls (1987) applied CCA to a set of explanatory variables comprising bimonthly Darwin pressures over a 22 month period. The particular set of 22 months was chosen to encompass a typical SO cycle of about a year and to locate the period of strongest correlation (July to December) in the middle of the set. This also enabled lagged correlations to be detected. The response variables were Tahiti Pressure, Willis Island air temperature and south-east Australian rainfall.

Nicholls (*op. cit.*) noted the usefulness of this procedure in detecting and exploring teleconnections. CCA will only detect linear relationships however, and may be thought of as a generalisation of multiple regression to the multi-response case.

2.1.7 Linear Discriminant Analysis

Drosdowsky & Chambers (1998) used linear discriminant analysis (LDA) to classify rainfall categories in terms of predictors selected via multiple regression. The first stage in their

analysis was to reduce the available sea-surface temperature anomaly (SSTa) data using a PCA. A step-wise multiple regression relating rainfall to principal components of SSTa and SOI was conducted, which gave reasonable predictions. It was however preferred to produce a probabilistic forecast, so the predictors identified in the regression modelling were used in a discriminant analysis. A better approach would be to apply step-wise variable selection directly in the discriminant analysis procedure. Such procedures are now widely available, as Proc StepDisc in SAS® software for example.

Drosdowsky & Chambers (1998) noted an unexpected poor performance of the discriminant analysis procedure in this case, and stated two potential explanations for this. First, the selection of too many predictors in the discrimination model and, secondly, "those selected by the multiple regression selection may not work best with the inherently nonlinear discriminant analysis procedure." The technique may be thought of as nonlinear in the response variable, which is categorical. However, this is a restrictive form of nonlinear model - the discriminant function is in fact linear when the rainfall category covariance matrices are assumed to be equal and quadratic when they are allowed to be different and two rainfall categories are allowed.

Note that the phenomenon of too many predictors in the discrimination model noted by Drosdowsky & Chambers (*op. cit.*) is an inadequacy of the LDA technique when many potential predictors are available. This situation is made worse if these predictors are also highly correlated. A potentially more useful approach is based on penalised discriminant and regression methods, which are discussed in section 3.5 below.

2.1.8   Analogs

The idea behind analogue forecasting is a simple one: given current meteorological conditions, we search back through the data record for the closest match. Our forecast is then the outcome from this match. There are numerous applications of this idea in the literature, an example of which is provided by Stone *et al.* (1997). This approach is based on identifying SOI phases, which has also been discussed by Casey (1995).

Nicholls & Katz (1991) note a number of world meteorological agencies that use analog forecasting techniques. In addition they note the additional inclusion of anti-analogs, where the past pattern is opposite to the present. These are used with the evolution observed in the

past reversed in order to make the forecast. Weighted averages of several analog forecasts have also been used, which was also noted by Drosdowsky (1994).

Drosdowsky (1994) discusses analog forecasting as a nonlinear method by adopting a state space approach to the analysis of climatological series. Analogs are then constructed by examining the historical record for states close to the current state. The approach draws heavily on the work of Sugihara & May (1990). This is a valuable idea, bringing a much needed focus on the underlying climate dynamics to the statistical analysis. Contemporary statistical methods based on nonlinear time series and ideas from dynamical systems (Tong, 1990) have much to offer in this light. Such methods provide a framework for examining nonlinear processes directly.

2.2   Opportunities for enhanced use of Statistical Methodology

A number of themes have emerged from the review. First, there is a heavy reliance in the literature on linear statistical methods. This has the potential to hide potentially useful predictive relationships that happen to be nonlinear in nature. We examine an alternative to the correlation coefficient in section 2.2.1, and develop a modelling framework in section 4. Climate data also tend to be spatially and temporally related, which is recognised by techniques such as PCA and CCA to some degree. More importantly, the assumption of stationarity required by conventional statistical methods is a very strong one. There are a number of more contemporary approaches that may be beneficial, and we discuss three of these below. In each case we begin with a motivating paragraph to establish the usefulness of each technique, and then provide some technical detail for the interested reader.

2.2.1   Nonparametric Screening Tools

In current practice the only tool in regular use to screen for significant relationships between climatological variables is the correlation coefficient. As noted in section 2.1.1, this tool is only suitable for detecting linear relationships. This screening process may be conceptualised as follows. We suppose that there is input data denoted X (e.g. an SST field) and an output ϕ (e.g. rainfall), and we seek those cases having a high correlation $\rho(\phi, X)$. We propose to generalise this to encompass more general, nonlinear relationships by seeking large values of $\rho[\phi, f(X)]$. In this formulation $f(\cdot)$ is an unknown function. Of particular interest are cases

where $\rho[\phi, f(X)]$ is large compared to the absolute value of $\rho(\phi, X)$, since this is indicative of nonlinear relationships characterised by $f(\cdot)$.

We show here how the loess technique of Cleveland & Devlin (1988) can be used to estimate the function $f(\cdot)$. Loess is a local regression procedure that robustly smoothes scatterplot data by fitting regression relationships locally using a window around each observation. Thus no global form of $f(\cdot)$ is assumed, and it is reconstructed using a set of local regressions. We use winter rainfall at Manjimup (station 9619) to illustrate the potential of this technique in searching for useful relationships; this is <u>not</u> a definitive analysis by any means.

We seek to relate rainfall to SSTs from the SO1 data set (231 SST locations) extracted from GISST2, to be used in section 3. We defer further description of these data to then. These data have correlations in the range –0.1752 to 0.1245. We explore here two cases in particular: what information might be hidden by very low correlations, and what extra information is to be gained from examining the relatively high correlations. The current data set admittedly provides a limited basis to address the second case in particular, where a linear component is present but may only be part of the true relationship. The first case is especially interesting since it corresponds to nonlinear relationships that would not be detected at all using a correlation coefficient.

SST location 137 has an observed correlation of –0.060, but after applying the loess smoother this becomes 0.356, and the results are shown in Figure 1 below. The options chosen were the default values. Loess can be made more sensitive to local features of the data by reducing the window size (known as the "span") in each local regression, and the results of reducing the span from 0.75 to 0.3 are shown in Figure 2 below. The correlation now is 0.512, and loess is clearly identifying some nonlinear features in this data set.

We now examine the SST location 21, which has the maximum correlation of 0.1245. After applying the default loess smoother this correlation is increased to 0.241, and the result is shown in Figure 3 below. There appears to be a threshold in the rainfall response above an SST of about 21°, which is associated with higher average winter rainfall. There are admittedly relatively few data points above this threshold. This nonlinear behaviour would not be reliably detected by a simple correlation approach.
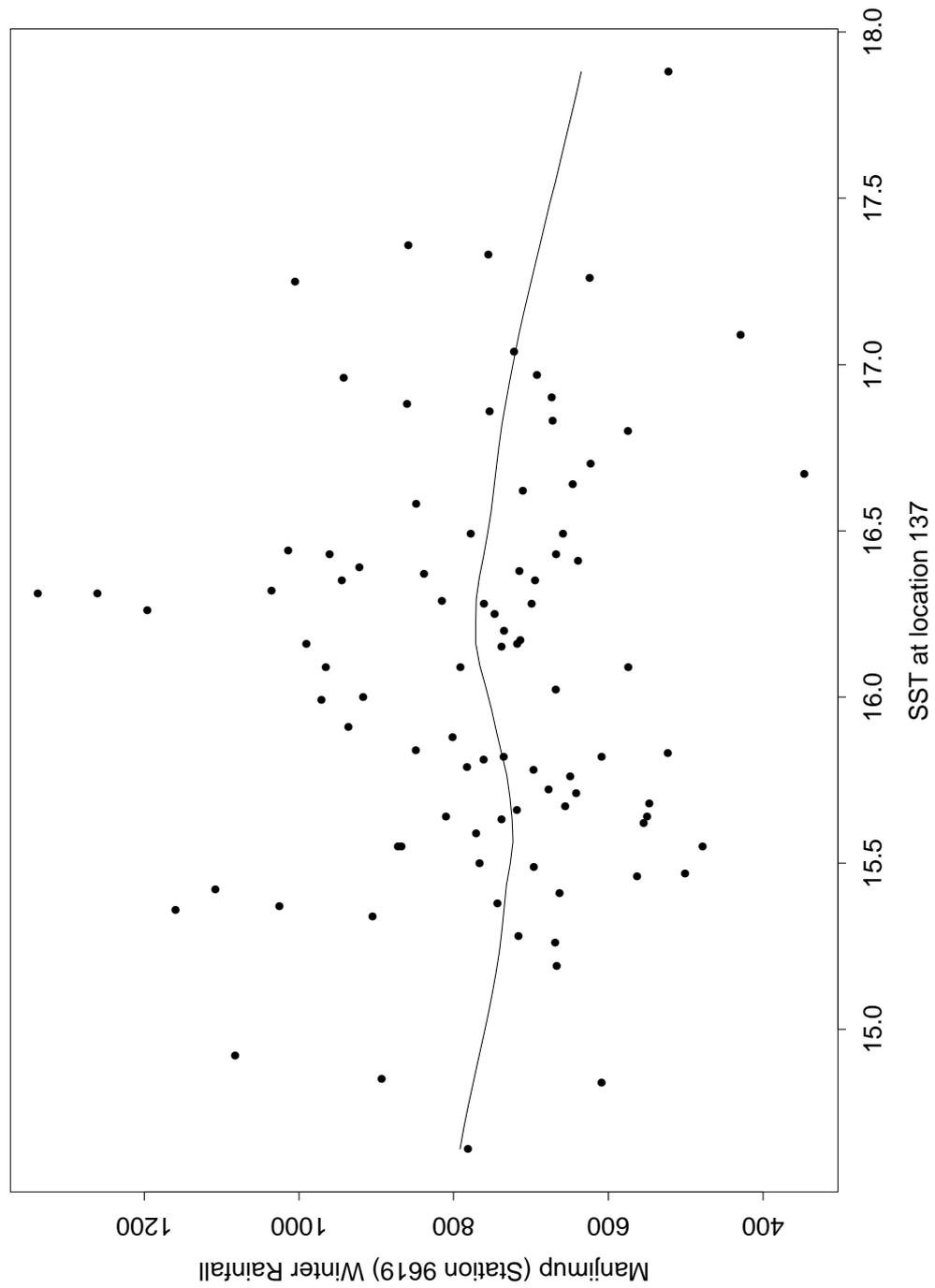
Figure 1   **Loess smoothing applied to the relationship between winter rainfall at Manjimup (station 9619) at SST at location 137 of the SO1 SST data set**
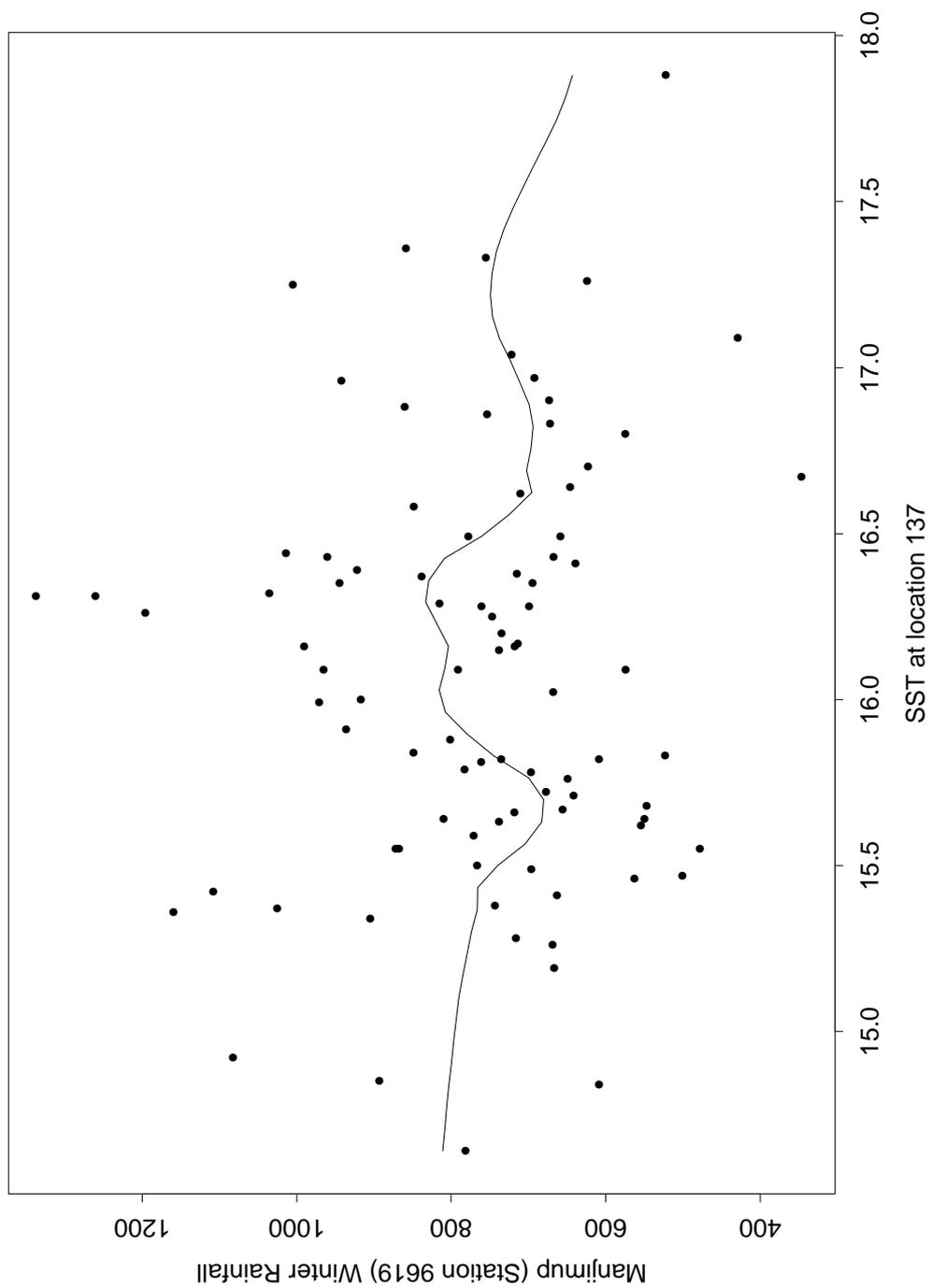
Figure 2   The effect of reducing the span in loess from 0.75 to 0.30
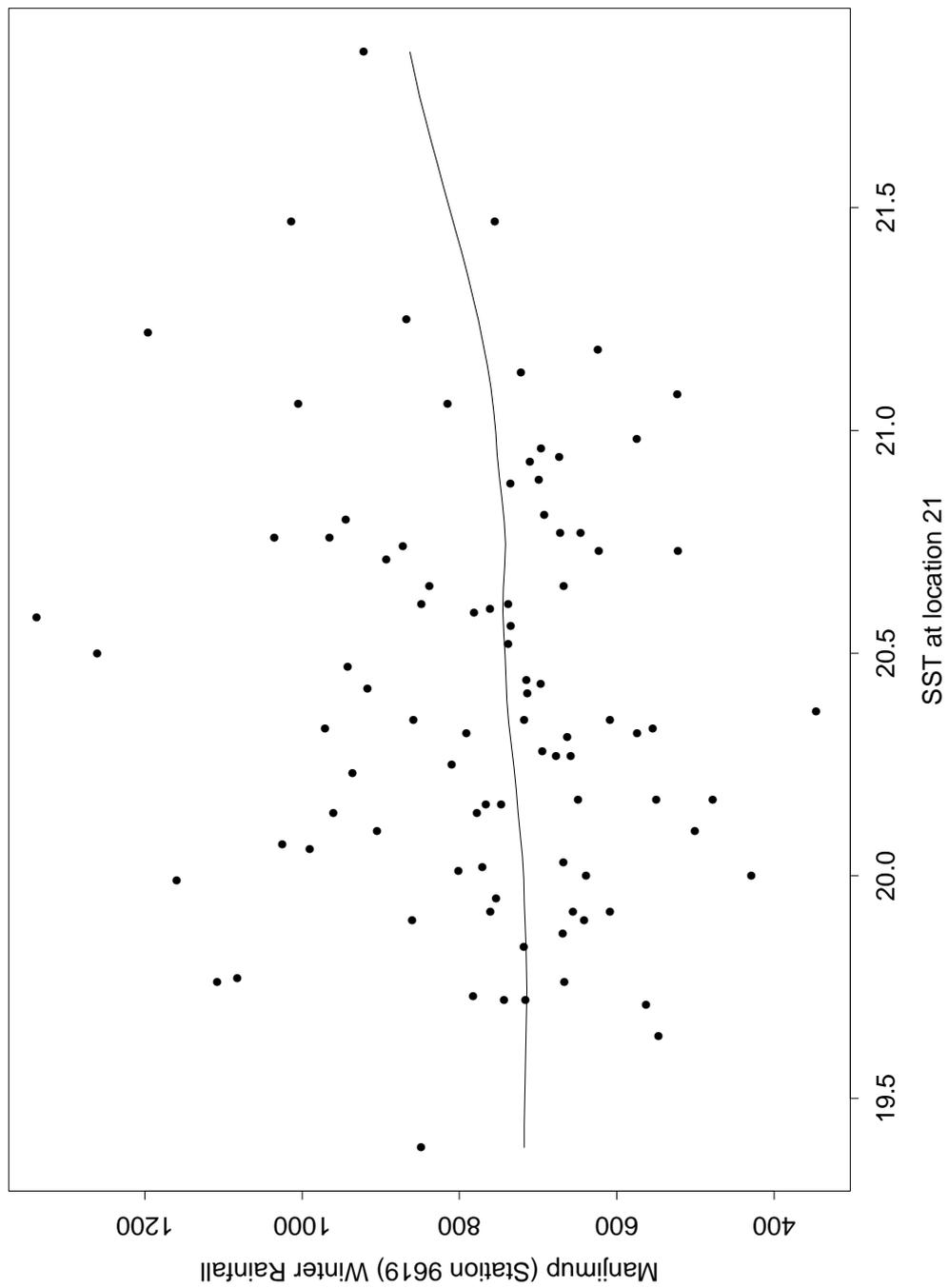
Figure 3   Loess smoothing for SST site 21 which has the highest correlation between SST and rainfall at Manjimup (station 9619)

This analysis has revealed some clear nonlinear features of the data, which would not be revealed by a simple linear correlation/regression analysis. Scatterplot smoothing is therefore a potentially very valuable tool. Subsequent research should identify suitable data sets to explore using this approach. Note that classical regression modelling can be generalised to use relationships defined by scatterplot smoothers such as loess, and such models are known as generalised additive models (GAMs) (Hastie & Tibshirani, 1986). Such an approach also allows a number of rainfall predictors to be included in a nonparametric regression model. This is a powerful contemporary statistical method that is potentially of great benefit in climate forecasting.

2.2.2 Functional Data Analysis

As defined by Meiring & Nychka (1998b), this is the study of curves represented by irregular samples from the curves, and the samples may be contaminated by measurement error. There are two principle differences with multivariate analysis. First, the underlying continuity of the domain in functional data analysis (FDA) and, secondly, the smoothness of the curves.

Meiring & Nychka (1998b) discuss an example where they apply FDA to stratospheric ozone modelling at particular locations resolved in a vertical profile. They consider a class of varying coefficient functional data models. The coefficients of a basis function expansion depend on covariates, such as the quasi-biennial atmospheric oscillation and season. This provides a means to model non-stationary phenomena by parameterising and fitting important physical processes. These models are found to be sensitive to the complex interactive effects of the covariates on the shape of the vertical ozone profile. Further details of this work are described in Meiring & Nychka (1998a).

FDA offers great potential where data naturally arise as curves, as illustrated by the above example. A number of papers have appeared in the statistical literature in recent times, such as Pezzulli & Silverman (1993), Ramsay (1991) and Silverman (1996) and references therein. A detailed introduction to FDA is provided by Ramsay & Silverman (1997). The fundamental objective of FDA is to decompose the underlying theoretical functional response into orthogonal components, or basis functions. These can be parameterised for the particular application at hand to provide physical insight to the processes involved. For example, in the vertical ozone profile application of Meiring & Nychka (1998b) described above, vertical ozone $Z(a,t)$ at altitude a and time t is decomposed as

$$Z(a,t) = \sum_{j=1}^{M} c_j(t, X(t))B_j(a) + \varepsilon(a,t) \qquad (1)$$

where

$$c_j(t, \mathbf{X}(t)) = \mu_j + \varphi_j(t)\alpha(t) + f_j(\mathbf{X}(t)) + \eta_j(t) \qquad (2)$$

Each basis function $B_j(a)$ is a continuous function of altitude; $\varepsilon(a,t)$ is a residual space-time process; μ_j is the mean of the jth coefficient; $\varphi_j(t)$ is a periodic function of time of year, used to model the quasi-biennial oscillation (QBO) in the equatorial wind direction that follows an approximate 28 month cycle and propagates down through the stratosphere; $\alpha(t)$ is a function of time t and $f_j(\mathbf{X}(t))$ is a non-parametric function of the covariates $\mathbf{X}(t)$; and $\eta_j(t)$ is a residual process on the coefficient scale.

On substituting (2) into (1) we see that "this model allows for an overall vertical mean, a periodically and vertically varying trend which is linear if $\alpha(t) = t$, complex interactive and non-linear effects of covariates $\mathbf{X}(t)$ on the vertical profile shape, and an overall space-time residual process."

2.2.3 Projection Pursuit

Projection pursuit in its exploratory mode may be thought of as a generalisation of methods such as principal component analysis (PCA) and canonical correlation analysis (CCA). These techniques seek to optimise criteria based on variance and correlation respectively, which are entirely statistical in nature. As discussed in the review, this often leads to poor interpretability of the results. This has led to approaches based on rotation of principal components to enhance interpretability (Richman, 1986).

We propose a quite different approach - make the optimisation criterion more climatologically 'interesting'. Given that we can define such a criterion, the resulting principal components or projections are, by definition, interesting. This technique is known as projection pursuit, and we describe it more detail below.

PCA seeks to maximise the variance of linear combinations of the form $Y = \mathbf{a}^T \mathbf{X}$ for a unit vector $\mathbf{a}$ and sample data $\mathbf{X}$. Thus $Var_{\mathbf{X}}(\mathbf{a}) = \mathbf{a}^T S \mathbf{a}$, where S denotes the sample variance matrix of the data $\mathbf{X}$. This leads to the optimisation problem

$$\arg\max Var_X(\mathbf{a}) \text{ subject to } \mathbf{a}^T \mathbf{a} = 1,$$

where "arg max" means that we seek the value of the argument $\mathbf{a}$ at the maximum. In projection pursuit we generalise the optimisation criterion to

$$\arg\max I_X(\mathbf{a}) \text{ subject to } \mathbf{a}^T \mathbf{a} = 1,$$

where the function $I_X(\cdot)$ is used to describe properties of the data that are *interesting*, rather than simply the variance. Clearly this can be generalised to compare more than one field, which is commonly achieved using an SVD of the covariance matrix of the respective fields (Cherry, 1996). In this more general setting there typically is no analytical solution and numerical techniques must be applied.

There do not appear to be any applications of projection pursuit in the climatological literature, but for an interesting illustration of its practical application see Walden (1994). In this paper a minimum entropy metric is applied to find distributions that are not uni-modal. The 'least interesting' distribution using this metric is a normal distribution, and we seek distributions that are far removed in entropy space. In the absence of a physically-based interestingness criterion an entropy-based criterion could be developed. In this case we would need to define 'least interesting' to develop a working approach. Nason (1995) describes the use and implementation of projection pursuit into three dimensions. For a more theoretical and philosophical treatment of projection pursuit see Jones & Sibson (1987).

The central idea of searching for interesting projections has found other applications in the statistical literature. Principal among these are *projection pursuit regression* and *projection pursuit density estimation*. Such methods offer a useful generalisation of conventional linear methods to allow more realistic modelling, and are related in spirit to the generalised additive models described earlier. For a description of projection pursuit regression and density estimation, see Friedman & Stuetzle (1981) and Friedman *et al*. (1984) respectively.

2.2.4   Space-Time Covariance Modelling

This is a very important theme in contemporary statistics, rather than a single method. The need for genuine spatio-temporal models in climatology has been recognised for some time. See, for example, Zhang *et al*. (1998), Glaseby (1998), Drosdowsky (1993a, 1993b) and references therein. The typical approach has been to consider spatial and temporal behaviour separately.

This area is complicated by the intricate mathematical modelling so often involved, as exemplified by Jones & Zhang (1996) for example. Indeed, it is difficult to make progress without simplifying assumptions such as separable covariance structures. That is, if we observe a random variable $Z(\mathbf{x},t)$, where typically $\mathbf{x} \in \Re^2$ denotes spatial location and $t \in \Re^+$ denotes time, then we assume that

$$Cov\left[Z(\mathbf{x},t),Z\left(\mathbf{x}^*,t^*\right)\right] = C_1\left(x,x^*\right)C_2\left(t,t^*\right).$$

Thus the process covariance function may be found by modelling space and time separately, which implies that they are independent. In most practical cases this is an unrealistic assumption. Examples of this approach are provided by Stein (1986) and Jones & Zhang (1996).

An alternative approach is to abandon an explicitly mathematical modelling approach for a more statistical one. In doing so we lose touch to some degree with the physics of the processes under consideration, but we gain the potential for increased data-driven physical insight. There has been a growing trend in climatology and meteorology to adopt genuinely spatio-temporal data analysis tools, as seen in the work of, for example, Burkhardt & James (1998), Glaseby (1998), Hutchinson (1995), Weare & Nasstrom (1982) and Barnett (1983). Much of the work in meteorology and climatology involves extensions to the EOF technique, which we have already discussed. In most cases this analysis assumes temporal stationarity and estimates a spatial covariance structure. In recent times the approach has been extended to include temporal structure as well – see Weare & Nasstrom (1982) for an introduction. To be readily interpretable, however, the process analysed must be stationary in space and time, which is rarely the case.

An alternative approach is to develop statistical models that can be fitted using available data without the assumption of stationarity. Nott & Dunsmuir (1998) provide an interesting introduction to this subject, in the context of modelling wind fields. They develop a model for the observed process of the form

$$Z(\mathbf{x},t) = \mu(\mathbf{x}) + \eta(\mathbf{x},t) + \varepsilon(\mathbf{x},t),$$

where $\mu(\cdot)$ denotes the spatial trend, $\eta(\cdot,\cdot)$ a zero-mean spatio-temporal process and $\varepsilon(\cdot,\cdot)$ a zero-mean measurement error. Nott & Dunsmuir (1998) describe approaches based on a spatial deformation technique (Sampson, 1986; Sampson & Guttorp, 1992) and a direct kernel estimation technique due to Oehlert (1993). They note a number of severe computational and theoretical drawbacks with the spatial deformation technique. Their own method is based on assumptions of local spatial stationarity, but temporal stationarity is required. They report inferior performance of the kernel technique to their own approach.

We are particularly interested in ideas of spatio-temporal modelling within the framework of stochastic differential equations, which may be seen as a means to bring the physical and statistical modelling together. For support of this view, see Dawson's contribution to the discussion of Eynon & Switzer (1983), where a stochastic differential equation can be written down. In climate physics problems it may be possible to do the same. The challenge then would be to derive important statistical-physical properties from such an equation. For an introduction to stochastic (partial) differential equations, see Ikeda & Watanabe (1989).

3. EXPLORATORY ANALYSIS OF RAINFALL AND SST DATA

In this analysis we have necessarily limited the results shown to demonstrate the points made in the discussion of the complete analysis undertaken. As for the IOCI98 workshop we have used results from Manjimup (station 9619) to display the main results found, augmented by results from other stations where necessary.

3.1 Description of Study Area and Data

We defined SWA as the region extending from about 29° to 35° south and 115° to 120° east, which is shown in Figure 4. The station numbers are identified in **Table 1**. The region experiences a 'mediterranean' climate with abundant winter rains that are nearly double that of any similarly exposed locality in any other continent, and intense summer drought. Eighty percent of annual precipitation falls in the period from May to October, and the majority of winter rains come from low pressure frontal systems (Wright, 1974; Gentilli, 1972).
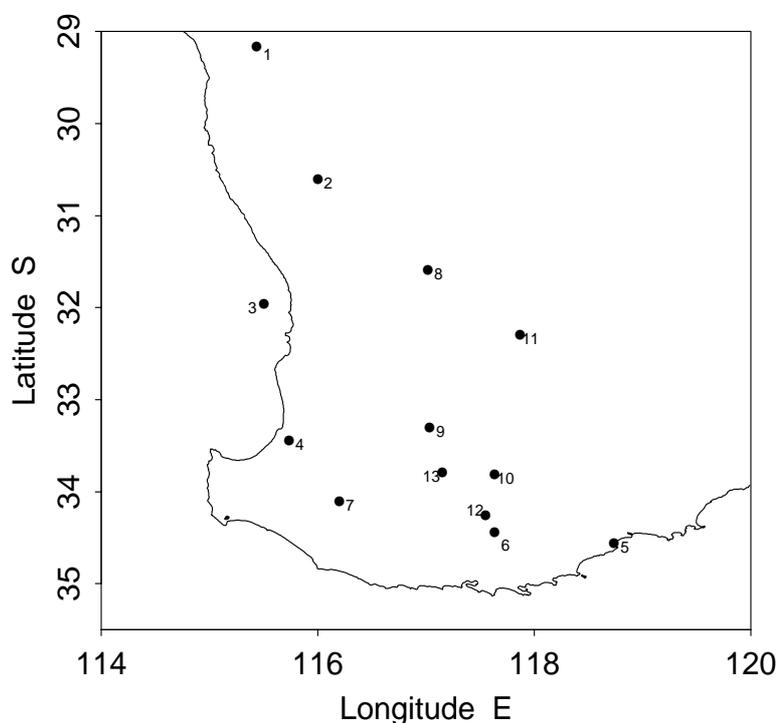
Figure 4   South-west Western Australia with rainfall stations used in this study

Daily precipitation data were obtained from the Commonwealth Bureau of Meteorology's high quality rainfall data set (Lavery *et al*., 1997). The SST data used in this study were drawn from the Global Sea-Ice and Sea Surface Temperature version 2 data set (GISST2), supplied by the UK Meteorological Office. The analysis shown here is designed for illustration only, and is not intended to be definitive. Subsequent research will use more recent versions of GISST, and will make greater use of more reliable satellite-based observation of SST.

3.2   Analysis of Monthly Rainfall Time Series

The monthly rainfall record for Manjimup (station 9619) is shown in Figure 5 below. An immediately obvious property of the series is its inherently 'spiky' nature, with seemingly an overall decline in the magnitude of these spikes since about 1960. The results of a complex demodulation are shown in Figure 6. This analysis does not assume that the time series is stationary, but instead assumes that it is fundamentally a sine wave having slowly varying amplitude and phase about some fundamental frequency. The amplitude and frequency can be estimated from the data and used to reconstruct a smoothed series. We see that the perceived recent decline in rainfall is not isolated in the historical record, which demonstrates considerable variations in amplitude.
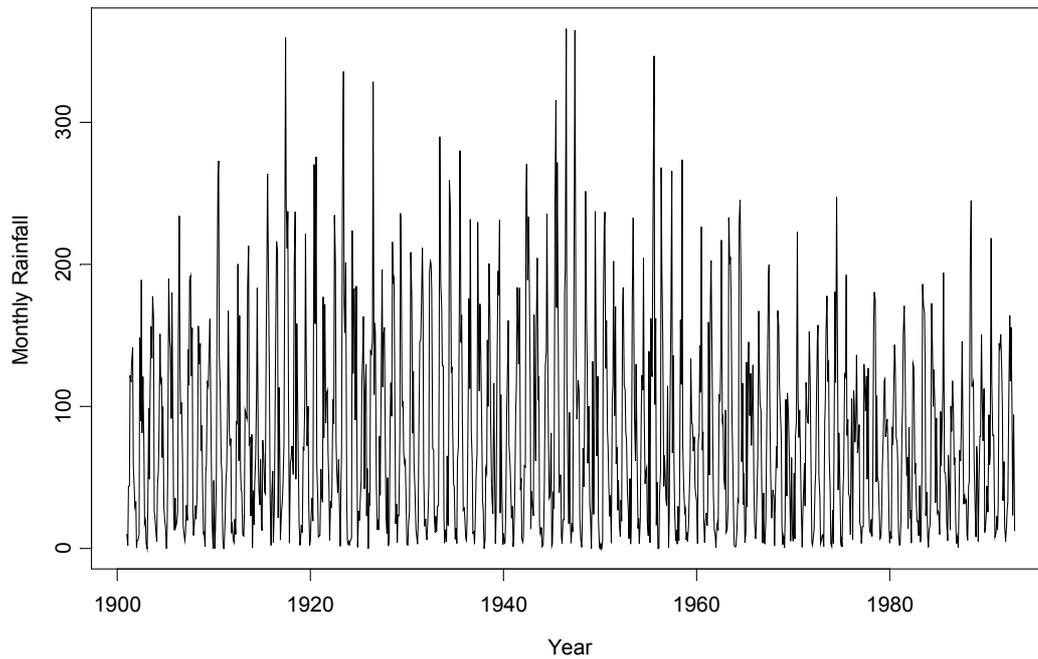
Figure 5   Monthly rainfall record for Manjimup (station 9619)

The complex demodulation can be used to construct a filter based on a given fundamental frequency. In this case we selected the spectral peak in the frequency range of the quasi-biennial peak evident in the multi-taper spectrum to be described in the next section. The resulting smoothed time series is shown in Figure 7. We can see a number of abrupt changes in mean level. Taken together, these results suggest that nonlinear dynamics are present in the rainfall time series.
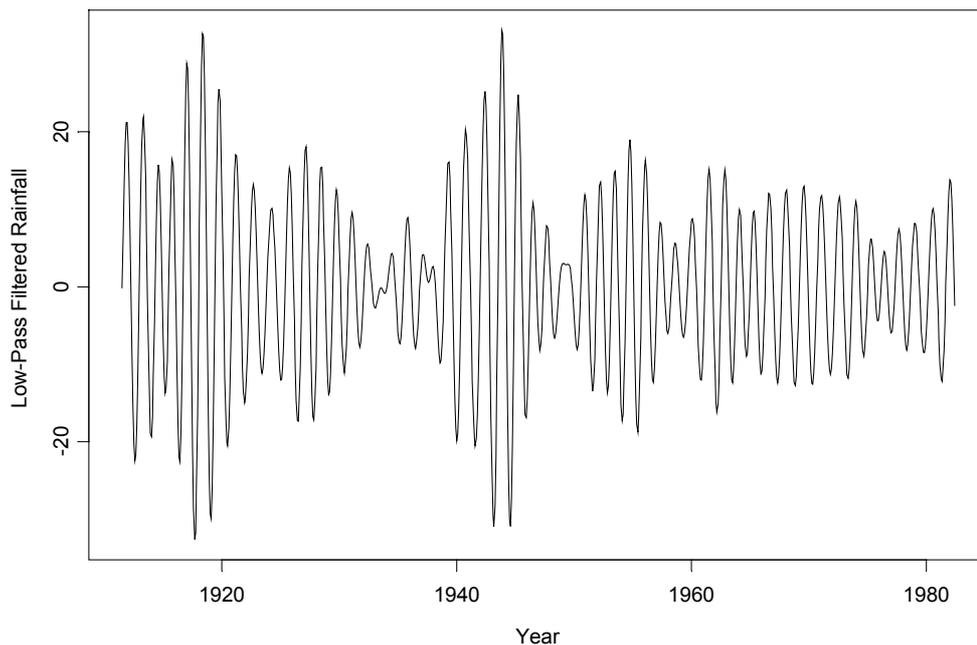
Figure 6   Complex demodulation of the Manjimup monthly rainfall series
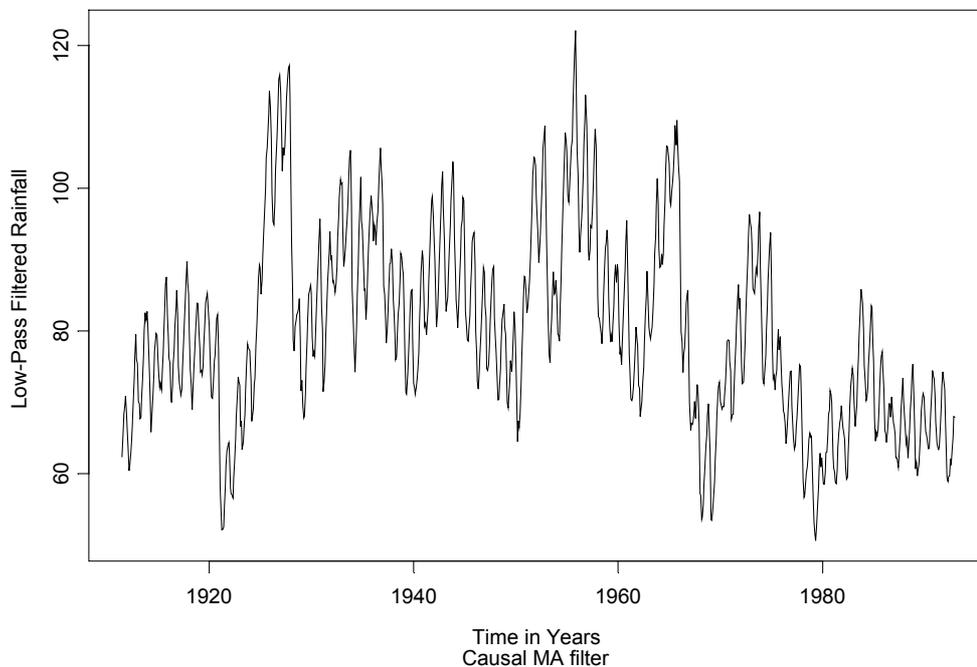
**Figure 7   Smoothed Manjimup monthly rainfall series using filter derived
from the complex demodulation**

3.3   Multi-Taper Spectral Analysis

In conventional spectral analysis a common problem is leakage between frequencies, particularly from frequencies having high power. That is, a few dominant frequencies tend to drown out signals from nearby frequencies. A solution to this is to smooth the data before calculating the spectrum, a process known as tapering. This process can however result in some features of the original data being lost. An alternative is to use a number of different tapers and estimate the spectrum as a weighted average. In this way a spectrum estimate is found that more truly reflects the original data whilst minimising leakage. This is especially useful for climate data where we will typically need to resolve a number of peaks.

The multi-taper spectral analysis of Manjimup (station 9619) is shown in Figure 8 below for periods greater than 1 year. We see that there is significant power at a wide range of frequencies, in keeping with a nonlinear dynamical system. The peak at around 0.4 cycles per year is common to most of the stations analysed, corresponding to a quasi-biennial oscillation. The peak in the region of 0.6 to 0.7 cycles per year is also common to most stations. The overall results are summarised in Figure 9 below, which shows the geographical spread of the spectral peak, expressed as a period.

No obvious pattern emerges from the analysis of the spectral peaks, although there does some appear to be a tendency for longer periods towards the south coast. Further work will seek to discover if there are links to local orography.

3.4   Wavelet Analysis

There is clear evidence from the analysis thus far that nonlinear dynamics are present in the rainfall time series. In practice this means that the assumption of stationarity is violated. Rather than viewing this as a problem to be solved by data pre-processing, for example, it should be viewed as a fundamental property of climate data. The statistical tools should be adapted to be appropriate to the data if we wish to gain any physical insights.
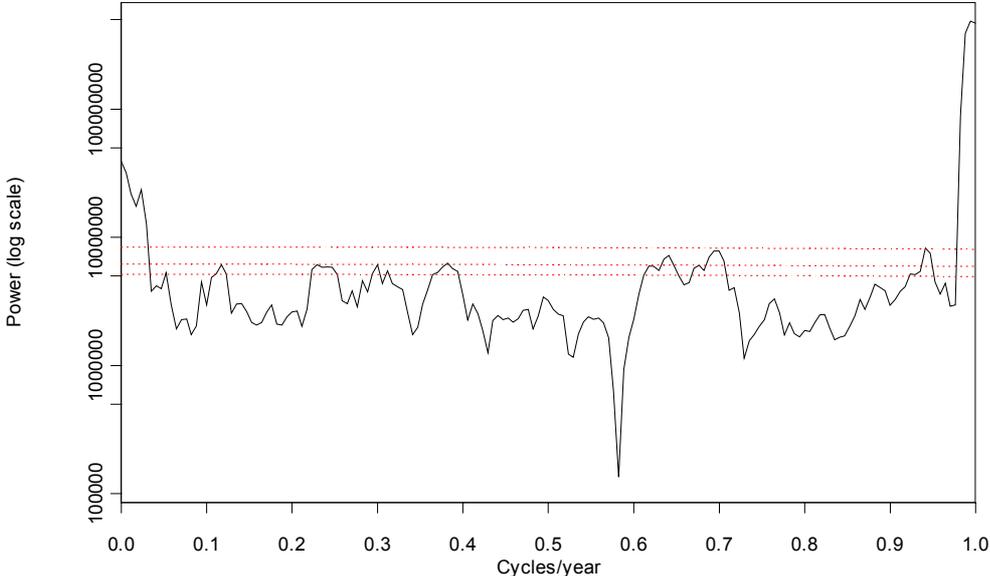
Figure 8 **Multi-taper spectral analysis of rainfall at Manjimup (station 9619), shown with 90%, 95% and 99% confidence limits**
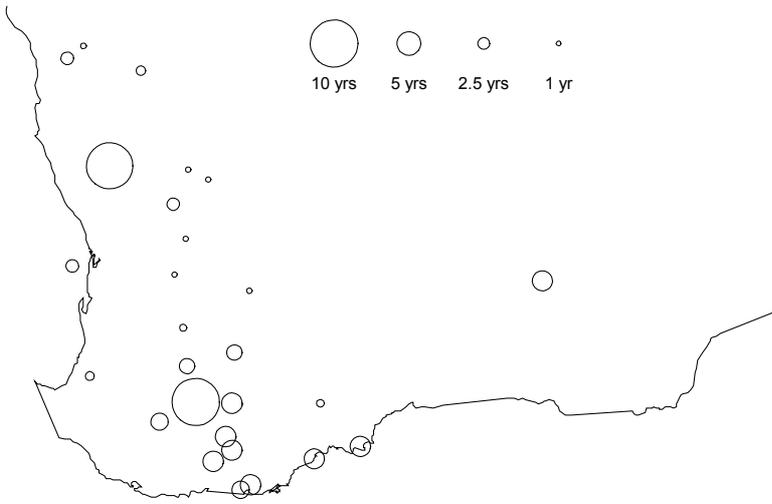
Figure 9 **Periods corresponding to the spectral maxima, including a number of extra stations outside the main study area**

Wavelets are a tool of great potential in the analysis of climatological data. Their fundamental properties include time (and space, in general) localisation, thus avoiding the leakage problems of spectral analysis, and a multi-resolution property that allows us to zoom-in and zoom-out from a time series. For more details of Wavelet techniques, Ogden (1997) is an excellent introduction.

We apply wavelets here to analyse the Manjimup monthly rainfall time series. The wavelet equivalent of the Fourier transform is the discrete wavelet transform (DWT), which is shown in Figure 10 below for these data. A symmlet (symmetric wavelet) originally constructed by Daubechies ("s8") is the default choice in S-PLUS software, and was used in this case.
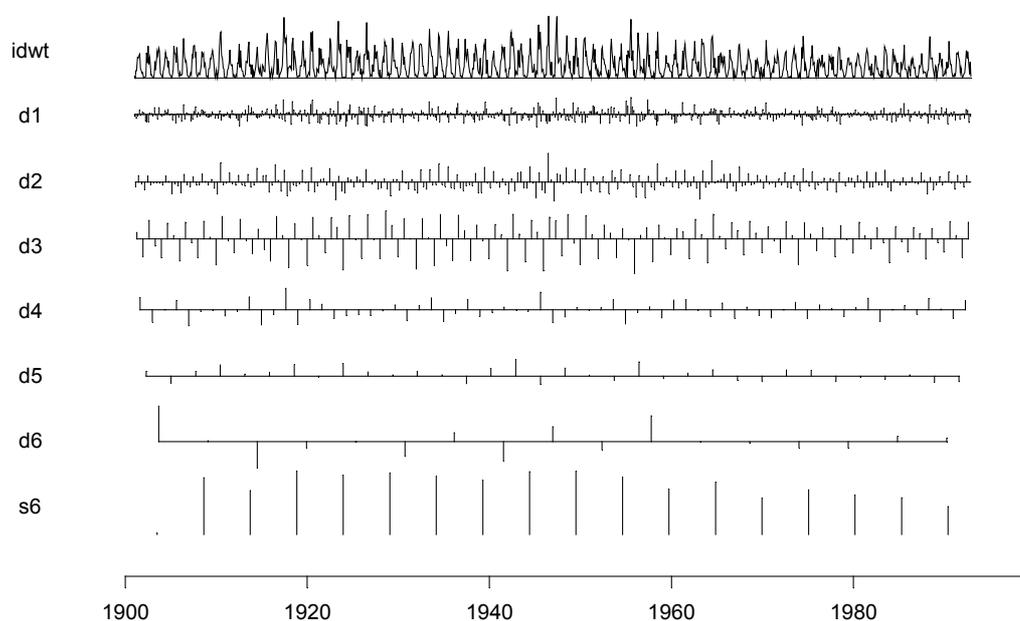
Figure 10   Discrete wavelet transform of the Manjimup (station 9619) monthly rainfall series

The representation of the DWT shown provides the magnitudes of the wavelet coefficients. Since they are localised in time this analysis also provides an indication of potential breakpoints. The first set of coefficients is labelled "s6", and corresponds to an overall rainfall trend. The subsequent d6, d5, …, d1 coefficients are successively more detailed components of the transform. Evidence of a break point is especially strong where these coefficients tend to group, such as before 1920, after 1940 and prior to 1960 for example.

The DWT can be used to reconstruct smoothed versions of the original data, and this process is depicted in Figure 11 below. The so-called multi-resolution analysis (MRA) is shown on the right hand side, and represents successively more detailed approximations to the original time series. The multi-resolution decomposition is shown on the left. The smoothed estimate *S*6 is found using the wavelet coefficients s6 of the DWT, and represents the crudest approximation. The next approximation is found by adding the next level of detail, *D*5, so we calculate *S*5 = *S*6 + *D*5 and so forth until Data = *S*1 = *D*1.
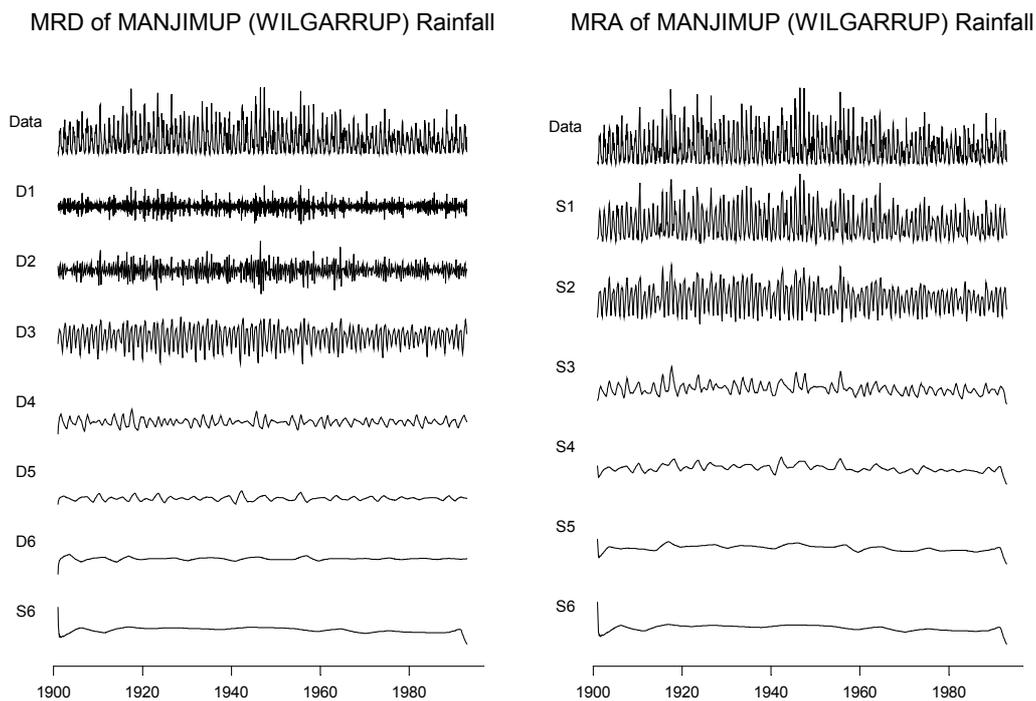
Figure 11   Multiresolution analysis of the Manjimup (station 9619) monthly rainfall time series

There is clear evidence of non-stationary behaviour in the crude smoothes S6 and S5. D3 brings in a strong high frequency component, whilst D2 and D1 seem to bring in some of the very large, isolated rainfall periods.

3.5   Penalised Discriminant Analysis

Penalised Discriminant Analysis (PDA) is an extension of classical linear discriminant analysis (LDA), developed by Hastie *et al.* (1995). LDA is a highly useful technique when carefully applied, but suffers from a number of deficiencies. When presented with large numbers of highly correlated predictors, to paraphrase Hastie *et al.* (*op. cit.*), LDA is too

flexible and tends to over-fit the data. An application of interest to us is relating SST fields to observed rainfall. In cases where the class boundaries in predictor space are complex and nonlinear, LDA tends to be inflexible and under-fits the data. For further discussion of approaches in this latter case, see Hastie *et al*. (1994).

The idea behind PDA is that these deficiencies can be overcome by appropriate 'regularisation' of the within-groups covariance matrix, denoted Σ_W. There are two distinct motivations for this:

1.   When the number of predictor variables is high relative to the number of observations, we cannot reliably estimate Σ_W;

2.   Even when the sample size is sufficient to reliably estimate Σ_W, coefficients of spatially smooth variables tend to be spatially rough. Interpretation would be greatly aided by a smooth version, given that the fit is not compromised.

Hastie *et al*. (1995) proposed that Σ_W be replaced by $\Sigma_W + \lambda\Omega$, where Ω is a "roughness"-type penalty matrix. The LDA analysis then proceeds as usual. Note that Hastie *et al*. (*op. cit.*) go on to show equivalence of PDA with penalised versions of canonical correlation analysis and optimal scoring. A key point here is that optimal scoring provides a link between nonparametric regression and discriminant analysis, providing a rich class of modelling tools in classification. For an alternative perspective on regularisation in discriminant analysis, this time using mixtures, see Hastie & Tibshirani (1996). Ultimately it might be most efficient to build penalised regression models, avoiding the artificial reduction of data by division into rainfall categories.

We show here how PDA may be used to relate SST fields to observed rainfall. We classify Winter Rainfall for a given year as 'Low', 'Normal' or 'High' if it lies in the first, second or third tercile respectively. An SST field is typically a spatially smooth two-dimensional field, and we choose a discrete second derivative-based penalty to reflect this. A suitable choice is the Neuman discretisation described by O'Sullivan (1991).

Given this choice, we obtain the results shown in Figure 12 below for Manjimup rainfall using SST data over a region denoted SO1 for ease of reference.
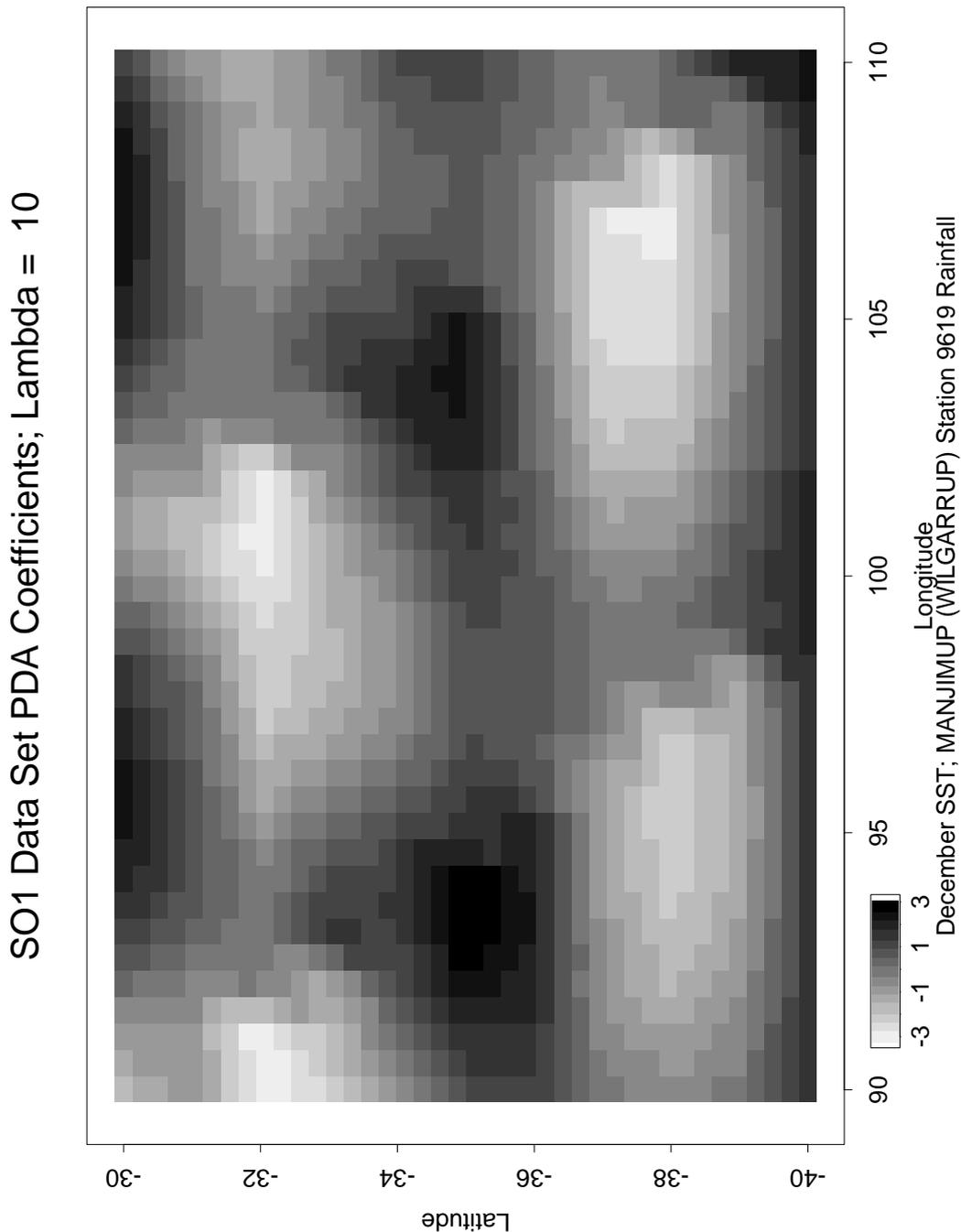
Figure 12   Penalised discriminant analysis relating the SO1 SST field to rainfall at Manjimup to 1971

The mean SST field is shown in Figure 13. Note that data up to 1971 have been used in this analysis, with the remainder intended for subsequent validation purposes. December SSTs seems to possess the strongest signal in terms of the magnitude of the coefficients. The value of λ was chosen by eye to obtain a clear image; it is possible to choose this parameter objectively using cross-validation. We reiterate that this analysis is designed to be illustrative

only, as there are many deficiencies in these SST data, both over the region and time span selected.
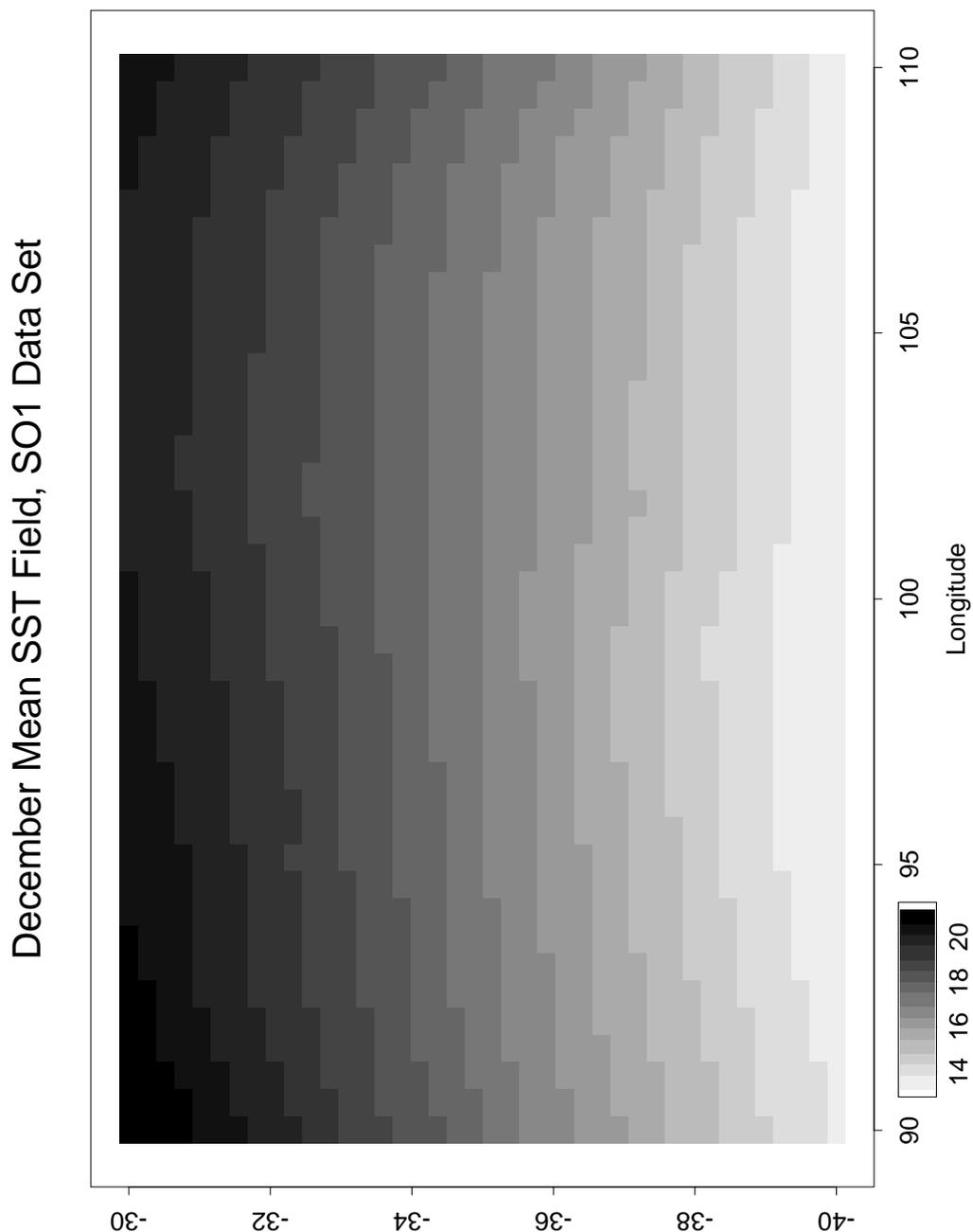
Figure 13   Mean SST field for the SO1 data set, using data to 1971

The spatial pattern of the coefficients seen in Figure 12 is common to most of the other stations, and suggests some relationship between SSTs in the SO1 region and rainfall at Manjimup (station 9619). The PDA has identified contrasts between relatively isolated ocean regions, and the mean SST field alone is clearly not driving this relationship. The spatial

pattern of PDA coefficients seen in Figure 12 is not unique however, and an illustration of this is provided in Figure 14 below for Mt Barker (station 9591). There seems to be a diagonal structure in the contrasts between regions for this station, reflecting a winter rainfall record that is different to Manjimup (station 9619). These features will require further investigation and refinement of the PDA technique. Some preliminary discussion of these results with CSIRO Marine Research (CMR) scientists has already taken place.

4.   A FRAMEWORK FOR NONLINEAR STATISTICAL ANALYSIS

4.1   Overview

The literature review and preliminary data analyses lead us to propose a nonlinear time series model for climate processes such as rainfall. We may express the model having forecast lead time *T* as:

$$X_{t+T} = f_T(\mathbf{X}_t, \mathbf{U}_t; \varepsilon_{t+T}), \qquad\qquad (2)$$

in which $\mathbf{X}_t$ is a vector[1] representing the history of the (rainfall) process $\{X_u : u \leq t\}$ at a particular site; $\mathbf{U}_t$ represents exogenous covariates such as SST or mean sea level pressure (MSLP), and so in general will be a data matrix. These quantities may be derived from some other analysis, such as PDA. The term ε_{t+T} represents a random error. In the preliminary modelling work to be presented below we assume that this term enters purely as a random additive error at time $t + T$. Our research will explore more general error structures as appropriate.

In conventional statistical approaches, the next step would be to assume some form, typically linear, for the mapping f_T. There have been many exciting developments in modern statistical practice during the past two decades, largely due to the advances in computing technology. It is now feasible to estimate f_T using available data, allowing us to recover the underlying dynamics, so that we may pose more insightful questions of the data than linear methods would allow. One possibility is to make use of local regression (Cleveland & Devlin, 1988), which is a technique for fitting general continuous relationships. There is considerable evidence however that a form of discontinuous modelling known as threshold autoregression is more appropriate for modelling nonlinear systems (Tong, 1990). These models can reproduce, for example, the occasional very large observed rainfall periods and many other nonlinear phenomena.

---

[1] We don't intend to develop multivariate models, for rainfall and temperature jointly say, until we have applied and validated univariate nonlinear models.
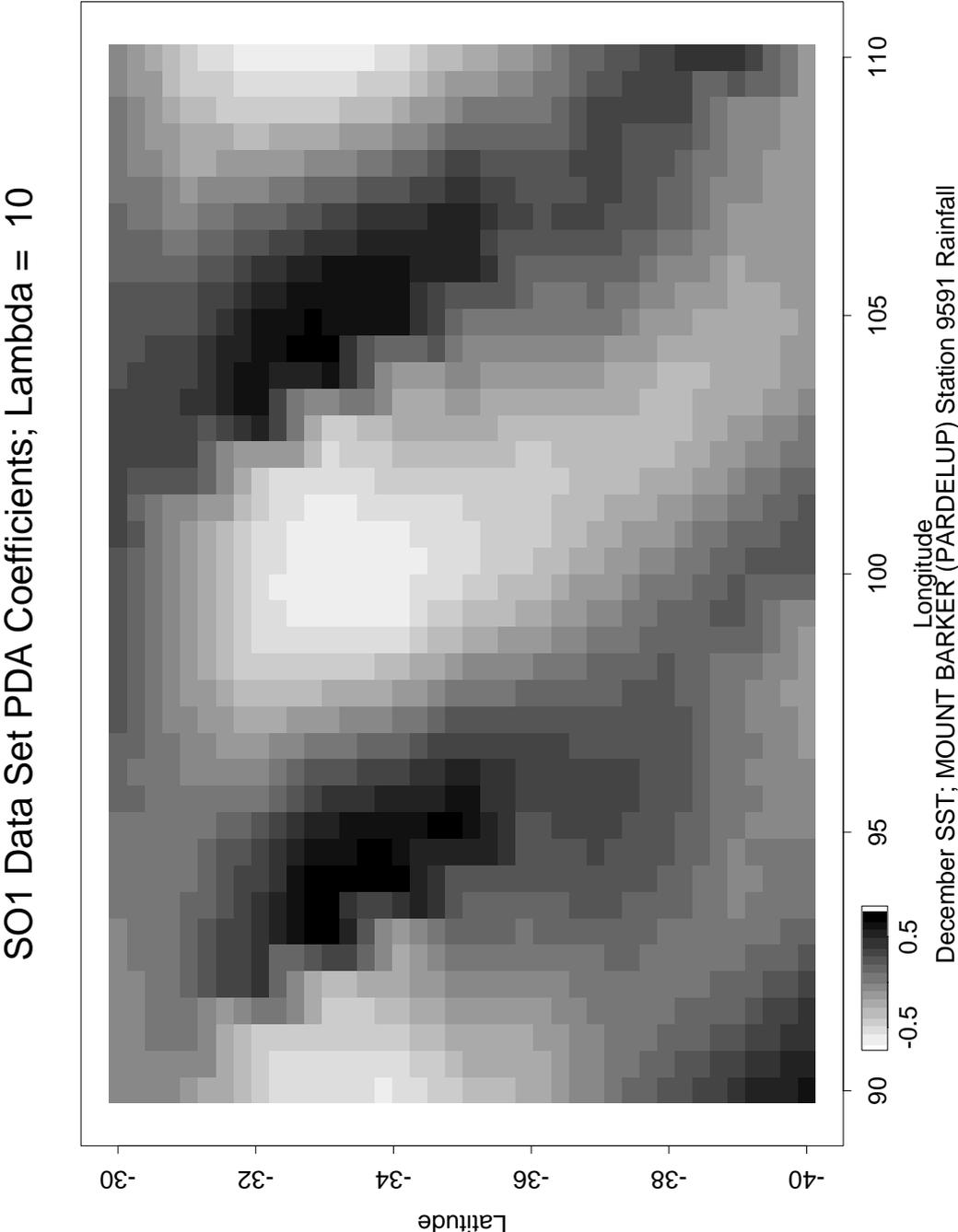
Figure 14   Spatial map of PDA coefficients for rainfall at Mt Barker to 1971 using SST data from region SO1

The most modern advances in this field use spline methods to estimate f_T (Lewis & Stevens, 1991; Lewis & Ray, 1997; Lall *et al.*, 1996; Denison & Mallick, 1998). The knot points of the spline correspond to the thresholds in threshold autoregression, so such methods may be thought of as a form of generalised threshold modelling. There is also evidence that spline models require fewer parameters (Lewis & Stevens, 1991). The reason for this is that interaction terms are easy to include, providing a more succinct description that could only be approximated by a high order linear autoregressive model. The best approach at this time would appear to be Multivariate Adaptive Regression Splines (MARS- see Friedman, 1991), a highly effective fully automatic technique. An excellent example of how such models can be used to gain physical insight is given by Lewis & Stevens (1991), in the context of modelling the sunspot cycle.

When viewed within a Bayesian framework the approach becomes especially powerful. Modern computational techniques based on reversible jump Markov chain Monte Carlo (MCMC) (Green, 1995) can be used to determine both the appropriate number of thresholds and their location. The methodology can also be used to select suitable rainfall predictors, such as SST predictors derived using PDA for example. This represents a significant advance in nonlinear time series analysis. Examples of this approach to estimating f_T are given by Denison *et al.* (1998) and Denison & Mallick (1998), but the proposed use in predictor selection is novel and will require a statistical research component.

Nonlinear methods are not as easy to apply as classical linear methods, but there are substantial potential benefits of this greater complexity in terms of enhanced predictability and physical insight. There are published approaches for fitting models described by equation (2) (Lin & Pourahmadi, 1998; Hardle *et al.*, 1997; Denison & Mallick, 1998; Lall *et al.*, 1996), so it is relatively straightforward to examine the potential utility of this approach. To be of great forecasting utility though there are significant research questions to address. First amongst these is the selection of predictor variables, both autoregressive and exogenous. We propose that a Bayesian framework be adopted, since this will provide a method for assessing prediction uncertainty, implemented using the reversible jump MCMC methodology. A research question that we defer for the moment is the analysis of multivariate time series, such as rainfall together with temperature. This is potentially of great benefit, but the approach should be proven on univariate time series first.

4.2   Preliminary Rainfall Modelling Results

We have used the TSMARS (Time Series MARS) approach of Lewis & Ray (1997) to conduct a preliminary examination of the MARS approach for modelling monthly rainfall. TSMARS has been implemented in the ITSMARS (Interactive TSMARS) software package written by Dan Ames of Utah State University. The rainfall data were scaled to have mean 0 and standard deviation 1 for computational convenience. No further data pre-processing was done.

The ITSMARS software leaves the last 10% of the data aside from model fitting for validation purposes. Using the remaining data we identified lags up to 12 months as being available for model fitting, and up to second order interactions were allowed. The fitted data for Manjimup (station 9619) are shown plotted against observed values in Figure 15 below. The correlation between observed and fitted values is 0.83, which seems quite satisfactory. Notice though the behaviour of the data towards the left-hand bottom corner of the figure, which corresponds to dry months. The fit to large rainfall events is less than satisfactory. These features will require further attention at a later stage.
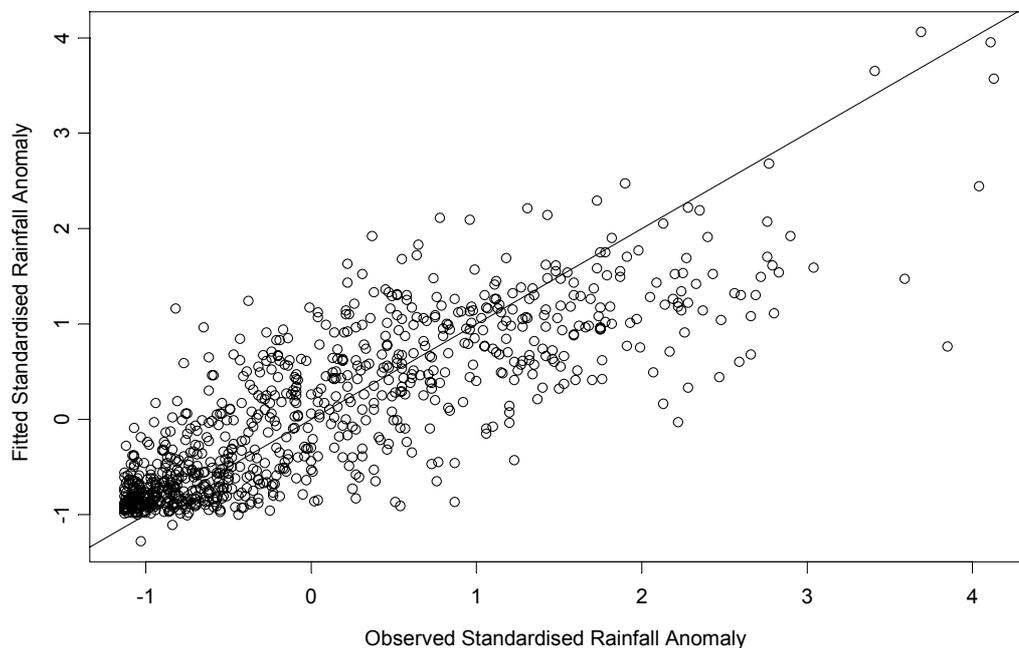
Figure 15   Fitted against observed rainfall anomalies using ITSMARS at Manjimup (station 9619)

The fitted time series is shown in Figure 16 below, which compares well visually with the observed series shown in Figure 5. In particular, the occasional very high rainfall values are present along with variations in amplitude.

The fitted model was used to predict the validation data left out of the model fitting, and the results are shown in Figure 17; the correlation coefficient in this case is 0.74. This is a very encouraging result given the crude nature of this initial analysis, and lends optimism that this approach has something to offer. Note that there is a poor fit once again to the high observed rainfall events. A complete list of correlation coefficients is shown in **Table 1**. Note that the worst case is Albany (station 9520), but due to its location on the south coast this station is expected to be a special case.

These results have shown that the TSMARS approach is able to reproduce successfully the qualitative features of rainfall that we require, which linear methods cannot. The results are also quantitatively encouraging, to the extent that this technique should now be implemented in a statistically rigorous manner and compared with existing approaches. As noted previously, this will require the development of a suitable predictor selection methodology. The potential benefit is nonlinear models having improved prediction capability over linear methods.
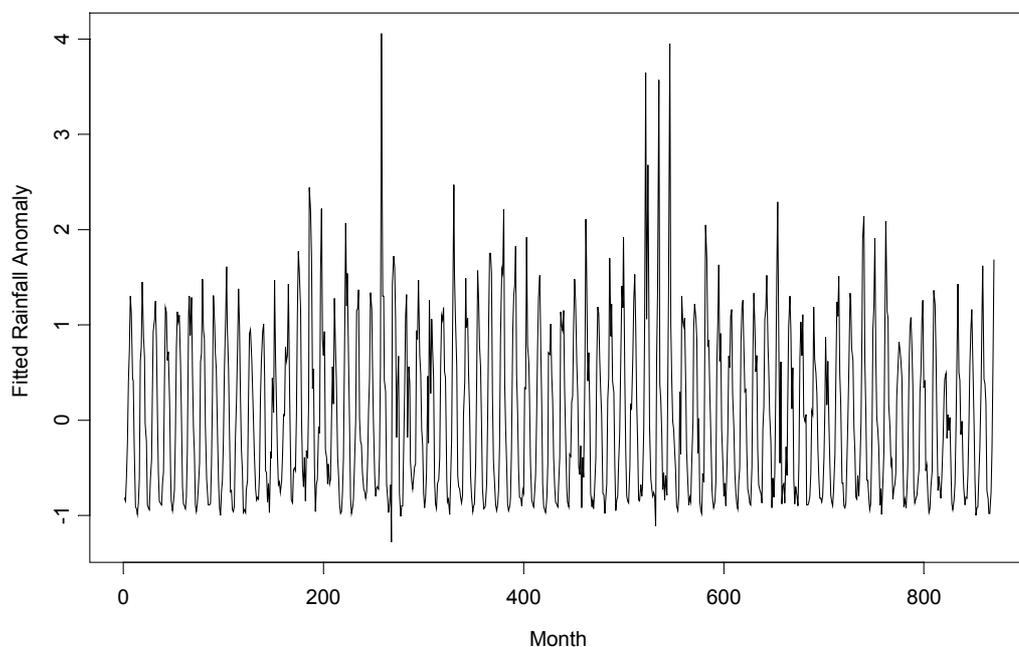
Figure 16   Fitted rainfall anomaly time series for Manjimup (station 9619) monthly rainfall
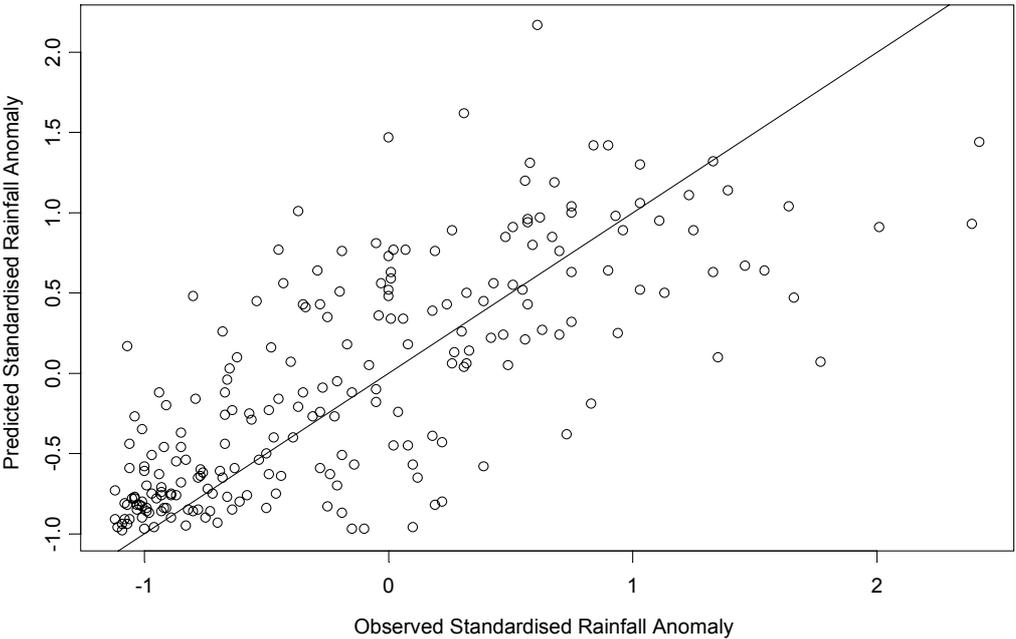
Figure 17   Predicted rainfall anomalies on the validation data for Manjimup (station 9619) monthly rainfall data

Table 1   **Correlation coefficients for initial study using ITSMARS in SWA**

Station Label in Figure 4	Station Name (Number)	Correlation Coefficient	
		Fitted Rainfall	**Predicted Rainfall**
1	Mingenew (8088)	0.75	0.43
2	Moora (8091)	0.75	0.59
3	Rottnest Is. (9038)	0.86	0.64
4	Capel (9503)	0.85	0.77
5	Albany (9520)	0.53	0.20
6	Mt Barker (9561)	0.70	0.42
7	Manjimup (9619)	0.83	0.75
8	Meckering (10091)	0.71	0.31
9	Arthur River (10505)	0.74	0.40
10	Broomehill (10525)	0.68	0.40
11	Corrigin (10536)	0.59	0.25
12	Cranbrook (10537)	0.67	0.46
13	Kojunup (10582)	0.76	0.52

5.   CONCLUSIONS

5.1   Summary of the Investigation

During the first year of the Indian Ocean Climate Initiative (IOCI), CSIRO Mathematical and Information Sciences (CMIS) and CSIRO Land and Water (CLW) have examined the use of statistical methods for climate forecasting in south-west Western Australia (SWA). We have progressed according to the following steps:

1. Survey the climate forecasting literature to assess the current state of knowledge from a statistical perspective;

2. Identify the main issues that need to be addressed using statistical methods, and assess the suitability of methods currently in use;

3. Identify a research program that has the potential to deliver enhanced statistical forecasting tools at inter-seasonal, inter-annual and decadal time scales.

Our progress against each of these themes has been as follows.

1. We have conducted an extensive review of the climate forecasting literature, and it is intended that this will be submitted to a peer-reviewed journal. The objectives of the review were: to gain an appreciation of the physical processes involved, particularly their impact on statistical modelling and analysis; to assess the current use of statistical methods; and identify potentially beneficial opportunities both for statistical research and enhanced application of contemporary statistical methods.

2. Perhaps the most pervasive method in climatology is that of empirical orthogonal function (EOF) analysis. This method assumes stationarity in time explicitly, but also implicitly assumes that the geophysical field of interest is spatially homogeneous. We discuss some more modern approaches to spatial-temporal modelling in this report. The criterion for extracting orthogonal functions is also purely statistical, and the patterns that emerge are usually predictable *a priori*. A great deal of work has been done on the use of oblique rather than orthogonal rotations in an attempt to obtain greater physical insight. That is, we no longer insist on the empirical functions being orthogonal by changing the optimisation criterion used to extract them. We suggest

projection pursuit as a potentially more useful general framework. In this approach we use physical knowledge to define an optimisation criterion, which is then optimised to derive low-dimensional, interpretable descriptors of complex, high-dimensional data.

Considerable use is made of regression and correlation methods, where linear relationships are typically assumed. This assumption is of particular concern since many potentially predictive relationships will never be discovered. For example, a quadratic relationship between two variables has a correlation of zero. However, if we correctly identify the quadratic relationship then we naturally find a high correlation between fitted and observed values. An emphasis on statistical model building using appropriate tools is required, and we suggest a number of these.

Statistical methods are typically of most value in data rich/knowledge poor scenarios or where uncertainty and noise are present. Climate forecasting is a combination of both of these, in varying degrees depending on the particular application. Therefore there is a clear need for appropriate statistical methods. Our review of the climate forecasting literature yielded two key themes: climate processes are inherently non-stationary and nonlinear. The statistical methods applied are typically linear in nature, and require assumptions of stationarity to be meaningful. Viewed in this light it is unlikely that such methods will reveal more than superficial physical insights and predictability.

3. We have sought to develop a statistical framework that has at its foundation a nonlinear dynamical statistical model, which we call a statistical-dynamical model. Results of preliminary model fitting to monthly rainfall data are very encouraging, with fundamental properties of the rainfall series reproduced with some degree of success.

5.2 Future Research

Our proposed research plan for Phase 3 of IOCI is as follows:

1. Completion of the rainfall and sea surface temperature (SST) exploratory data analysis commenced in Phase 1 of IOCI;

2. Development of the nonlinear modelling approach within a Bayesian statistical framework. This effort will be focused initially on rainfall forecasting;

3. Apply a range of contemporary statistical methods to identify potential ocean and atmospheric rainfall predictors for the nonlinear model;

4. Develop an approach to predictor selection in nonlinear models using the reversible jump Markov chain Monte Carlo methodology;

5. Apply the nonlinear modelling approach to the downscaled 1000-year CSIRO9 GCM run.

Outcomes from this work will include:

1. A Bayesian statistical framework that uses probability distributions to represent uncertainty about model parameters. This will provide a probabilistic risk assessment approach to climate forecasting. An example output would be a predictive probability distribution for winter rainfall some months ahead;

2. Physical insights into the nonlinear mechanisms that generate rainfall. In particular, the modelling results will provide some information on the factors that influence the switching of rainfall between different regimes;

3. A statistical approach to identifying important climate predictor variables, applied to rainfall in the first instance; and

4. Statistical insights into the nonlinear mechanisms producing GCM output, which may be helpful when interpreting observational data.

Our proposed research linkages are:

1. *Bureau of Meteorology (Research Centre and Perth Regional Office)* – Identification of potential rainfall predictors; physical interpretation of nonlinear modelling results; forecasting skill comparison of nonlinear models and existing approaches.

2. *CSIRO Atmospheric Research* – Application of nonlinear modelling approach to GCM output; physical interpretation of data-based nonlinear modelling results.

3. *CSIRO Land and Water* – Downscaling of GCM output; physical interpretation of nonlinear modelling results.

4. *CSIRO Marine Research*- Identification of potential rainfall predictors; physical interpretation of nonlinear modelling results.

REFERENCES

Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 49, 803-821.

Barnett, T.P. (1983). Interaction of the Monsoon and Pacific trade wind system at interannual time scales. Part I: The equatorial zone. *Mon. Wea. Rev.,* 111, 756-773.

Bretherton, C.S., Smith, C. & Wallace, J.M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *J. Climate,* 5, 541-560.

Burkhardt, U. & James, I.N. (1998). Measuring the intensity of a storm track- An EEOF approach. *7th International Meeting on Statistical Climatology*. Whistler, BC, Canada.

Casey, T. (1995). Optimal linear combination of seasonal forecasts. *Aust. Met. Mag.,* 44, 219-224.

Cherry, S. (1996). Singular value decomposition analysis and canonical correlation analysis. *J. Clim.,* 9, 2003-2009.

Cleveland, W.S. & Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Assoc.,* 83, 596-610.

Cohen, A. & Jones, R.H. (1969). Regression on a random field. *J. Am. Statist. Assoc.,* 64, 1172-1182.

Denison, D.G.T. & Mallick, B.K. (1998). A nonparametric Bayesian approach to modelling nonlinear time series. Department of Mathematics, Imperial College, Technical Report.

Denison, D.G.T., Mallick, B.K. & Smith, A.F.M. (1998). Automatic Bayesian curve fitting. *J. R. Statist. Soc. B,* 60, 333-350.

Drosdowsky, W. (1993a). An analysis of Australian seasonal rainfall anomalies: 1950-1987. I: Spatial patterns. *Int. J. Climatol.,* 13, 1-30.

Drosdowsky, W. (1993b). An analysis of Australian seasonal rainfall anomalies: 1950-1987. II: Temporal variability and teleconnection patterns. *Int. J. Climatol.,* 13, 111-149.

Drosdowsky, W. (1993c). Potential predictability of winter rainfall over southern and eastern Australia using Indian Ocean sea-surface temperature anomalies. *Aust. Met. Mag.,* 42, 1-6.

Drosdowsky, W. (1994). Analog (nonlinear) forecasts of the Southern Oscillation index time series. *Weather and Forecasting,* 9, 78-84.

Drosdowsky, W. & Chambers, L. (1998). Near global sea surface temperature anomalies as predictors of Australia seasonal rainfall. Bureau of Meteorology Research Centre, Research Report 65.

Drosdowsky, W. & Williams, M. (1991). The Southern Oscillation in the Australian region. Part I: Anomalies at the extremes of the Oscillation. *J. Climate,* 4, 619-639.

Eynon, B.P. & Switzer, P. (1983). The variability of rainfall acidity. *Can. J. Statist.,* 11, 11-24.

Freiberger, W. & Grenander, U. (1965). On the formulation of statistical meteorology. *Rev. Int. Statist. Inst.,* 33, 59-86.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Ann. Statist.,* 19, 1-50.

Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression. *J. Am. Statist. Assoc.,* 76, 817-823.

Friedman, J.H., Stuetzle, W. & Schroeder, A. (1984). Projection pursuit density estimation. *J. Am. Statist. Assoc.,* 79, 599-608.

Gentilli, J. (1972). *Australian Climate Patterns,* Nelson, Melbourne.

Glaseby, C.A. (1998). Modelling multivariate spatio-temporal weather data using latent Gaussian processes. *7th International Meeting on Statistical Climatology.* Whistler, BC, Canada.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* 82, 711-732.

Hardle, W., Lutkepohl, H. & Chen, R. (1997). A review of nonparametric time series analysis. *Int. Statist. Rev.*, 65, 49-72.

Hastie, T., Buja, A. & Tibshirani, R. (1995). Penalised discriminant analysis. *Ann. Statist.,* 23, 73-102.

Hastie, T. & Tibshirani, R. (1986). Generalised additive models. *Statist. Sci.,* 1, 297-318.

Hastie, T. & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc. B,* 58, 155-176.

Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Assoc.,* 89, 1255-1270.

Horel, J.D. (1984). Complex principal component analysis: Theory and examples. *J. Clim. Appl. Meteorol.,* 23, 1660-1673.

Hutchinson, M.F. (1995). Stochastic space-time weather models from ground-based data. *Agric. For. Meteorol.,* 73, 237-264.

Ikeda, N. & Watanabe, S. (1989). *Stochastic Differential Equations and Diffusion Processes,* North-Holland/Kodansha, Tokyo.

Jolliffe, I.T. (1989). Rotation of ill-defined principal components. *Appl. Stats.,* 38, 139-147.

Jolliffe, I.T. (1998). Cluster analysis: Some recent developments and their relevance to Climatology. *7th International Meeting on Statistical Climatology*. Whistler, BC, Canada.

Jones, M.C. & Sibson, R. (1987). What is projection pursuit? *J. R. Statist. Soc. A,* 150, 1-36.

Jones, R. H. & Zhang, Y. (1996) Models for continuous stationary space-time processes , Unpublished Manuscript, .

Lall, U., Sangoyomi, T. & Abarbanel, H. D. (1996). Nonlinear dynamics of the Great Salt Lake: Nonparameteric short-term forecasting. *Water Resour. Res.,* 32, 975-985.

Lavery, B., Joung, G. & Nicholls, N. (1997). An extended high-quality historical rainfall dataset for Australia. *Aust. Met. Mag.,* 46, 27-38.

Lewis, P. A. W. & Ray, B. K. (1997). Modeling long-range dependence, nonlinearity, and periodic phenomena in sea surface temperatures using TSMARS. *J. Am. Statist. Assoc.,* 92, 881-893.

Lewis, P. A. W. & Stevens, J. G. (1991). Nonlinear modelling of time series using Multivariate Adaptive Regression Splines (MARS). *J. Am. Statist. Assoc.,* 86, 864-877.

Lin, T. C. & Pourahmadi, M. (1998). Nonparametric and non-linear models and data mining in time series: a case-study on the Canadian Lynx data. *Appl. Statist.,* 47, 187-201.

Meiring, W. & Nychka, D. W. (1998a). Functional data analysis for vertical profiles. *Interface 98.*

Meiring, W. & Nychka, D. W. (1998b). Functional data analysis of vertical Ozone profiles. *7th International Meeting on Statistical Climatology.* Whistler, BC, Canada.

Monahan, A. H. (1998). Nonlinear principal component analysis. *7th International Meeting on Statistical Climatology.* Whistler, BC, Canada.

Nason, G. (1995). Three-dimensional projection pursuit. *Appl. Statist.,* 44, 411-430.

Nicholls, N. (1986). Use of the southern oscillation to predict Australian sorghum yield. *Agric. For. Meteorol.,* 38, 9-15.

Nicholls, N. (1987). The use of canonical correlation to study teleconnections. *Mon. Wea. Rev.,* 115, 393-399.

Nicholls, N. (1989). Sea surface temperatures and Australian winter rainfall. *J. Clim.,* 2, 965-973.

Nicholls, N. (1991). The El Niño / Southern Oscillation and Australian vegetation. *Vegetatio 91: Vegetation and climate interactions in semi-arid areas.*

Nicholls, N. & Katz, R. W. (1991) In *Teleconnections linking worldwide climate anomalies* (Eds, Glantz, H. and Nicholls, N.) Cambridge University Press, New York, pp. 511-525.

Nott, D. & Dunsmuir, W. T. M. (1998). Analysis of spatial covariance structure from monitoring data. Department of Statistics, UNSW, Technical Report S98-6.

Obled, C. & Creutin, J. D. (1986). Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *J. Clim. Appl. Meteorol.,* 25, 1189-1204.

Oehlert, G. W. (1993). Regional trends in sulfate wet deposition. *J. Am. Statist. Assoc.,* 88, 390-399.

Ogden, R. T. (1997) *Essential Wavelets for Statistical Applications and Data Analysis,* Birkhäuser, Berlin.

O'Sullivan, F. (1991). Discretised Laplacian smoothing by Fourier methods. *J. Am. Statist. Assoc.,* 86, 634-642.

Paterson, J. G., Goodchild, N. A. & Boyd, W. J. R. (1978). Classifying environments for sampling purposes using a principal component analysis of climatic data. *Agric. Meteorol,* 19, 349-362.

Pezzulli, S. & Silverman, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Comp. Statist.,* 8, 1-16.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis,* Springer-Verlag, New York.

Ramsay, J. O. D., C.J. (1991). Some tools for functional data analysis. *J. R. Statist. Soc. B,* 53, 539-572.

Richman, M. B. (1986). Rotation of principal components. *J. Climatol.,* 6, 293-335.

Rimmington, G. M. & Nicholls, G. M. (1993). Forecasting wheat yields in Australia with the southern oscillation index. *Aust. J. Agric. Res.,* 44, 625-32.

Russell, J. S., McLeod, I. M., Dale, M. B. & Valentine, T. R. (1993). The southern oscillation index as a predictor of seasonal rainfall in the arable areas of the inland Australian subtropics. *Aust. J. Agric. Research,* 44, 1337-49.

Sampson, P. D. (1986). Spatial covariance estimation by scaled-metric scaling and biorthogonal grids. Department of Statistics, University of Washington, Technical Report 91.

Sampson, P. D. & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Assoc.,* 87, 108-119.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.,* 24, 1-24.

Smith, I. (1994). Indian Ocean sea-surface temperature patterns and Australian winter rainfall. *Int. J. Climatol.,* 14, 287-305.

Stein, M. (1986). A simple model for spatial-temporal processes. *Water Resour. Res.,* 22, 2107-2110.

Stone, R. C., Smith, I. & McIntosh, P. (1997). Statistical Methods For Deriving Seasonal Climate Forecasts From GCMs. *The symposium on applications of seasonal climate forecasting in agricultural and natural ecosystems.* Brisbane.

Sugihara, G. & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurment error in time series. *Nature,* 344, 734-741.

Tong, H. (1990) *Non-linear time series. A dynamical systems approach,* Oxford University Press, New York.

Walden, A. T. (1994). Spatial clustering: using simple summaries of seismic data to find the edge of an oil-field. *Appl. Statist.,* 43, 385-398.

Wallace, J. M., Smith, C. & Bretherton, C. S. (1992). Singular-value decomposition of sea surface temperature and 500-mb height anomalies. *J. Climate,* 5, 561-576.

Weare, B. C. & Nasstrom, J. S. (1982). Examples of extended empirical orthogonal function analyses. *Mon. Wea. Rev.,* 110, 481-485.

Wolter, K. (1987). The Southern Oscillation in surface circulation and climate over the tropical Atlantic, eastern Pacific, and Indian Oceans as captured by cluster analysis. *J. Clim. Appl. Meteorol.,* 26, 540-558.

Wright, P. B. (1974). Seasonal rainfall in southwestern Australia and the general circulation. 102, 219-232.

Zhang, X., Sheng, J. & Shabbar, A. (1998). Modes of interannual and interdecadal variability of Pacific SST. *J. Climate,* 11, 2556-2569.

APPENDIX A - GLOSSARY

Cross-referenced terms and acronyms are shown in italics.

Bayesian	A statistical framework that expresses uncertainty using probability distributions. Bayesian statisticians explicitly combine data with subjective knowledge to learn about physical processes. This is accomplished using *Bayes' theorem*.
Bayes' Theorem	As implemented in scientific practice, this theorem essentially states that uncertainty conditional on available data and expert knowledge is proportional to the product of the uncertainty in the data and the uncertainty in expert knowledge.
Correlation	A measure of the strength of a linear relationship between two variables.
Covariate	A variable used in a statistical model to predict the value of a response variable. For example, in predicting rainfall a suitable covariate might be *SST* at a particular location. Covariates are often known as predictors, and these two terms are used interchangeably.
Linear	A general term to describe relationships that can be represented as straight lines between two variables, or hyperplanes for many variables.
Loess	A technique for smoothing two-dimensional data that is robust to very noisy signals.
Markov chain Monte Carlo	A computationally intensive technique that uses simulation techniques to implement *Bayesian* statistical methods. This term is universally known by the acronym *MCMC*.

Nonlinear	A general term to describe relationships that cannot be described as straight lines or hyperplanes, as is the case for *linear* relationships.
Reversible Jump MCMC	A methodology for choosing optimal statistical models in a *Bayesian* statistical framework, motivated by *MCMC* ideas.
Spectral Analysis	A technique for detecting regular patterns in *time series*.
Spectral Leakage	In spectral analysis, strong signals can often obscure weaker signals that have a similar period. This is known as leakage.
Spline	A technique for approximating functions, typically accomplished by breaking the domain of the function into segments within each of which some simple function is fitted to the data.
Stationary	A physical process is said to be stationary if its generating mechanism does not change in time and/or space.
Time Series	A set of data recorded sequentially in time.
Wavelet Analysis	A technique similar to *spectral analysis*, but with some additional benefits. In particular, wavelet analysis suffers much less from the problem of *spectral leakage* and can be applied to data that are not *stationary*.

APPENDIX B - LIST OF ACRONYMS

CCA   Canonical Correlation Analysis.

CLW   CSIRO Land and Water.

CMIS   CSIRO Mathematical and Information Sciences.

CMR   CSIRO Marine Research.

DWT   Discrete Wavelet Transform.

EOF   Empirical Orthogonal Function.

FDA   Functional Data Analysis.

GCM   Global Climate Model.

GISST   Global Sea-Ice and Sea Surface Temperature data set prepared by the Hadley
                Centre of the United Kingdom Meteorological Office.

IOCI   Indian Ocean Climate Initiative.

ITSMARS   Interactive TSMARS- a program to implement TSMARS.

LDA   Linear Discriminant Analysis.

MARS   Multivariate Adaptive Regression Splines.

MCMC   Markov chain Monte Carlo.

MRA   Multi-Resolution Analysis.

MSLP   Mean Sea Level Pressure.

PCA   Principal Component Analysis.

PDA   Penalised Discriminant Analysis.

SST   Sea Surface Temperature.

SWA   South-west Western Australia.

TSMARS   Time Series MARS.