Chapter 1 Introduction

As demands upon existing global water resources increase, the accurate measurement of water availability assumes greater importance. The efficient management of future water supplies will demand a clear understanding of the many interactions within the hydrological cycle, in particular the impact of climate variability upon the spatial and temporal distribution of rainfall. Interactions between various global climate phenomena produce protracted wet and dry cycles, a characteristic referred to in this work as hydrological persistence. Knowledge of the likely timing, duration and severity of persistent low rainfall periods will likely assist in reducing possible social and economic impacts. In this context, there is a clear requirement for statistical models that improve both the identification and the explanation of coherent patterns within rainfall and streamflow records. This thesis demonstrates the efficacy of hidden Markov models (HMMs) to encapsulate hydrological persistence primarily at monthly time scales.

1.1 The influence of climatic persistence

The climate of Australia experiences high interannual variability, with rainfall being sensitive to anomalies in climate modes such as the El Niño Southern Oscillation (ENSO) and ocean-atmospheric interactions across both the Indian and Southern Oceans. Although slowly-varying global circulation phenomena are known to have strong teleconnections with hydrologic observations, many stochastic models designed to simulate and forecast such data fail to provide a useful description of these climate influences. The supposition of the broader climate having a tendency to fluctuate between a discrete number of stable regimes (or “states”) has a long history in meteorological studies. This concept lends itself directly to a rationalisation of hydroclimatic persistence.

The natural persistence of wet and dry spells within Australian hydrology is illustrated in Figure 1.1 using reconstructed natural flows for the River Murray. These time series demonstrate deviations from median values for records of both annual totals and monthly variates that are produced by removing annual seasonality from the record of monthly flows. Extended periods either side of these median thresholds are apparent at each time scale, however this characteristic is ostensibly more dramatic in the 10-year sample of monthly variates. Within this sample period, a protracted two-year period during which monthly variates remain below the long-term median is followed shortly after by a wet period of similar duration. Characteristics such as these are apparent across a wide range of hydrologic data for Australia, revealing an underlying tendency towards hydrological persistence, particularly at a monthly time scale.

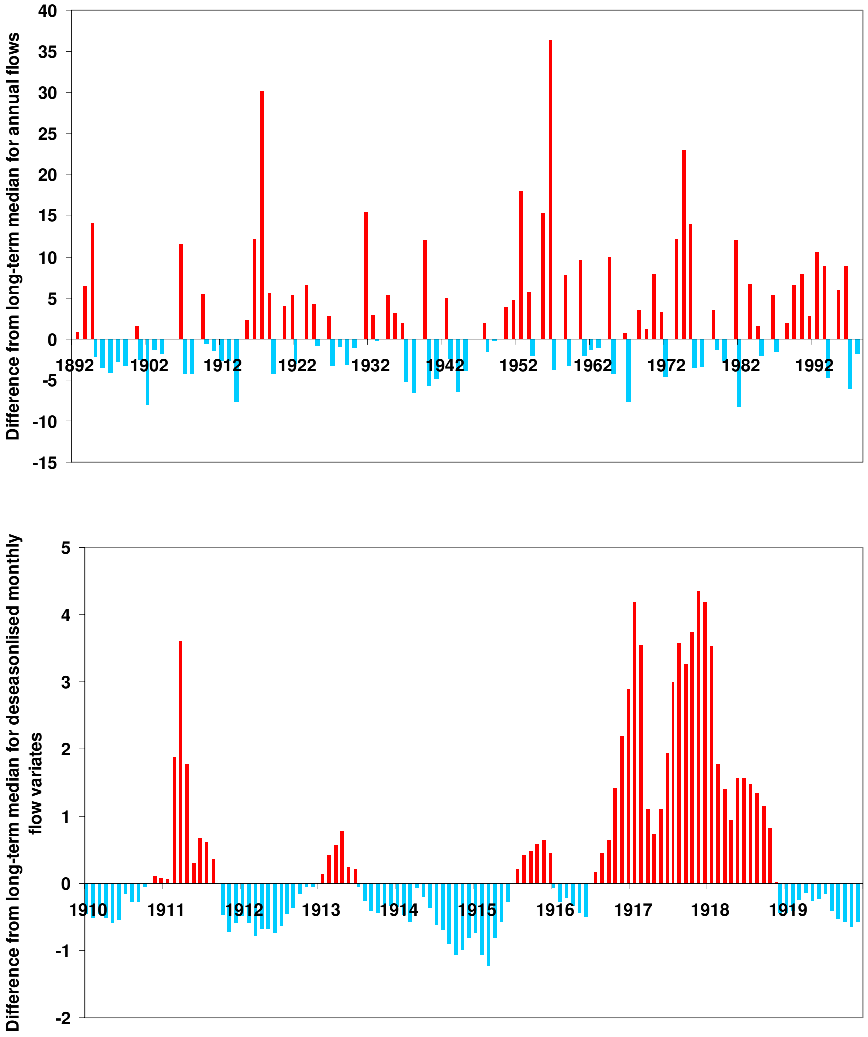

Figure 1.1 Deviations from median values of reconstructed natural flows in the River Murray evaluated for annual totals (x10^3 Gigalitres, top) and deseasonalised monthly variates (bottom)

The design and management of water resource infrastructure requires accurate simulations of monthly rainfall and streamflow totals. Furthermore it is important that stochastic rainfall models also replicate observed patterns of wet and dry spells. The design of hydrological time series models however raises a number of questions. Firstly, are prominent features within hydrologic records predictable over the long-term? Secondly, can the relationship between these features and climatic fluctuations be substantiated?

Temporal changes in climate proxies, such as sea surface temperature and atmospheric pressure records, are used to classify variations in the broader climate. Although these variations may also indicate changes in rainfall observations, clear linkages between these hydroclimatic variables are often concealed through the influence of atmospheric instabilities. Even if the magnitude and structure of climate anomalies were similar across each climate event, the hydrologic response to each would differ. Furthermore statistics of localised weather conditions are poorly represented within simulations and seasonal forecasts of atmospheric global circulation models (GCMs). Therefore rather than attempting to use the output of such broad-scale models to produce hydrologic simulations, an alternative approach is to design time series models that capture localised hydroclimatic oscillations.

Hidden Markov models (HMMs) have enjoyed extensive use across a wide range of scientific fields. Within these parsimonious models, the probability of localised rainfall at a monthly scale is conditioned upon a small number of discrete weather states that are unobserved (i.e. "hidden"), yet estimated through the calibration of this model to observed data. HMMs offer an improved conceptual approach to describing hydrologic responses to broader-scale climatic oscillations than linear time series models such as ARMA models.

HMMs have been promoted as suitable time series models for hydrological persistence at an annual scale (eg Thyer and Kuczera, 2000; Srikanthan *et al.*, 2002b), yet rarely for the description and simulation of monthly data, which is a more appropriate time scale with regard to the dominant climate frequencies. Moreover the analysis of monthly data provides an obvious increase in sample size than the analysis of annual totals. In order to improve the application of HMMs, many aspects of their structure can be developed, using a rigorous statistical framework to quantify uncertainties in both the observation data and the calibration.

1.2 Objectives of Thesis

The main goal of this thesis is to develop stochastic approaches to assist in the identification and explanation of hydrological persistence within observed data setes measured at a range of time scales. This goal comprises the following objectives:

- To use a rigorous statistical framework to identify hydrological persistence in over various time scales, and to reconcile this persistence with climatic fluctuations
- To analyse distinctions between hydrological persistence and standard time series definitions of persistence and investigate models that are specific to the former
- To develop hidden Markov models (HMMs) to provide adequate descriptions of hydrological persistence at a monthly scale
- To reduce the reliance of conventional HMMs on assumptions concerning the parametric form of conditional observations
- To analyse hydrological persistence within various observed data
- By using an explicit framework for hydrological persistence, to produce accurate simulations of monthly rainfall data that would assist in numerous water resource applications

1.3 Outline of Thesis

Chapter 2 provides the background to the analysis of hydroclimatic persistence, summarising interactions between sources of climate variability and their teleconnections with hydrologic observations across a range of time scales. By reviewing previous studies into the dominant sources of climate variability, this chapter develops relationships between climate fluctuations and hydrological persistence. Climate indices are introduced as one possible approach to relate changes in dominant atmospheric circulation modes to hydrological variability.

Chapter 3 focuses upon the theory of hydrological persistence, defining this in terms of spells analysis. A range of statistical tests used to quantify hydrological persistence are introduced, alongside the interpretation of mathematical persistence revealed as the Hurst phenomenon. Existing approaches to the stochastic modelling of hydrologic data are then presented together with a discussion of the inadequacy of such models to provide a clear description of temporal persistence. Against this discussion, the benefits of the HMM approach is reviewed alongside the rigorous statistical methodology that can be embraced to fully account for parameter uncertainty. Previous applications of both the HMM approach and the Bayesian paradigm to stochastic hydrology are discussed.

Chapters 4 and 5 present a clear case for modelling persistence in Australian hydrologic series. Using spatially-averaged rainfall data and streamflow records from across the country, evidence for significant persistence at a monthly scale is presented through various runs statistics. Sources of climate variability, characterised through various indices, are shown to modulate these hydrologic data. The hydroclimatic influence of global circulation modes is further demonstrated through the analysis of various arid-zone hydrological data series. Relationships between climate indices and statistics derived from spells analyses show evidence that sources of broad-scale climate variability explain some of the persistence in hydrologic data that is modelled encapsulated with HMMs. Importantly the Hurst phenomenon, an alternative interpretation of persistence, is shown to be inconsistent with a spells-based interpretation of hydrological persistence. This provides further a further demonstration of the applicability of HMMs for modelling time series displaying significant clustering of hydrologic data.

Chapter 6 describes the calibration of HMMs to annual rainfall data for six mainland capital cities of Australia, together with Alice Springs, and the spatially-averaged data for the meteorological district surrounding Sydney. By improving the statistical approaches used in previous investigations of persistence in these data, the results in this chapter demonstrate that there is insufficient data at the annual scale to identify significant persistence. The circumstances under which the construct of persistence in HMMs becomes redundant are developed, in the process presenting a straightforward method to identify persistence accurately. Using these results, HMMs are then calibrated to monthly data (with intra-annual variability removed) in Chapter 7, and illustrate statistically significant persistence at this time scale. The monthly scale is consistent with the dominant frequencies of climatic persistence, and is suggested to be more appropriate than an annual scale for investigating hydrological persistence. Importantly, previous studies into the usefulness of HMMs to describe persistence have not investigated the calibrations of these models to monthly rainfall or streamflow data.

Chapter 8 focuses upon a novel development of the conventional HMM formulation, a non-parametric approach in which assumptions about the form of state conditional distributions are relaxed. This model provides an unbiased description of persistence, and adapts existing methods for HMM calibration. The robust and flexible structure of this model is illustrated through its calibration to hydrologic data from across Australia, with results being consistent with spells analyses as well as climate observations. The ability of this non-parametric HMM to be calibrated to both discrete and continuous-valued data is illustrated through an analysis of persistence in both monthly rainfall totals and time series of monthly rain-days. The identification of persistence in related data including short-duration pluviograph data provides further evidence for hydrological persistence at a monthly scale.

The conventional structure of parametric HMMs is developed in Chapter 9 through the relaxation of certain statistical assumptions. Related statistical models such as hidden semi-Markov models (HSMMs) and autoregressive hidden Markov models (ARHMMs) are described, and calibrated to the monthly rainfall totals for Sydney. These results demonstrate the efficacy of these models to describe hydrological persistence, although such application has been rarely described. Furthermore a novel hierarchical HMM, which describes persistence at monthly and annual frequencies simultaneously, is introduced and calibrated to monthly rainfall data. Chapter 10 provides a more thorough investigation into the nature of persistence in Australian hydrology, combining the capabilities of both parametric and non-parametric HMMs to provide estimates for underlying probability distributions in various persistent data. These results demonstrate the benefit of Bayesian model selection methods to identify the most appropriate models for monthly data series.

Finally, Chapter 11 illustrates how the calibration of HMMs can be adapted in order to generate multiple simulations of monthly hydrological data, the accuracy of which is measured against simulations from conventional linear models. These results demonstrate the efficacy of the HMM approach to describe monthly persistence and to provide accurate simulations of hydrologic data at a range of temporal aggregations. Furthermore, statistics describing characteristics of persistence are simulated accurately with HMMs. Catchment-scale rainfall simulations from HMMs are assessed in terms of reservoir reliability, demonstrating another useful application of this modelling approach in stochastic hydrology. The final chapters of this thesis summarise the main conclusions, and discuss possible avenues for future research.

Chapter 2 Identifying persistence within the global climate

Stochastic models for hydrologic processes have rarely incorporated the influence of slowly-varying climate modes. Hydrologic observations reveal extended periods of above and below-average values, and oscillations between such periods are closely related to low frequency climate phenomena (eg Kiem *et al.*, 2003). Protracted dry periods may develop into drought conditions that can ultimately place increased risk upon water resources. The analysis of wet and dry sequences has long been the focus of hydrologic studies, and time series models that incorporate climatic persistence may improve simulations and forecasts, both of which are fundamental to risk assessment in water resource management. The focus of this thesis is on the identification and modelling of hydrological persistence. In order to recognise the prevalence of this persistence, it is important to illustrate the mechanisms through which fluctuations in global climate modes impact upon the hydrological cycle. This chapter presents the context for the wider study by examining the complex hydroclimatic interactions that produce persistence in rainfall and streamflow observations. A thorough review of literature that examines global circulation modes is included here, along with descriptions of techniques used to quantify climatic changes.

2.1 Observed modes of climatic persistence

The global climate system contains modes of variability that persist over characteristic periods. Oscillations within these modes have substantial impacts upon precipitation and temperature across the world. Major circulation phenomena have large-scale teleconnection patterns (Allan *et al.*, 1996); the best known is the El Niño Southern Oscillation (ENSO), which develops in the Pacific Ocean yet impacts globally. Other modes include the North Atlantic Oscillation (NAO) and the Pacific Decadal Oscillation (PDO), both of which affect primarily regions of the Northern Hemisphere, and oscillations in the tropical Atlantic that influence climatic variability across South America and western Africa. In order to analyse persistence within hydrologic records it is important to focus upon the climatic source of this persistence.

Time series indices that describe these oscillatory phenomena tend to show prolonged periods of above- or below-average values. This feature demonstrates an intrinsic pattern of climatic persistence at a range of time scales. Lockwood (2001) described the climate system as a dissipative, non-linear system, with many sources of instabilities. Slowly varying surface conditions, such as sea surface temperatures and land surface moisture levels, can act as boundary conditions that significantly influence the climate of the atmosphere (Arnell, 2002).

Feedback mechanisms between the atmosphere and these surface boundary conditions can amplify anomalous values into persistent cycles. In some years, regional climates may therefore display a certain array of characteristics and in other years a different set.

The climate of Australia includes numerous modes of variability. Anomalous conditions in the Pacific, Indian and Southern Oceans, which surround this continent, contribute to variability within the surface climate. Baines (1998) used *Cerberus*, the name of a three-headed dog of Greek mythology, to describe the three major sources of variation in the Australian climate. The El Niño Southern Oscillation, the Indian Ocean dipole and the Antarctic Circumpolar Wave interact over different time scales to augment and to moderate their individual effects. The interaction of climate modes produces a forcing mechanism in the hydrological cycle that is revealed through persistent periods of high and low observations. In order to elucidate the influence of these different climate modes upon persistence in Australian hydrology, an overview of their characteristics is presented.

2.1.1 El Niño Southern Oscillation (ENSO)

The El Niño Southern Oscillation (ENSO) is the primary source of global climate variability acting over the 2- to 7-year time scale (eg Ropelewski and Halpert, 1987; Katz, 2002; Viles and Goudie, 2003). Its cyclic patterns of warming and cooling in the surface waters of the tropical Pacific produce prominent global teleconnections with regional rainfall and temperature patterns (Hamlet and Lettenmaier, 1999). The quasi-periodicity of ENSO, which sees the appearance of warm or cool water in the equatorial eastern and central Pacific at intervals of 3 to 5 years, is one of its most fundamental aspects (Graham and White, 1988), and is intimately linked with the conservation of latent heat through the waters of the Pacific. Importantly, ENSO is the most prominent source of variability within rainfall and streamflow records across much of Australia.

There is a range of interpretations in the literature on the use of the terms El Niño, La Niña and ENSO. The term El Niño has evolved in its meaning over the years, being originally applied to the appearance of a warm current flowing southward along the Pacific coast of South America each year around Christmas. However as a result of the frequent association between South Pacific ocean temperatures and interannual basin-wide equatorial warm events (McPhaden *et al.*, 1998), the term has become synonymous with larger-scale climatically significant ocean-atmosphere interactions, which are significant features of the global climate. The El Niño coastal warming is actually part of a broad weather system that affects both the Pacific and Indian Oceans, and the oscillation of high pressure cells across the Pacific to which it is related is the true cause of the global climatic phenomenon (Diez, 2004). The opposite conditions to El Niño, termed La Niña events, consist of basin-wide cooling of the tropical Pacific, with the coupled ocean-atmospheric processes termed the El Niño Southern Oscillation (ENSO).

Figure 2.1 (after McPhaden *et al.*, 1998) provides an overview of ocean and atmospheric changes that occur with El Niño events in the Pacific. Under "normal" conditions shown at the top of this figure, the warmer surface layer of the tropical Pacific created from atmospheric heat gain flows westward under the influence of easterly trade winds. This piling up of warmer water creates a large sea surface temperature (SST) gradient along the equatorial Pacific, and an east-west contrast in atmospheric pressure across the Pacific that has lower surface pressures over the warm pool in the west. This "see-saw" in atmospheric pressures across the Pacific (Chiew *et al.*, 1998) is the Southern Oscillation (SO). The convective loop of atmospheric circulation that is known as the "Walker Circulation" sees warm moist air rise in the west, move to the east and subside in the high-pressure zones of the eastern Pacific.

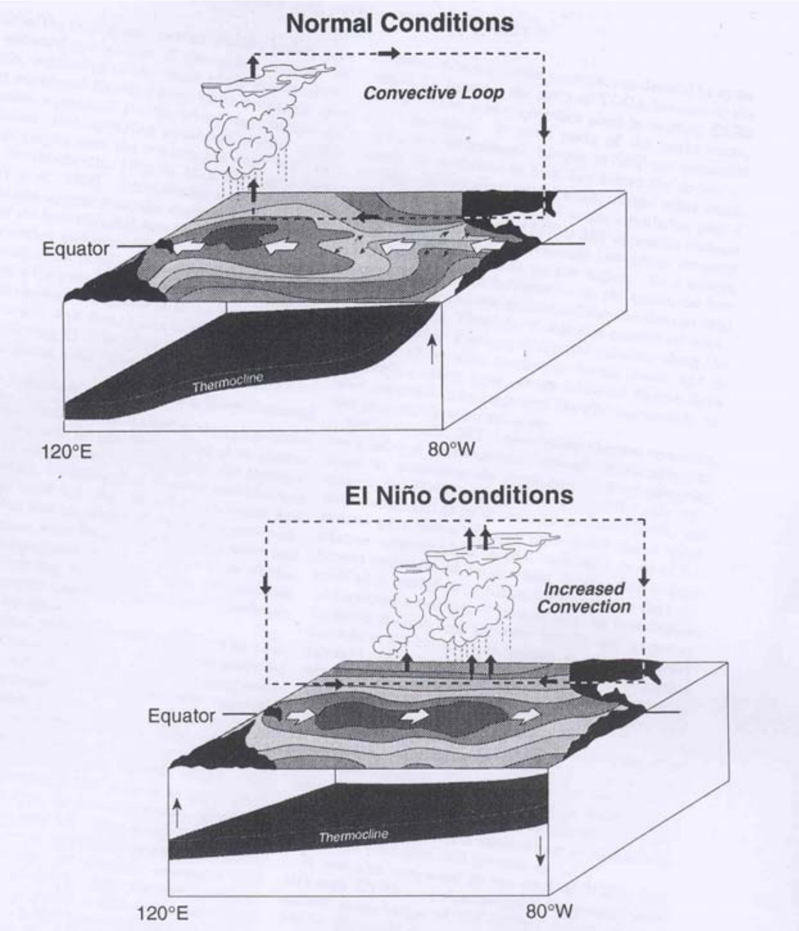

Figure 2.1 Mechanisms of ENSO (after McPhaden et al., 1998)

El Niño events, shown in the lower diagram of Figure 2.1, are characterised by a large-scale weakening of the trade winds, easterly movement of surface waters and the subsequent warming of the eastern and equatorial Pacific (McPhaden *et al.*, 1998). The easterly acceleration of surface currents helps to depress the thermocline to the east, and causes anomalously warm SSTs to be observed near the coast of South America. The area of anomalously warm surface water at the peak of an El Niño episode can reach 30 million km^2 (Rial *et al.*, 2004), so it is clear that the size of the latent heat exchange at the ocean-atmosphere interface is large enough to alter global climate patterns. Temporal changes in SSTs and atmospheric pressures are incorporated into a range of climate indices that define the ENSO phenomenon, and these are described further in Section 2.2.

The approximate periodicity of the ENSO phenomenon has long been recognised (eg Bjerknes, 1969), with recent evidence suggesting its quasi-cyclic behaviour is due to the action of a natural coupled oscillator of the ocean-atmosphere system (Graham and White, 1988; Tziperman *et al.*, 1994; Wang, 2001). In such a model, warm and cool episodes are regarded as alternate phases of a self-sustaining cycle (Graham and White, 1988), rather than discrete episodes overlying a mean background state. Tziperman *et al.* (1994) view the irregular oscillations as those of a low-order chaotic system, driven by a seasonal cycle.

The global ENSO signal consists of a global standing mode in covarying sea level pressure (SLP) and SST anomalies (White and Annis, 2004), evolving in association with the intense warm SST anomalies in the eastern equatorial Pacific termed El Niño episodes. Recent observations of ENSO show that the global standing mode associated with the El Niño cycle is superimposed upon an eastward propagating global ENSO wave (GEW) (White and Cayan, 2000). Tourre and White (1997) identified the GEW through interannual SST and SLP anomalies originating in the western tropical Indian Ocean, and propagating slowly eastward through the Indian and Pacific Oceans, taking 4 to 6 years to circle the globe. The GEW can then provide a positive feedback to El Niño and La Niña events in the eastern equatorial Pacific (White and Cayan, 2000), supplying an additional source of tropical ENSO variability.

2.1.2 Low frequency climate variability and abrupt climate changes

The interannual scale of climate variability that is associated with ENSO is widely recognised, however the global climate contains various other circulation phenomena that have a range of frequencies and durations. The combined influence of these different modes of variation is expressed as a forcing mechanism upon regional climates, and is revealed through a tendency for stable conditions over extended periods. Coherent patterns of variability at multi-decadal time scales are observed in a range of climate data, with low frequency changes in the behaviour of ENSO shown to coincide with changes in global SST patterns. In spectra of globally-averaged SST and night-time marine air temperature for the period 1856-1981, Folland *et al.* (1984) found peaks at periods of 16 years and 21 years. These low frequency effects reflect circulation phenomena that modulate the influence of shorter period modes such as ENSO on regional climates. Importantly, these longer period climate patterns rapidly change phase, which in turn influences the persistent patterns of hydrologic variables.

Extensive literature has discussed the widespread changes in the Pacific climate during the late 1970s (Graham, 1994; Trenberth and Hurrell, 1994; Mantua *et al.*, 1997), that produced an extended ENSO-like warming of the tropical Pacific and cooling in the North Pacific (Arblaster *et al.*, 2002). Over the period since, SST-based indices display a tendency toward higher values. Of the various recorded examples of multi-decadal SST variability, climatic shifts that occurred around 1976 were the most widespread (Yonetani and Gordon, 2001), with Francis and Hare (1994) noting that abrupt changes in sea level pressure and surface air temperature occurred across the Pacific. These abrupt changes in different aspects of the Pacific climate have been termed a "regime shift" (Zhang *et al.*, 1997), which introduces a notion of stable climate states. At an atmospheric scale, Christiansen (2003) showed evidence for two such stable regimes in stratospheric circulations of the Northern Hemisphere, with an abrupt and statistically significant shift between these identified in the latter half of the 1970s.

Alongside distinct shifts that occurred in physical indices around 1976-7, evidence for widespread ecological changes has been presented (Mantua *et al.*, 1997), including dramatic shifts in marine and terrestrial ecosystem variables detected in the western Pacific through the mid 1970s. Reid *et al.* (1998) showed that implications of oceanic regime shifts for fisheries and oceanic CO_2 uptake are profound with significant changes in North Atlantic phytoplankton levels occurring over the second half over last century. Scheffer *et al.* (2001) also present time series of various marine ecosystem variables that show "conspicuous" jumps between apparent states, suggesting that climate state shifts may be reflected more consistently by biological data.

From observing consistent changes in a range of climate variables, it appears that the Pacific climate underwent a major transition during the 1970s and remained in a quasi-stable state for multiple decades. Although these changes might be interpreted as an abnormal fluctuation of a generally stable climate system, results presented by Mantua *et al.* (1997) suggest that the regime shift of the mid 1970s was not unique. Signatures of interdecadal climate variability are in fact widespread and detectable in various climate and ecological systems of the Pacific basin, at a number of times throughout recorded history. For example, similar abrupt changes occurred during the 1920s and 1940s. Zhang *et al.* (1997) note the climatic changes in the Pacific that occurred during the mid-1970s were analogous, but in the opposite direction, to those observed during around 1942-43. These authors showed that three independent indices, derived from SLP, air temperatures and SST data, displayed increases in mean values around 1976-77, also decreases during the 1940s. Wide-scale climatic changes during the 1940s were not confined to the Pacific however, with Allan *et al.* (1995) showing evidence for Indian Ocean mid-latitude SSTs to be consistently warmer over the period 1942-1983 than for the period 1900-1941. Krishna Kumar *et al.* (1995) demonstrate the significance of this change by showing that correlations between all-India summer monsoon rainfall and seven climatic predictors were insignificant until the early 1940s, after which point they become significant.

Abrupt regime shifts in the regional climate of Australia are manifested through step changes in dependent hydroclimatic variables, such as significant increases in summer and annual rainfall across New South Wales (NSW) during the 1940s (Cornish, 1977). Franks (2002b) tested the annual maximum discharges from 40 gauged streams in NSW for evidence of single step changes, interpreted as evidence for distinct climate states in the flood records. The results of this study indicated that 19 of the 40 gauges showed single step changes that were significantly different at a 99% level, with 16 of these changes occurring during the 1940s. By deriving a regional index of flood frequency, Franks also demonstrated a dramatic change in flood risk corresponding to 1945.

Scheffer *et al.* (2001) suggest a range of nonlinear mechanisms that explain rapid shifts between regimes. Abrupt changes, particularly through the mid-latitude atmospheric circulation, may be triggered through enhanced convective activity associated with persistent warm SSTs through the tropics (Graham, 1994). Latif and Barnett (1994) however suggest that the coupled ocean-atmosphere interaction is the true cause of inter-decadal variability in the North Pacific. Using this hypothesis, a study by Yonetani and Gordon (2001) showed that decadal variability simulated by a coupled general circulation model (GCM) contained abrupt climate changes, which are likely to influence hydrological cycles dramatically.

In light of the abundant evidence that suggests wide-scale dramatic changes to have occurred in the climate at various times, it is important to note the uncertainty associated with interpreting low-frequency variability in short time series. Using over 100 physical and biological time series, Hare and Mantua (2000) developed a composite analysis to argue definitively that a regime shift occurred through the North Pacific in 1976. However, with the lengths of many of these records being approximately 20 years, Rudnick and Davis (2003) suggested that there was simply not enough data to adequately confirm the existence of nonstationarity, expressed through significant regime shifts. To illustrate this, Rudnick and Davis (2003) simulated multiple independent and stationary time series, with frequency content identical to a low frequency series obtained from principal component analysis of Pacific SST data. By analysing 20-year spans of these simulated series in an identical manner to Hare and Mantua (2000), Rudnick and Davis (2003) showed that comparable regime shifts are detected about half the time. These authors found that the composite analysis used by Hare and Mantua (2000) identified changes in regime that were at a higher frequency than those simulated.

The danger in interpreting characteristics such as stable regimes in short climate records that display low-frequency behaviour is also addressed by Wunsch (1999), who noted that the purely random behaviour of stationary processes can appear “visually interesting” with regards to regime-like patterns, particularly over short periods. In particular, high serial correlation can produce stochastic trends such as random walks. Furthermore an assumption that climate data are non-normally distributed (typified in physical processes such as tropical rainfall) could explain regime-like nonstationarity. By increasing the resolution of observed data, such as through the analysis of monthly totals rather than annual totals, finer modes of nonstationarity may be identified. ENSO periods remain for an average of 15 months, and these periods would be difficult to identify through the use analysis of annual totals. Other persistent climate modes are discussed in the following sections.

2.1.3 Antarctic Circumpolar Wave (ACW)

The Southern Ocean is the only body of water that encircles the globe and contains a strong eastward flow known as the Antarctic Circumpolar Current that moves at a rate of approximately 0.01 m/s (Baines, 1998). It is the unifying link for exchanges of water masses between the major ocean basins (White and Peterson, 1996), and therefore expected to play a major role in the transmission of climate anomalies across the globe. Associated with this current is a recently discovered system of coupled sea-surface temperature anomalies, termed the Antarctic Circumpolar Wave (ACW), consisting of two large regions of relatively warm water alternating with regions of relatively cooler water, as shown in Figure 2.2. These anomalies propagate eastwards with a period of 4 to 5 years, and are allied with significant interannual variations in sea-level pressure, wind stress and sea-ice extent across the Southern Ocean (White and Peterson, 1996). The ACW occurs concurrently with the slow eastward propagation of the global ENSO wave (GEW), taking approximately 8 years to circle the globe (White *et al.*, 2002). Qiu and Jin (1997) showed the ACW to be independent of the GEW, based on their different zonal average speeds, although White *et al.* (2002) found the two phenomena to be linked in selected longitudinal domains.

ACW anomalies originate in the western Pacific, spreading to the south and east in the Southern Ocean. Therefore the slow eastward propagation of the ACW around the Southern Ocean is influenced by tropical ENSO in the central and southeastern Pacific (White and Cherry, 1999). The GEW reinforces the ACW in the eastern Pacific and western Atlantic sectors of the Southern Ocean (White *et al.*, 2002), such that this imposed GEW signal is propagated throughout the Southern Ocean. The warm SST anomalies of the ACW influence the local Australian climate of Australia when flooding into the Great Australia Bight and surrounding Tasmania. Southerly winds across the continent during these periods tend to be warmer and to carry more moisture than average, such that the rainfall regime of southern Australia is likely to be influenced by the ACW.

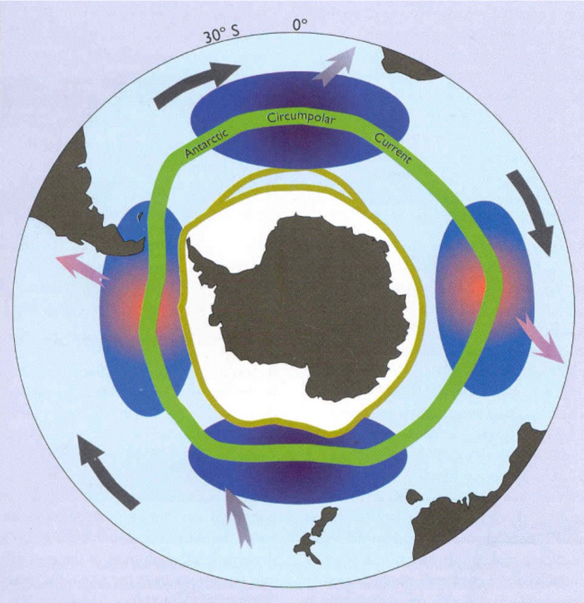

Figure 2.2 Schematic diagram of the ACW (after Baines, 1998)

Various studies have identified decadal, interdecadal and multi-decadal modes of variability in the ACW (eg Carril and Navarra, 2001). Low frequency changes in the ACW observed during the second half of the twentieth century may have occurred in response to similar changes in the central tropical Pacific and Indian Oceans. As described previously, the evolution of El Niño has been fundamentally different since 1976/7, consistent with its multi-decadal modes of variability. White and Annis (2004) showed that the evolution of the ACW to be fundamentally different before and after this time, suggesting a connection of climate modes with SST changes across the Indian and Pacific Oceans.

White (2000) examined the influence of three separate sources of climate variability on interannual precipitation anomalies across Australia; the ACW south of Australia, the northern branch of the ACW in the Indian Ocean and the GEW in the tropical north of Australia. In association with these sources of variability, covarying anomalies in SST and tropospheric moisture flux over interannual time scales were found to propagate eastward across the Indian Ocean, taking 2-3 years to progress from Africa to Australia. White (2000) developed a statistical climate prediction system that could significantly predict precipitation anomalies at lead times of 1-2 years. Eastward propagation of SST anomalies in the ACW south of Australia and the north branch of the ACW west of Australia can predict more than 50% of the total interannual variance over Western Australia, Victoria and New South Wales south of 20°S. In an earlier study, White and Cherry (1999) found autumn-winter temperature and precipitation records from across New Zealand, which appear to be independent of ENSO, to fluctuate with a 3-6 year mode of variability that was associated with the ACW. This ocean circulation phenomenon is an important influence upon the rainfall regime of Australia, producing interannual variability that is distinct from the effects of ENSO.

2.1.4 Indian Ocean Dipole (IOD)

Apart from internal modes of variability that have been identified in both the Pacific and Atlantic, a pattern of ocean-atmospheric interaction causing interannual climate variability in the Indian Ocean was recently distinguished (Saji *et al.*, 1999). A dipole mode in the Indian Ocean, with anomalously low SSTs off the coast of Indonesia and high SSTs in the western Indian Ocean is associated with wind and precipitation anomalies across the Indian Ocean basin. During ENSO events, basin-scale SST anomalies cover the tropical Indian Ocean, with this mode explaining up to 30% of the total SST variability. The dipole mode appears independent of ENSO and accounts for about 12% of the variability in this region (Saji *et al.*, 1999).

The positive phase of the IOD leads to decreased rainfall over the southeastern Indian Ocean (being linked to drought patterns across Indonesia) and increases in rainfall over the western Indian Ocean including tropical eastern Africa. Various studies have linked changes in Indian Ocean SSTs to east African rainfall variability (eg Landman and Mason, 1999), and the results of the Saji and Yamagata (2003) study indicated that the IOD accounted for over 40% of rainfall variability across parts of eastern Africa. The negative dipole mode is associated with a reversal of these anomalous conditions, with warmer waters around the Indonesian archipelago and a large pool of cooler water across the southeastern Indian Ocean, as represented in Figure 2.3. Negative dipole events are correlated with rainfall events across Australia, being associated with the development of northwestern cloud bands (Kuhnel, 1990) which are the principle source of rain to the dry centre and to the southeast of this country.

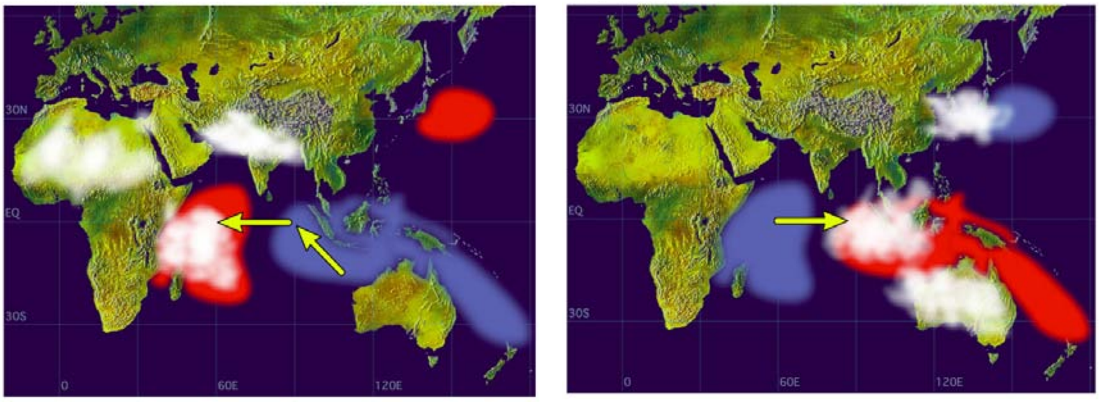

Figure 2.3 Schematic diagrams of the positive phase (on left) and negative phase (on right) of the Indian Ocean dipole (after Rao, 2005)

This section has described a number of global circulation phenomena that produce characteristic spatio-temporal signatures on the Australian climate. The persistence of these climate modes produces clear patterns of variability within time series of hydrologic observations. The simulation of hydrologic series should therefore incorporate similar climatic shifting patterns.

2.2 Measuring changes in the Pacific climate

The identification of coherent patterns of hydroclimatic persistence is dependent upon the computation of climatic variability. Climate proxies such as atmospheric pressures and sea surface temperatures are a means to quantify temporal climate changes. These indices reveal modes of persistence in the broader climate, and the most significant of these in Australia are described in this section.

2.2.1 ENSO indices

Physical changes in ENSO can be monitored through variables that are associated with ocean-atmospheric interactions across the tropical Pacific. One of the more commonly-used measures of ENSO variability is the Southern Oscillation Index (SOI), which quantifies east-west differences in sea-level pressure across the Pacific, in turn signifying variations in equatorial winds. The Troup SOI (Troup, 1965) standardises monthly differences in this atmospheric pressure gradient between Tahiti and Darwin by long-term values for each month. El Niño events are characterised by extended periods of negative SOI values, with extended positive values related to La Niñas. Although the SOI is widely-used to define ENSO events, the fact that it is derived from observations at only two stations leaves this index susceptible to numerous small-scale and high frequency atmospheric phenomena that influence pressures yet do not reflect the Southern Oscillation (Trenberth, 1997).

Sea surface temperature changes in the eastern equatorial Pacific provide an alternative method for detecting ENSO variability. Indeed, SST-based indices are seen as a more direct measure of the temperature anomaly related to El Niño events (Kiem and Franks, 2001), although there is some uncertainty as to the regions of the Pacific Ocean that provide the most relevant information. One of the more widely used SST-based indices is the NINO3 index of temperature anomalies in the region between 5^{o}S to 5^{o}N in latitude and 90^{o}W to 150^{o}W in longitude, termed the “Niño-3” region. SST anomalies in this region are the major ENSO-related quantity to be predicted by models verified with observed data (Trenberth, 1997), with warm SSTs in this region indicating El Niño events.

A more recently developed indicator is the Multivariate ENSO Index (MEI: Wolter and Timlin, 1993), which uses data from a number of different variables and is therefore able to better reflect the complex nature of the ocean-atmospheric interactions involved (Kiem and Franks, 2001). The variables used in its calculations include sea-level pressure, sea surface temperatures, zonal and meridional components of the surface wind, surface air temperature and total cloudiness fraction of the sky. After spatially filtering of these variables into clusters, the MEI is then calculated through Principal Component analysis. All seasonal values are standardised to the 1950-1953 reference period. Extended positive values of the MEI indicate El Niño periods, similar to SST-based indices, however reliable values for this index are only available from 1950 onwards. The quasi-periodicity of ENSO is demonstrated in Figure 2.4, the spectral density for the time series of monthly MEI. A major peak is shown to extend over a periodicity of between 2 to 4 years, highlighting the dominant frequency of ENSO events in the Pacific, although error bounds are large due to the relatively short length of this record.

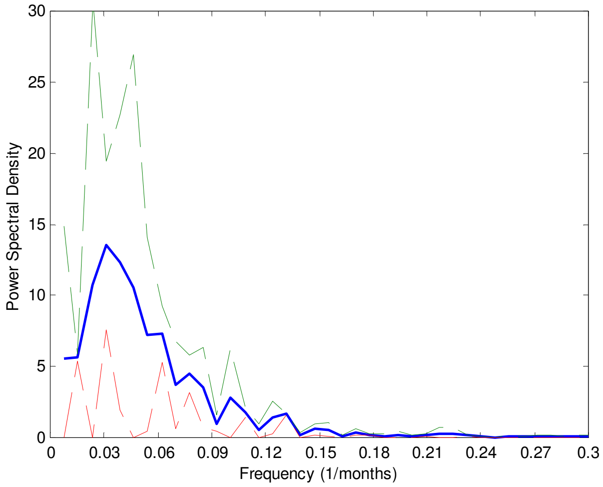

Figure 2.4 Power spectrum for monthly MEI totals (solid line), with 95% confidence limits (dashed lines), using averaged periodogram method

Although El Niño events are generally characterised by positive SST anomalies (and higher atmospheric pressures) in the eastern equatorial Pacific, there is much conjecture over the most suitable definition for individual ENSO events. Some scientists restrict this to the original coastal phenomenon of South America, whereas others extend a definition to include Pacific-wide characteristics (Trenberth, 1997). Glantz (1996) attempted to define El Niño in the manner of a dictionary description, however this lacked a quantitative basis and is therefore inadequate for use in calculations. The complex mechanics of ENSO clearly impede the development of a single definition of events, although classification methods based on changes in ENSO anomalies have been advocated from a variety of sources. The method of Ropelewski and Halpert (1996) uses five-month running-means (5-mrm) of monthly SOI values, defining El Niño events as any year when this 5-mrm remains below -0.5 standard deviations for a duration of 5 months or longer. La Niña events were defined as periods of 5 month duration when the 5-mrm remains above 0.5 standard deviations. Rather than using running-means of monthly SOI values, Chiew *et al.* (1998) employed 12-month average values of this index, taken over the April-March period. ENSO classifications by these authors were then based on average monthly

SOI values over annual periods either above 5 or below -5. By stratifying rainfall and runoff time series across Australia based on ENSO phases as defined with various indices, Kiem and Franks (2001) showed that the MEI out-performed both the SOI and NINO3 indices for discriminating ENSO effects.

The Japan Meteorological Agency (JMA) produced a working definition of ENSO events that was based on SST anomalies in the Niño-3 region. By comparing individual ENSO events obtained from the JMA definition with those obtained from other classification methods, Trenberth (1997) suggests SST anomalies in the "Niño-3.4" region (5°N-5°S, 120°-170°W), provide a superior description. Under this classification, periods during which the 5-mrm of monthly SST anomalies are at least +0.4°C for at least six consecutive months are El Niño events, with La Niñas defined with a threshold of –0.4°C for the same duration. The 5-mrm of SST anomalies are used in order to smooth out the possible intra-seasonal variation of the tropical ocean (Trenberth, 1997). A time series plot of 5-mrm values of Niño-3.4 SST anomalies since 1950 is shown in Figure 2.5, with values exceeding thresholds of ±0.4°C shaded. Sequences exceeding six months in duration are then defined as El Niño and La Niña events.

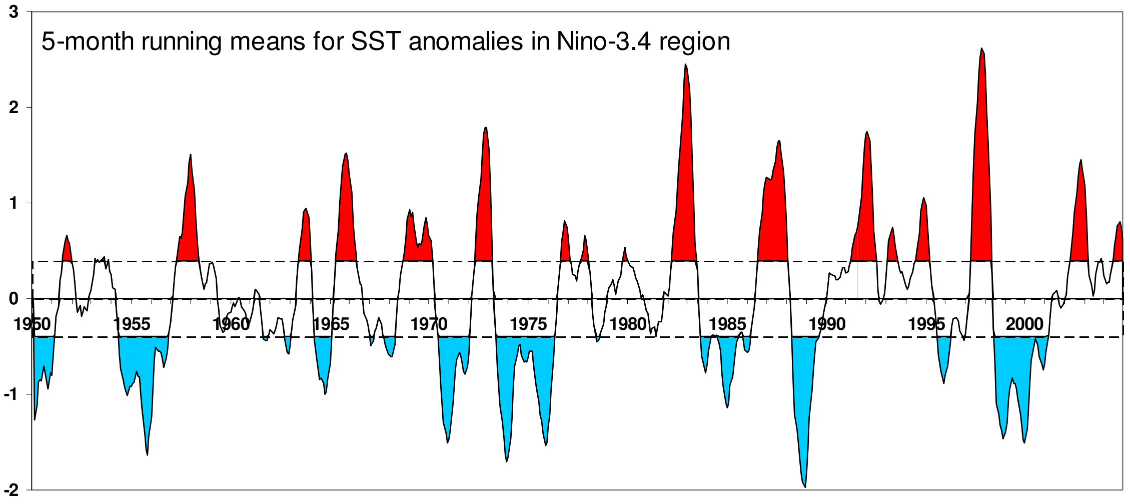

Figure 2.5 Time series of five-month running means of monthly Niño-3.4 values

Using this quantitative assignment of ENSO events, it is apparent that over the time period shown in Figure 2.5 (658 months), there are 178 months classified as El Niño (27%) and 193 months classified as La Niña (29%). Consequently, for 56% of the months on record, either El Niño or La Niña conditions occur. Using an extended period of SST data from 1856-2004 (1786 months), the same classification procedure identifies 488 El Niño months (27%) and 509 La Niña months (29%). This relates to 34 El Niño events and 31 La Niñas, therefore a total of 65 ENSO events, leading to a mean event length of approximately 15 months.

The Indian Ocean dipole is characterised by a strong reversal in sign for basin-wide SST anomalies. This sign reversal forms the basis for a simple time series index that describes the temperature difference between the tropical western Indian Ocean (50°E-70°E, 10°S-10°N) and the tropical southeastern Indian Ocean (90°E-110°E, 10°S-Equator). This series is available over the period 1869-2002, and explains 36% of the variance of rainfall across southwestern Australia (Saji and Yamagata, 2003). Although correlations between the dipole mode index (DMI) and monthly ENSO indices are significant, ENSO explains less than 25% of the variance of the former series. A significant proportion of IOD events have occurred independently from ENSO, with a significant proportion of ENSO episodes occurring in the absence of IOD events (Saji and Yamagata, 2003). To illustrate this independence, the significant dipole mode events of 1961, 1967 and 1994 coincided with an ENSO neutral, a La Niña and a weak El Niño event respectively. Following this, it is likely that the IOD is generated by ocean-atmospheric interactions inherent to the Indian Ocean, and correlation between ENSO and IOD represents mutual interaction with neither climatic mode being the dominant.

The analysis of fluctuating El Niño and La Niña episodes, together with IOD phases, demonstrates strong persistence within the dominant modes of interannual climatic variability. In Chapter 4, ENSO is shown to produce strong hydrologic responses across Australia. These interactions are therefore a clear mechanism for persistence in the broader climate to produce persistent characteristics within time series of hydrologic totals. Moreover, these hydroclimatic interactions underlie the requirement for stochastic models for such series to have capabilities for replicating such persistence.

2.2.2 Quantifying multi-decadal Pacific variability

Section 2.1 outlined evidence for possible climate shifts in the Pacific climate at multi-decadal intervals. Using Pacific-wide SST data, Pierce *et al.* (2000) showed that low-frequency variability was not confined to the tropics in the manner of the higher-frequency ENSO, rather it has significant expression in the mid-latitudes. This point is important as it suggests that this source of variability may influence regional climates around the entire Pacific. Power *et al.* (1999a) used the term *Interdecadal Pacific Oscillation* (IPO) to describe the coherent pattern of Pacific SST variability acting over decadal to interdecadal periods, although Mantua *et al.* (1997) had used the term *Pacific Decadal Oscillation* (PDO) for the same feature. Rial *et al.* (2004) described the IPO as an atmosphere-ocean phenomenon associated with persistent, bimodal climate patterns in the Pacific Ocean.

Indices for inter-decadal variability in the Pacific climate regime have been derived from a variety of sources (eg Mantua *et al.*, 1997; Zhang *et al.*, 1997; Folland *et al.*, 1998), using Principal Component (PC) analysis of various SST data sets. The indices are clearly related to

ENSO signatures, both temporally and spatially, such that the PDO and IPO can be viewed as "ENSO-like" interdecadal climate variability. The PDO index of Mantua *et al.* (1997) is calculated as the leading PC of monthly mean SST data for the tropical and northern Hemisphere extra-tropical regions. This spatial bias is revealed with its signature being more clearly observed in the wintertime surface climate record of North America than any other continent (Latif and Barnett, 1994). The PDO index is similar to an index calculated by Zhang *et al.* (1997) from annual mean anomalies of unfiltered global SST data. The IPO index of Folland *et al.* (1998) is obtained from near-global data sets of seasonal SSTs, using a low-pass filter with half-power at a period of 13.3 years, which produces a smoothed time series. This filter scale eliminated the variability of not only the ENSO signal, but also a decadal-scale mode of variability in the North and tropical Atlantic (Folland *et al.*, 1998). This latter index is taken as being representative of the IPO throughout the remainder of this work, as it provide continuous seasonal data extending to 1857 and is furthermore likely to represent a Pacific-wide manifestation of Mantua's PDO index (Houghton *et al.*, 2001). Folland *et al.* (1998) also provide a multi-decadal index derived from monthly air temperature anomalies, however this is likely to be less accurate and susceptible to localised variations than SST-derived indices (Power *et al.*, 1999a).

The four indices discussed here are significantly correlated, although seasonal averages of IPO have linear correlation of only 0.57 with seasonal averages of the monthly PDO index. Cross-correlations are not as strong at any other time lags. The interdecadal component of variability is common amongst these three indices, and since they have different geographical extents it is likely that they represent true climate variability (Power *et al.*, 1999a).

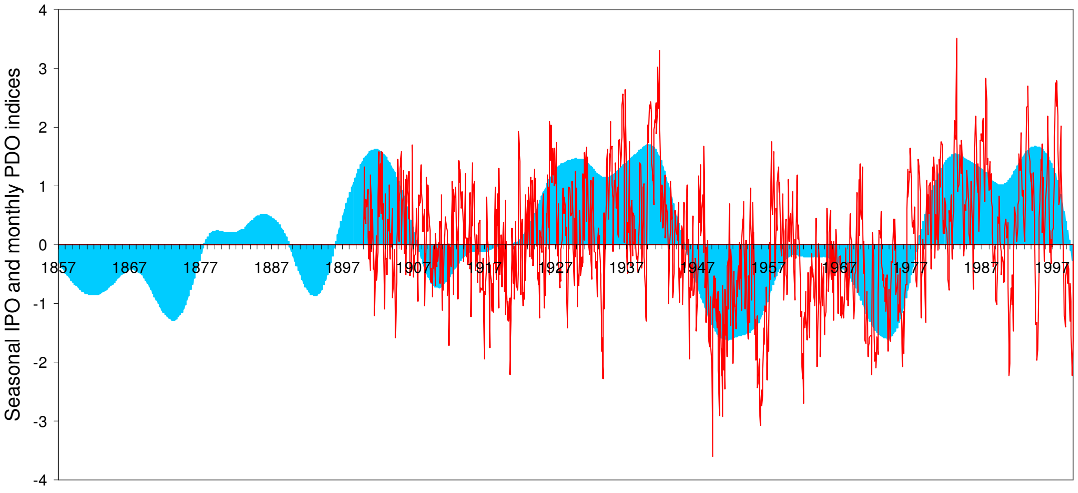

Figure 2.6 Time series of seasonal IPO values (blue) and monthly PDO values (red)

Seasonal values of the IPO index are displayed in Figure 2.6, showing the IPO to experience three major phases during the 20th century: positive phases between 1922-1944 and 1978-1998 and a negative phase between 1945 and 1977. The unfiltered monthly PDO series is shown alongside the IPO values, and it is clear that the major phases visible in the latter series are not as apparent in the former, although there is a clear tendency for negative PDO values during the IPO negative phase between 1945 and 1977.

When the IPO is in a positive phase, temperature anomalies in the tropical Pacific are positive, whilst those both near New Zealand and over the North Pacific are negative (Salinger *et al.*, 2001). ENSO indices are correlated with the IPO (and therefore PDO), such that El Niño conditions tend to coincide with years of positive polarity (Mantua *et al.*, 1997). It is important to note that the transition from negative to positive phase of the IPO occurred in the mid-1970s, the same period noted for dramatic changes in various Pacific climate indicators. The time series of IPO represents a shifting low frequency pattern in Pacific SSTs, such that its fluctuations reflect persistence inherent in the broader climate.

Hydrologic responses to oscillations in both the IPO and ENSO are discussed in Section 2.3.3, and although the former series identifies climatic persistence over a multi-decadal time scale, it is unlikely that persistence at a similar frequency will be observe in time series of monthly rainfall and streamflow totals. A more likely scenario is that the IPO modulates the frequency of ENSO oscillations, in the process influencing the frequency and magnitude of monthly hydrological persistence.

2.2.3 Stable climate states

Climate indices quantify fluctuations in global circulation phenomena, and provide a means to evaluate hydrologic responses to these fluctuations. The indices detailed here, which reflect different aspects of the complex ocean-atmosphere interactions operating on a global scale, show oscillations between different phases with characteristic periodicities. Climatic changes of varying magnitude can be observed on a variety of time scales. For instance, the low frequency variability identified through the IPO and PDO indices suggests that the Pacific climate switches repeatedly on multi-decadal scales to seemingly different climate modes. It is therefore important to investigate whether such changes reflect an underlying forcing mechanism in the atmosphere to oscillate between stable climate states. Time series of hydrologic variables may respond in a similar manner if the characteristics of these states reflect changes in rainfall-generating mechanisms.

An investigation such as this can be approached though monitoring the tendency for the atmosphere to exist within persistent modes (“states”) of circulation. From an atmospheric perspective, stable circulation patterns are generally regarded as quasi-stationary on a large scale, superimposed upon which are various smaller (localised) disturbances (eg Charney and DeVore, 1979). Although such a description enables qualitative explanations for many observed features of the atmosphere, it cannot explain the existence of persistent or recurrent regional weather patterns. Baur (1951) raised the notion many decades ago that persistent or recurrent regional weather patterns (which he termed *grosswetter*) should be the primary focus of long-range weather forecasting. Lorenz (1963) showed theoretical results that supported the views of Baur, by suggesting that atmospheric circulations may contain more than one equilibrium flow regime. Charney and DeVore (1979) followed this through evidence of multiple stationary states in atmospheric circulation regimes. Hansen and Sutera (1995) investigated the amplitude of planetary-scale atmospheric pressures, and identified two modes in the probability density function (pdf) of such amplitudes, where a minimum in the pdf implies a source of instability. Therefore this bimodality, corroborated with simulations from an extended general circulation model (GCM), indicates a tendency for the atmosphere to exist in at least two stable regions.

The supposition of multiple stable regimes in the climate can also be tested through the identification of relatively rapid and sharp transitions in various indices. Schwing *et al.* (2003) note that the atmospheric teleconnection between the North Atlantic and North Pacific appears to have multiple modes, and oscillations in the standing wave patterns of this teleconnection would then lead to abrupt changes in indices such as the NAO and PDO. Christiansen (2003) presented further evidence for two regimes in the interannual variability of the stratosphere, and furthermore an abrupt shift between these to occur in the mid-to-late 1970s. The statistically significant regime transition in the atmospheric field lends support to the notion of changes in IPO and PDO indices reflecting major climatic shifts. Christiansen (2003) also notes that this regime shift is coincident with a change in the relationship between the atmosphere and solar radiation, the primary forcing mechanism for the climate system.

These observations substantiate the assumption that the hydrological cycle is influenced by oscillations between stable regimes, as suggested by the action of persistent climate processes described earlier. The following section demonstrates the impact of this persistence upon rainfall and streamflow observations.

2.3 Hydrological impact of climatic persistence

Over recent decades, many studies have attempted to describe hydrologic variability, both within catchments and across regions, in terms of climate anomalies (Arnell, 2002). Global climate anomalies are characterised by the recurrence of distinct precipitation, temperature and atmospheric pressure patterns. ENSO is the most prominent mode of climatic variability, affecting weather patterns around the world due to its displacement of various significant atmospheric features, such as the speed and direction of ocean currents and the winds that drive them, and the surface temperatures of the ocean. From a hydrological perspective, the cumulative effects of these changes can be extensive. The quasi-periodicity of the ENSO phenomenon impacts upon the seasonal patterns of rainfall and temperature across many regions of the world, leading to a tendency for above and below average values over extended periods. This natural source of persistence links climate anomalies to observed hydrologic variability, and has important implications for the management of water supplies.

2.3.1 Relationships between ENSO and rainfall variability

Relationships between the SO and precipitation were first explored early last century, beginning with the seminal work of Sir Gilbert Walker (1923; 1924), who linked global pressure anomalies to Indian monsoons. Although each warm ENSO episode is distinctive, precipitation and temperature anomalies appear to characterise all El Niño events (Viles and Goudie, 2003), and teleconnections between ENSO and regional climates are the basis for long-range forecasts of rainfall and streamflow across many parts of the world (Chiew *et al.*, 1998). Nicholls (1988) showed that areas influenced by ENSO displayed substantially higher variability in annual rainfall, and Linacre and Geerts (1997) noted that the role of ENSO in modulating Australian rainfall variability may be at least as important as its role in moderating the rainfall mean.

Ropelewski and Halpert (1987) documented large-scale patterns of above- and below-average precipitation patterns being associated with both phases of the ENSO. These results were expanded upon in a later study (Ropelewski and Halpert, 1996), in which shifts in the distribution of rainfall associated with extremes of the SO were identified in various parts of the world. The most extreme case of such a shift was found in the central Pacific, where the 90th percentile of seasonal spatially-averaged rainfall during high SOI periods was far less than the 10th percentile during low SOI episodes. Warm ENSO episodes lead to an increase in rainfall in the tropical eastern Pacific, although reduced rainfall is found to the east and west of this equatorial wet anomaly (Diaz *et al.*, 2001). Within the 19 regions studied by Ropelewski and Halpert (1996), El Niño periods were also related to increased seasonal precipitation in parts of southwestern North America, including the Great Basin region and the Gulf of Mexico, parts of southern India and Sri Lanka, and parts of South America, including Peru and Chile. Conversely, warm ENSO extremes are related to anomalously dry conditions over southeastern Asia and northern and eastern Australia. The variable influence of ENSO on African precipitation (Lindesay, 1988; Mason, 2001) has been extensively documented, with El Niño events being associated with reduced rainfall in the wet season (May-Sept) of tropical Africa, and enhanced rainfall during the rainy season (Oct-Feb) south of the Equator. El Niño conditions suppress the development of tropical storms and hurricanes in the Atlantic, yet increase the numbers of tropical storms over the eastern and central Pacific Ocean (Viles and Goudie, 2003).

The strong relationship between the SO and precipitation across Australia has been the focus of several studies (eg McBride and Nicholls, 1983; Kiladis and Diaz, 1989; Allan, 1991; Ropelewski and Halpert, 1996). El Niño episodes tend to be accompanied by reduced rainfall across the interior of eastern Australia, especially during the winter, whereas La Niña episodes are generally associated with periods of higher rainfall. Chiew and McMahon (2003) used monthly rainfall and streamflow data over a period of 98 years from 284 catchments across Australia to demonstrate that El Niño generally leads to dry conditions across Australia in the latter half of calendar years.

Figure 2.5 indicated a high degree of variability in the magnitude of the SST anomalies associated with individual ENSO events. Furthermore, the teleconnection patterns between ENSO events and regional climate variability also display a highly irregular behaviour. This irregularity can be illustrated by comparing the impact of the two most recent El Niño events (disregarding the period of warm SSTs occurring at the end of 2004) in 2002/03 and 1997/98. Although time series of both SST anomalies and SOI values show the more recent El Niño event to be much weaker than the 1997/98 event, these two events had markedly different effects on the Australian climate. The latter El Niño had a very strong impact across the country, leading to rainfall deficiencies over the period from March 2002 to January 2003, with severe drought conditions affecting almost all of Australia. These deficiencies are demonstrated in Figure 2.7, produced by the Australian Bureau of Meteorology, which shows accumulated rainfall for the 2-year period ending January 2003 to be amongst the lowest on record. In contrast to this, the considerable increases in tropical SSTs and associated changes in atmospheric pressure during the earlier event failed to translate into similarly dry conditions.

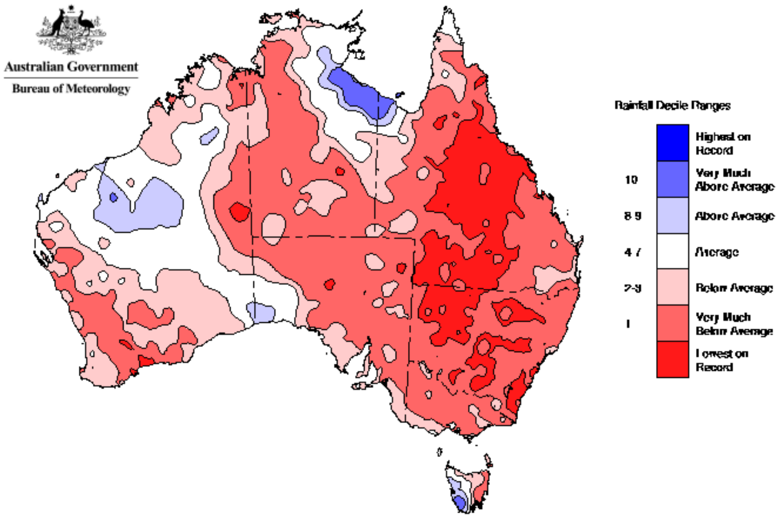

Figure 2.7 Rainfall deficiencies across Australia during 2002/3

The mild impact of the 1997/98 El Niño episode was not confined to Australia though, with a number of other regions around the globe, in which climate variability is well correlated to

ENSO events, experiencing climate conditions that could not be considered typical (Kane, 1999). Dry conditions across southern Africa experienced during earlier El Niño events were not experienced at this time, and the typically wet conditions brought by El Niño to Peru remained from this event throughout the following (dry) La Niña period of 1998/99.

Table 2.1 Comparison of the strength of the most recent 10 El Niño periods and their impacts upon the hydrology of Australia (source: Bureau of Meteorology)

El Niño event	Strength as measured by SOI and SSTs	Overall impact on Australia
2002/2003	**SOI:** Weak, **SST:** Weak to Moderate	Very Strong
1997/1998	**SOI:** Strong, **SST:** Very Strong	Weak
1994/1995	**SOI:** Strong, **SST:** Weak to Moderate	Strong
1991/1992	**SOI and SST:** Moderate to Strong	Strong
1987/1988	**SOI:** Weak, **SST:** Weak to Moderate	Weak
1982/1983	**SOI and SST:** Very Strong	Very Strong
1977/1978	**SOI:** Moderate, **SST:** Weak	Moderate to Strong
1972/1973	**SOI:** Moderate, **SST:** Moderate to Strong	Strong
1969/1970	**SOI and SST:** Weak	Weak
1965/1966	**SOI:** Moderate to Strong, **SST:** Moderate	Moderate

The irregular impacts upon the Australian climate shown in the two most recent El Niño episodes are compared with the 10 most recent episodes in Table 2.1. Summarising detailed analyses of the Bureau of Meteorology, this comparison provides an overall impact of each El Niño upon reductions in rainfall across Australia. These analyses show that the disparities between changes in climate indices and the impacts upon Australian rainfall for the 2002/2003 and 1997/1998 El Niños were more extreme than any of the preceding eight episodes. In seeking to explain observations such as these, Mason and Goddard (2001) assert that even if the magnitude and structure of SST anomalies were identical from one climate event to the next, the general problem of inherent unpredictability within the atmosphere would lead to differences in the climate anomalies observed during each event. This notion underlies the importance of a stochastic approach that incorporates oscillations within hydrologic time series without reliance upon climate index values. Further analysis on this topic is presented in Section 3.2.

2.3.2 Relationships between ENSO and streamflow variability

Rainfall is the dominant factor in the hydrologic cycle, and year-to-year variation in its spatial distribution is an important characteristic of hydrologic regimes. Catchment runoff is an effective measure of the spatial variability in rainfall, with the intrinsic “memory” of watershed storage amplifying anomalous changes in precipitation input. Consequently, streamflow records can more clearly detect high and low cycles that are related to modes of climatic variability such as ENSO as opposed to rainfall records. Hydrologic responses to ENSO events may be delayed due to the innate lag between regional climates and variability in the Pacific Ocean, where El Niño and La Niña events are recorded, or simply from the hydrological response to climate anomalies. For instance, in regions where precipitation is stored as snow, winter climate anomalies can manifest themselves in the streamflows of spring and summer (Arnell, 2002).

Many studies have focused upon establishing links between ENSO and streamflow across the globe, identifying large changes between the hydrologic conditions of El Niño and La Niña periods. Viles and Goudie (2003) noted that the effect of ENSO on streamflow is generally amplified over that on precipitation, and showed the recurrence intervals of discharge events in areas of southwestern North America and Chile to differ greatly between ENSO extremes. An analysis of streamflows in these same regions by Cayan *et al.* (1999) supports these effects, with several basins being at least ten times as likely to experience extremely high flows during El Niño periods than during La Niñas.

As with precipitation, ENSO is strongly linked to streamflow across much of Australia, with ENSO indices explaining over 20% of annual variability in the records from various southeastern rivers (Simpson *et al.*, 1993). Chiew and McMahon (2003) identified clear El Niño-streamflow teleconnections across most of Australia, and these were almost always stronger than El Niño-rainfall teleconnections. The inter-annual variability in Australian streamflow is amongst the highest in the world (McMahon *et al.*, 1992), and ENSO-streamflow and ENSO-rainfall teleconnections are also stronger in this country than in other parts of the world (Chiew and McMahon, 2002b; 2003). In particular, streamflow records from semi-arid regions can clearly show the influence of irregular atmospheric circulation phenomena (Puckridge *et al.*, 2000). Hydrographs of dryland rivers are strongly influenced by aseasonal factors (Walker *et al.*, 1995), with rainfall in these regions tending to be intense and very localised. The effects of ENSO are particularly evident in the Lake Eyre Basin of arid central Australia (Kotwicki and Allan, 1998), and an investigation of the flow regime for Cooper Creek within this basin by Puckridge *et al.* (2000) showed a tendency for important flood clusters to correspond to La Niña periods.

Arnell (2002) provides a concise summary of the published associations between ENSO and streamflow across the globe, which predictably reflect the spatial influence of ENSO on precipitation. During El Niño events, streamflow tends to be higher than average across much of South America and the southwest areas of the USA, with lower than average flows observed across Australasia, central America, and the equatorial zone of Africa. The influence of periodic climate phenomena on streamflow totals has important effects upon the management of water supplies, with ENSO events being associated with extended periods of high and low levels in water storages. Figure 2.8 summarises the location of global streamflow anomalies that are due to ENSO, with the shaded areas indicating lower streamflows being received during El Niño episodes, and areas that receive higher streamflows outlined.

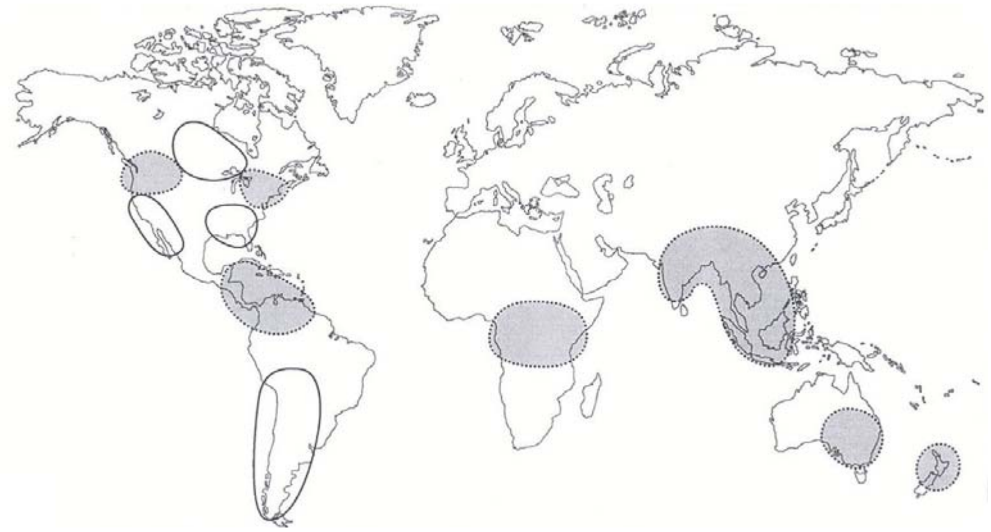

Figure 2.8 Global streamflow anomalies during El Niño (after Viles and Goudie, 2003)

The most relevant outcome from these analyses is that climate oscillations encapsulated by the ENSO phenomenon provide a direct source of persistence within hydrologic data. This hydroclimatic interaction is stronger within streamflow time series than with rainfall observations, having a broad range in the strength of its modulations. The analysis of hydrologic totals in regions that are modulated strongly by ENSO phases are likely to benefit from a modelling approach that provides an explicit description of hydroclimatic persistence.

2.3.3 Low frequency modulation of ENSO impacts

With a broad range of ocean-atmospheric phenomena present in the climate regime of the Pacific, it is important to distinguish their cumulative effect upon regional climates. ENSO and IPO signatures are both revealed through time series of Pacific SST anomalies, and it is likely that the impact of individual ENSO events upon regional climates will be modulated by the longer period IPO.

Observed teleconnections between ENSO and regional climates around the Pacific are complicated by large inter-El Niño or inter-La Niña variability (Kumar and Hoerling, 1997; Gershunov and Barnett, 1998), due to changes in the IPO. Although the influence of ENSO is significant, variability associated with this generally explains less than 25% of interannual variations in global mean temperature and rainfall (Diaz *et al.*, 2001), attesting to the strong role played by other modes of variability. The strength of teleconnections between precipitation and ENSO increases with the amplitude of Pacific SST anomalies (Kumar and Hoerling, 1998), with stronger teleconnections expected to occur during periods when tropical Pacific SSTs display larger interannual variance. Diaz *et al.* (2001) demonstrates the influence of the mid-1970s climate change on ENSO teleconnections, by showing that as much as 50% of the variability in

North Pacific atmospheric circulation was driven by ENSO after 1977, as opposed to only 15% for the period 1948-1977.

Gershunov and Barnett (1998) investigated the influence of the coherent longer-period mode of climate variability on North American ENSO teleconnections. Using the term North Pacific Oscillation (NPO) to describe the same low frequency effects that were defined as the PDO by Mantua *et al.* (1997), these authors presented evidence for El Niño signals to be strong and stable during high NPO phases, when North Pacific SSTs are cold. This suggests that the more slowly-evolving mode of variability can modulate ENSO-related climate predictability in North America. In light of the widespread use of ENSO-based precipitation forecasts, the decadal state of the climate is the best indicator of whether ENSO teleconnections can be used as reliable predictors of seasonal precipitation (McCabe and Dettinger, 1999).

Interdecadal periods of warm SST anomalies in the tropical Pacific are associated with a breakdown of interannual teleconnections between ENSO and rainfall in various other areas surrounding the Pacific (Arblaster *et al.*, 2002). Power *et al.* (1999a) showed that the two phases of the IPO strongly modulate ENSO-precipitation signals across Australia, with a larger fraction of climate variability related to ENSO evident in periods of negative IPO regimes. These authors used four important climate-related variables to assess interdecadal variability in the influence of ENSO on Australia: continent-averaged rainfall and temperature data, reconstructed natural flows in the River Murray (with effects of irrigation, land changes, etc. removed) and an estimate of the domestic wheat crop yield. In order to observe changes in the relationship between these variables and ENSO, correlations with SOI were stratified by the polarity of the IPO index, using thresholds of IPO < -0.5 and IPO > +0.5. In periods of IPO < -0.5, correlations between the SOI and each of the four variables were significant yet none of these were significant in periods where the polarity of the IPO was reversed.

Power *et al.* (1999a) demonstrated that when correlations between annual SOI values and annual values of the four climate variables are taken over 13-year moving windows, the time series of correlations are strongly related to annual IPO values. This result further suggests that the IPO relates to the low-frequency variability in ENSO-climate relationships across Australia. In an earlier study, Nicholls *et al.* (1996) found that the relationship between the SOI and Australian rainfall had diminished since the mid-1970s, corresponding to the start of a positive IPO epoch. This demonstrates that individual ENSO events have stronger impact upon Australia during the negative phase of the IPO. Furthermore, Kiem *et al.* (2003) showed the negative phase of the IPO to be biased towards an increased frequency of La Niña episodes. In light of such observations, one approach to the analysis of hydrologic records in regions influenced by multi-decadal climatic persistence is to treat each IPO epoch separately. Alternatively stochastic models that are formulated around the interaction of hydroclimatic modulations, at various time scales, offer a suitable approach.

There is considerable uncertainty surrounding the relationship between inter-annual and multi-decadal modes of Pacific climate variability. The results of Power *et al.* (1999b) suggest that the phases of the IPO tend to modulate the rate of occurrence and magnitude of El Niño and La Niña events. Meehl *et al.* (2001) also presented evidence for a direct modulation of ENSO characteristics by the IPO, showing the importance of a sharp thermocline structure in the eastern tropical Pacific (in the Niño-3 region) for capturing realistic ENSO variability in coupled models. Changes in this thermocline due to low frequency variability in ocean dynamics could affect the up-welling of deeper waters in this region, which in turn can alter the amplitude of ENSO variability.

A second explanation for this low-frequency modulation, proposed by Arblaster *et al.* (2002), involves IPO modulating the mechanisms by which ENSO is communicated to regional climates. This can be explained through the accepted "delayed action oscillator" model for ENSO, which involves the internal interaction of oceanic and atmospheric processes in the eastern equatorial Pacific (Franks, 2004). Interannual variability of ENSO arises through anomalous perturbations in this coupled system, which can then propagate via positive feedback in equatorial (Walker Cell) circulations. These subsequently interfere with the location of the Inter Tropical Convergence Zone (ITCZ) and the South Pacific Convergence Zone (SPCZ), the latter being one of the most important climatic features in the subtropical Southern Hemisphere by delivering rain-bearing cloud bands to eastern and central Australia.

Small shifts in the location of the SPCZ can produce large rainfall anomalies (Salinger *et al.*, 2001), and alterations in the latitude of the SPCZ due to ENSO events in the Austral summer are an obvious mechanism by which ENSO alters rainfall patterns across the Australian continent. Warm El Niño events disrupt the propagation of the SPCZ over southern latitudes, causing reduced rainfall across the Australian continent. Cold La Niña periods however produce an enhanced southern movement of the zone, resulting in higher rainfall across Australia. Kumar *et al.* (1999) determined a southerly shift in the Walker circulation in the recent two decades to be the cause of a weakening in the relationship between ENSO and the Indian monsoon. Although these authors noted such a shift could be due to global warming, natural variability on an interdecadal time scale could also be related. Folland *et al.* (2002) reinforced this latter point by showing that alterations in the latitude of the SPCZ are also significantly related to multi-decadal IPO variability, independent of ENSO influences. During La Niña events that occur in IPO negative epochs, the SPCZ is at its southwestern extreme, thus bringing higher amounts of rainfall to Australia. As Franks (2004) states, this rationalises the results of Power *et al.* (1999a) who showed the enhancement of La Niña conditions in IPO negative phases.

Extreme hydrologic events (eg droughts and floods) in eastern Australia are generally associated with extreme ENSO events, with year-to-year flood and drought risk varying according to ENSO conditions (Franks, 2004). Kiem *et al.* (2003) showed that the IPO modulation of ENSO produced multi-decadal epochs of elevated flood risk, with state-wide flood frequency indices for New South Wales showing much higher flood risk during La Niña events as opposed to El Niño events but also during the negative phase of IPO as opposed to the positive. This insight further demonstrates the role of IPO and ENSO in modulating hydrologic risk, and strengthens suggestions that the climate can fluctuate between stable regimes at a range of time scales.

2.4 Summary of chapter

This chapter has summarised the major global circulation phenomena that influence the climate of Australia. Ocean-atmospheric interactions originating in each of the three major bodies of water surround this continent include persistent cycles that modulate the hydrological cycle. One approach to linking persistence in the climate to hydrological persistence is to analyse relationships between climate indices and hydrologic data. An alternative approach to the analysis of hydroclimatic interactions that underlie hydrological persistence is through the modelling assumptions of HMMs. Their hypothesis of fluctuations between small numbers of generally stable climate states leading to persistence in hydrologic responses is consistent with a substantial body of research. The following chapter investigates the quantification and stochastic modelling of persistence in hydrologic data.

Chapter 3 Hydrological persistence

A number of studies that have attempted to quantify the relationships between large-scale climatic fluctuations and hydrological yield have focused upon the issue of persistence. A thorough review of literature on this topic is presented in this chapter, together with theoretical results concerning statistical interpretations of persistence. Various methods to incorporate this feature into hydrological time series models are also discussed, with particular attention focused upon hidden Markov models (HMMs). These models have a parsimonious structure that is intuitively better for modelling shifting levels in hydrological time series, thus presenting an innovative approach to the modelling of hydrological persistence. Techniques to account for uncertainties caused by both a lack of data and by model calibration methods are critically reviewed, including benefits provided to these models by using a Bayesian approach.

3.1 Estimating persistence in hydrological time series

Hydrologic series display important and characteristic responses to large-scale climate modes, with records from across Australia showing a tendency to cluster either above or below "normal" levels for extended periods of time. Persistence in rainfall or streamflow observations can be attributed to atmospheric, basin or sub-surface storage (Matalas, 1997), and has major implications for the design and management of water supply infrastructure. A standard time series definition of persistence is generally given in terms of a slow decay of the autocorrelation function, modelled as a hyperbolic decay. Throughout this thesis however, *hydrological persistence* is interpreted through the lengths of wet and dry periods, termed *spells*, exceeding a threshold duration that could be expected for non-persistent data (such as white noise). The standard time series definition of persistence is not used in the thesis for this interpretation of persistence.

Studies that have focused upon spell analysis of hydrologic time series (eg Yevjevich, 1967a; Saldarriaga and Yevjevich, 1970; Peel *et al.*, 2004b) have been noted for their usefulness in water resources planning and for investigating the stochastic and deterministic nature of observed data. The characterisation of spells requires the specification of threshold levels that divide a sample into clusters of similar values. The sample mean offers a natural choice for dividing wet values from dry, although the use of the sample median (as used by Peel *et al.*, 2004b) reduces the influence of skewed data on the length and magnitude of runs. A similar threshold is applied in order to undertake runs analysis throughout this work.

3.1.1 Statistical tests to identify persistence

A number of statistical tests are available to estimate the occurrence and extent of persistence in hydrologic records. Some of these tests are designed to investigate the nature of spells, which is generally termed runs analysis, while other tests consider the temporal dependence of persistent series through techniques such as autocorrelation and spectral analysis. In their description of runs analysis, Saldarriaga and Yevjevich (1970) advocated the use of the mean run length of a sample in order to test for persistence in an observed data series. Peel *et al.* (2004a) developed an alternate summary statistic, based on the frequency distribution of run lengths to compare a range of observed streamflow series. The raw skew of the frequency distribution of run lengths in a series of length *T*, where *m(s)* is the frequency of run length *s* and *L* is the length of the longest observed run, is estimated by

$$g = \frac{\sum_{s=1}^{L} m(s) \times s^3}{(T-1)} \tag{3.1}$$

Gold's length of runs test (LORT) (Gold, 1929; Srikanthan *et al.*, 1983) identifies persistent series through comparing the length of runs defined using a median threshold to lengths expected for observations generated from a purely random process. The LORT calculation also uses the frequency of run lengths *m(s)*, with expected value of *m*(*s*) for a random process

$$E[m(s)] = \frac{(T+3-s)}{2^{s+1}} \tag{3.2}$$

and the sum Q,

$$Q = \sum_{s=1}^{L} \frac{\{m(s) - E[m(s)]\}^2}{E[m(s)]} \tag{3.3}$$

is the length of runs test-statistic, distributed as a χ^2 variable with (*L-1*) degrees of freedom. Values of *Q* below the 95th percentile of the corresponding χ^2 variable provide statistically significant evidence for dependence.

The linear dependence between any two observations in a series can be measured through the calculation of sample autocorrelation, with short-term dependence in a time series usually measured by the magnitude of low-order autocorrelation coefficients. Peel *et al.* (2004b) suggest that the significant variable in run length equations, apart from truncation level, is the magnitude of the lag-1 autocorrelation. The correlation between observations a distance *k* apart in a time series $\{y_t\}$, termed the autocorrelation at lag *k* or $r(k)$, can be estimated by:

$$r(k) = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{3.4}$$

where T is the sample size and $\bar{y} = \frac{1}{T}\sum_{i=1}^{T} y_i$.

The use of an autocorrelation function as a description of the temporal dependence in a time series is usually used when the distribution of residuals approximates Gaussian. If conditions of normality and homoscedasticity are not met, the autocorrelation function may be an unreliable measure and should be used with care (Sen, 1978). In order to examine the nature of sequential dependence in hydrologic time series, Sen (1978) developed *autorun analysis* as an alternative methodology to autocorrelation analysis. This technique is based purely upon the theory of wet and dry runs and is a more appealing statistic for identifying persistent processes. By using the variable *m* to describe the threshold separating high (*wet*) variables from low (*dry*), the autorun coefficient at lag *k* is given by

$$r(k) = \frac{P(y_{i-k} > m, y_i > m)}{P(y > m)} \tag{3.5}$$

This can also be expressed in terms of the threshold exceedance probability $p = P(y_i > m)$, such that the median truncation level, for which $p = 0.5$, is a special case. Sen *et al.* (2003) show that a small sample estimate of Eq. 3.5 is given in terms of the number of joint events (n_k), in which observations lag-*k* apart are simultaneously greater than *m*. The sample estimate is given as

$$r(k) = \frac{n_k}{p(T - k)} \tag{3.6}$$

Although correlations are rather insensitive to the conditional distribution of residuals, the calculations involved in autorun analysis are distribution-free such that the autorun coefficient will not distort the dependence structure of the sequence considered (Sen *et al.*, 2003). In the case of a mean threshold applied to a sequence of normally distributed stochastic variables, the autorun coefficient is equivalent to the autocorrelation coefficient.

3.1.2 Hurst phenomenon

In any consideration of persistence in hydrology, it is important to consider the contribution to this field of Hurst, whose seminal papers (1951; 1957) followed an extensive investigation of the long-term characteristics of reservoir storage. Following a study of geophysical records, Hurst *et al.* (1965) described the persistence in these observations in the following manner:

> "the persistency which exists in many natural time series as well as others of social origin is not well described by serial correlations… periods occur when on the whole values are above the long-term average, and others when they are below it, to a greater extent than is to be expected from a series of independent variables."

By using the rescaled range statistic, evaluated as the cumulative sum of departures of annual totals from their mean, Hurst (1951) identified a

> "definite difference between many natural time series and those where the terms are independent of each other".

The evaluation of the rescaled range statistic requires the estimation of the adjusted range $R(t,k)$ for subsets of the time series $\{y_t\}$ that start at t and have size $(k+1)$. The mean value of this subset is

$$\mu(t,k)=\frac{1}{k+1}\sum_{i=t}^{t+k} y_i \qquad \textbf{(3.7)}$$

and the corresponding standard deviation is

$$S(t,k)=\left[\frac{1}{k+1}\sum_{i=t}^{t+k}\left(y_i-\mu(t,k)\right)^2\right]^{\frac{1}{2}} \qquad \textbf{(3.8)}$$

By defining $D(t,k,m)$, where $0 \le m \le k$, as

$$D(t,k,m)=\sum_{i=t}^{t+m} x_i-(m+1)\mu(t,k) \qquad \textbf{(3.9)}$$

the adjusted range $R(t,k)$ is then calculated as the difference between the maximum and minimum values for $D(t,k,m)$. The rescaled range is found as $R(t,k)/S(t,k)$, simplified to R/S, and is repeated for randomly chosen values of t and k. Hurst showed that if the $\{y_t\}$ were purely random and thus drawn without serial dependence, the expected value of the rescaled range for a large data set is given by

$$E\left[\frac{R(t,k)}{S(t,k)}\right]=\left(Ak\right)^{0.5} \qquad \textbf{(3.10)}$$

where A is constant. Exponents of (Ak) significantly exceeding 0.5 are evidence against independence. After taking logarithms, Eq. 3.10 is rewritten as

$$\log E[R/S] \approx A + H \log k \tag{3.11}$$

Hurst *et al.* (1965) analysed many geophysical time series to determine a range of values for the exponent of (Ak), referred hereafter as *H*, that would be expected to occur in nature. The results of this study showed H to have a mean value of 0.72 and standard deviation of 0.09, thus significantly exceeding the theoretical 0.5. The characteristic for natural series to produce $0.5 < H < 1.0$ has been termed the Hurst phenomenon, with H termed the Hurst exponent. In double logarithmic plots of $R(t,k)/S(t,k)$ against k, Hurst behaviour is revealed as a straight line arrangement of points corresponding to the size of different subsets, the slope of which is the Hurst exponent. A range of heuristic methods to estimate the Hurst exponent *H* have been suggested in the literature, including the semi-variogram, a double logarithmic plot of the correlogram and least squares regression in the spectral domain (see Beran, 1994), although the rescaled adjusted range statistic is perhaps the best known. Recently, Abry and Veitch (1998) described a wavelet-based tool that provides a natural, statistically and computationally efficient estimator of the Hurst exponent.

An intuitive explanation for the adjusted range is derived in terms of its original application in reservoir storages, as discussed by Borgman and Amorocho (1970) (cited in Bras and Rodriguez-Iturbe, 1993). Suppose y_t is the total inflow to a reservoir during period t, with water flowing out of the reservoir at a constant rate, being such that the total amount of water in the reservoir at time $(t+k)$ is the same as at time t. $R(t,k)$ is then the minimum capacity of a reservoir such that it will not overflow during this period, calculated as the difference between the largest surplus and greatest deficit. Statistical behaviour of the rescaled range provides insight into the volumes that need to be maintained (Bras and Rodriguez-Iturbe, 1993).

Extensive literature has been devoted to explaining the Hurst phenomenon, which has been of great interest to statisticians, hydrologists and geophysicists in the years since its discovery. A range of possible explanations have been presented, suggesting that it cannot be attributed to one physical cause. One line of thought is that the Hurst phenomenon is a transitory behaviour, and that series of hydrologic observations are not long enough to test the steady-state behaviour of the adjusted range, which will approach 0.5. A second theory stated by Klemes (1974) involves the possibility of low-frequency variability in the underlying mean of the process. The third dominant theory proposed in the literature suggests that the Hurst phenomenon is derived from stationary processes having very large memory, such that their autocorrelation functions decay very slowly in time. Bras and Rodriguez-Iturbe (1993) note that in the limit, such an argument claims infinite memory for natural processes.

The Hurst phenomenon is a puzzling attribute of geophysical time series, and although it describes a natural long-term characteristic, it is not analogous to the interpretation of persistence taken in this work. The characteristic for Hurst exponents exceeding 0.5 is referred to in this work as *mathematical persistence* in order to differentiate it from the characteristic of various hydrologic data that show extended periods in stable regimes, which is termed *hydrological persistence*.

Mathematical persistence, also referred to as long-range dependence or long memory, is revealed through correlations that slowly decay at a hyperbolic rate with $r(k) \approx k^{-\alpha}$ as k tends to infinity, such that the dependence between events at distant time points diminishes slowly as this distance increases. The relationship between the Hurst phenomenon and this hyperbolic decay of autocorrelations is shown through the relationship $H = 1 - \alpha/2$. Beran (1994) notes that the condition of hyperbolic decay of autocorrelations is essentially equivalent to the spectral density having a pole at zero. The characteristic of long-range dependence is in contrast to the short-range dependence shown by ARMA and Markov processes, in which the asymptotic decay of autocorrelations is exponential such that $r(k) \approx \alpha^{k}$.

In a frequency domain, long-range dependence of a series is revealed when its spectral density is unbounded at the origin. The low frequency power spectra of these series display power-law behaviour $f^{-\alpha}$ that is consistent with the time-domain characteristic of slowly-decaying autocorrelations. In frequency-domain models for mathematically persistent time series, this feature is referred to as "$1/f$" noise (Bak *et al.*, 1987), which can be observed over vastly different time scales. It is important to note that mathematical persistence (hence the Hurst phenomenon) is an asymptotic quality, providing an indication of the rate of convergence but not the magnitude of autocorrelations or frequencies. With correlations at specific lags for long-range dependent series not specified, Beran (1994) makes the point that the detection of this feature in finite samples is difficult. Furthermore even though ARMA models display an asymptotic value of $H = 0.5$, finite realisations from such processes can show estimates for the Hurst exponent of $0.5 < H < 1$.

The Hurst phenomenon describes a statistical phenomenon that is not examined closely in the remainder of this thesis. This work is focused upon the identification and estimation of time series displaying short-term dependence and fluctuations between "stable" regimes, rather than a slow decay of autocorrelations (or frequencies) or high values of the Hurst exponent. A range of suitable time series models for hydrological persistence are described in Section 3.2.

3.1.3 Previous studies of persistence in hydrology

A considerable number of hydrological and climatological studies have used runs analysis to evaluate the impact of persistence in hydrologic records. The most common application of this method is in drought studies, which use thresholds to define periods of consecutive dry years that are consistent with drought conditions. Although there are various definitions for drought (eg Dracup *et al.*, 1980; Wilhite, 2000), runs analysis provides a tool for investigating the associated water deficit.

Yevjevich (1963; 1964; 1967b) provided much of the early investigations into the importance of runs analysis in hydrology. Through these early studies, drought severity was defined as the sum of negative deviations from a threshold for given dry spell lengths. As a result, run length and run magnitude become important parameters for investigating the impact of drought conditions. Peel *et al.* (2005) cites many papers that investigate the expected frequency and average recurrence interval of wet and dry periods of certain magnitude and severity, noting that in comparison to run length, run magnitude has received considerably less attention in the hydrologic literature. Peel *et al.* (2004b; 2005) produced extensive analyses into runs of annual precipitation and streamflow on a global scale. These investigations showed run lengths defined by median values to be similar across all continents and Köppen climate zones, expect for arid and tropical North Africa, which tends towards longer (more persistent) runs. Annual run magnitude was closely related to interannual variability in rainfall and streamflow observations, with continental differences consistent with differences in interannual variability.

Importantly, most hydrologic studies using runs analysis have investigated wet and dry fluctuations at annual time scales. There has been little focus upon wet and dry spells at a monthly scale, or of drought impact at sub-annual intervals. This forms a major focus of this current work, which although not following previous investigations in merely identifying hydrological persistence, describes valuable results that demonstrate the importance of runs analysis. The economic and social stresses that accompany drought conditions are a major consequence of hydrological persistence, and with this feature being a significant aspect of global hydrologic regimes, it is vital that it is incorporated into stochastic time series models. The following section examines approaches to model hydrologic series, and also methods to incorporate wet and dry spells into such models.

3.2 Stochastic modelling of hydrological time series

Stochastic modelling plays an important role in hydrology, with data generation and forecasting being used extensively in the planning and management of water resources. Synthetic hydrology is an approach to counter limitations presented by historical hydrologic inputs, through the generation of random hydrologic sequences that retain correct probabilistic behaviour (Bras and Rodriguez-Iturbe, 1993). The generation of synthetic streamflow or rainfall series for use in simulation studies arose through a dissatisfaction with earlier techniques of hydrologic design (Jackson, 1975a). The development of formal stochastic modelling in hydrology began during the 1960s with the application of autoregressive (AR) models to time series of annual streamflows (Salas and Boes, 1980). The earlier parts of this chapter have described persistence within the different aspects of the hydrological cycles, and it is therefore vital that a method to produce accurate simulations of persistent processes is available.

Stochastic models play a major role in testing aspects of physical systems, so the foundations of these models must therefore preserve the overall physical relevance of such systems in order to gain the correct perspective from modelling results. The hydrological cycle is the environmental system describing the distribution and circulation of water in all its forms (Hipel and McLeod, 1994) so when modelling parts of this cycle, such as rainfall or streamflow time series, it is desirable that the key physical characteristics of the entire system are adequately described. Physically founded models can explicitly account for interactions within a watershed, and Salas and Smith (1981) state that it is

> "desirable that… physical considerations be used for aiding in
> the identification of the type of model".

This view is shared by Sen *et al.* (2003), who state that one way of decreasing uncertainty between true and estimated models is by selecting a stochastic model that best represents the physical reality of the system.

The development of alternative models for stochastic hydrologic simulation needs to be made with regard to whether the model has physical and/or operational justification (Salas and Boes, 1980). Jackson (1975a) addresses this question by outlining a distinction between *descriptive* and *prescriptive* hydrologic models; the former being those models that seek to describe a process or system, hoping to provide insight into its "actual operation", with the latter intended only to extract results for planning use. In this way, Jackson (1975a) states that prescriptive models are not intended as "phenomenological" descriptions as they do not profess to accurately model the actual workings of a system, and assumptions made for the model are not necessarily in accordance with real physical phenomena (Salas and Boes, 1980).

The issues surrounding the development of probability models are also confronted by Cox (1990), who introduced the term *substantive* to represent Jackson's *descriptive* models, and *empirical* to represent Jackson's *prescriptive* models. In many ways, the most appealing models are substantive models that connect directly with subject-matter considerations, and attempt to explain the observed system in terms of its mechanisms, often with variables that are not directly observed. While substantive models have a descriptive endeavour, there is no requirement for such models to have a complex structure. To illustrate this, Cox (1990) cites the formulation of simple five-parameter schemes for high frequency rainfall series (see Rodriguez-Iturbe *et al.*, 1988), assuming Poisson cluster processes for individual rain cells. Although such models are highly idealised, they offer a means by which a complex physical process can be represented in a parsimonious manner, using parameters of physical significance. Conversely, empirical models such as multiple regression models are not based on any specific subject-matter considerations, rather aiming to preserve certain characteristics of the original time series. In seeking to develop suitable stochastic models for hydrological persistence, it is desirable that the physical context for this variability is maintained. A range of approaches to modelling hydrologic time series are presented in the following section, with their justification in terms of observed persistence also discussed.

3.2.1 Univariate time series models

The main concern in the early development of stochastic models was the preservation of observed values for the mean, standard deviation, skew and serial correlation coefficients, and to incorporate these into stationary linear time series models (Salas and Boes, 1980). These are termed short-memory (or short-dependence) models as their correlation functions decay rapidly as lags increase. The correlations of long-memory models decay more slowly. The family of stationary linear models that includes autoregressive (AR), moving average (MA) and mixed autoregressive moving average (ARMA) process are intuitively appealing, and have been applied in a range of different fields. The mathematical definitions and properties of these models have been investigated thoroughly by various authors (eg Box and Jenkins, 1970; Chatfield, 1996), and only a brief description is presented here.

Models that specify first-order correlation coefficients, termed autoregressive models of order one or AR(1) models, are perhaps the most common model of time-series simulation and forecasting. The AR(1) process, which is commonly called a *Markov* process, or the *Thomas-Fiering* model in hydrology, can be represented in terms of a zero mean process as

$$y_t = \phi_1 y_{t-1} + z_t \tag{3.12}$$

where z_t is a white noise sequence with zero mean and variance σ_z^2. To ensure stationarity, the serial correlation parameter is required to maintain the condition of $|\phi_1| < 1$. The autocorrelation of the process $\{y_t\}$ at lag k is related to the serial correlation parameter by $r(k) = \phi_1^k$. Furthermore, the variance of the process is related to both the lag-1 autocorrelation and to the variance of the white noise sequence by

$$\sigma_y^2 = \frac{\sigma_z^2}{(1 - r(1)^2)} \tag{3.13}$$

In hydrologic applications, the AR(1) model is often expressed in terms of the central moments of a random process (Bras and Rodriguez-Iturbe, 1993),

$$(x_t - \mu) = r(1)(x_{t-1} - \mu) + \sigma_x (1 - r(1)^2)^{0.5} w_t \tag{3.14}$$

Here, μ is the stationary mean of the process $\{x_t\}$, and $(x_t - \mu)$ is y_t in Eq. 3.12. The random term z_t is replaced by a term containing w_t, a zero mean normally distributed variable with unit variance, with $\sigma_x (1 - r(1)^2)^{0.5}$ being the standard deviation of $\{z_t\}$. The spectra of AR(1) processes with positive parameters are dominated by low-frequency fluctuations, while negative parameters see high frequencies dominate. The Markov process with a single ϕ parameter is a special case of autoregressive models. A process $\{y_t\}$ is said to be an autoregressive process of order *p*, abbreviated to AR(*p*), if

$$y_t = \varphi_1 y_{t-1} + ... + \varphi_p y_{t-p} + z_t \tag{3.15}$$

A stationary linear process, termed a moving average process, can be described in terms of current white noise terms, as well as one or more previous innovations (Hipel and McLeod, 1994). Given that $\{z_t\}$ is a purely random process with zero mean and variance σ_z^2, a process $\{y_t\}$ is said to be a moving average (MA) process of order *q*, abbreviated to MA(*q*) if

$$y_t = \varphi_1 z_{t-1} + ... + \varphi_q z_{t-q} + z_t \tag{3.16}$$

The MA(*q*) process is assumed stationary regardless of the values of the φ parameters, as y_t is formed from a finite linear combination of the z_t terms. However, by imposing the restriction that $|\varphi_q| < 1$ for all q, invertibility of the process is ensured.

A key principle to follow in time series modelling is to have as few parameters as possible. When fitting stationary linear models, it may be advantageous to combine AR and MA processes to form a mixed autoregressive/ moving average (ARMA) process. The importance of these mixed processes lie in the fact that a stationary time series may often be described by an ARMA model involving fewer parameters than a pure MA or AR process by itself (Chatfield, 1996). An ARMA(*p*, *q*) process containing *p* AR terms and *q* MA terms, is expressed as

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + z_t + \varphi_1 z_{t-1} + ... + \varphi_q z_{t-q} \tag{3.17}$$

It is clear that an ARMA(p, 0) is identical to an AR(p) process, and an ARMA(0, q) process can also be written as MA(q).

Although linear autoregressive models have been presented as physically based models for stochastic hydrology (eg Yevjevich, 1963), their benefit in this context is closer to a prescriptive or empirical sense. These models are intended to preserve only certain characteristics, such as the mean, standard deviation and the correlation structure, and do not attempt to model nature exactly. Notwithstanding this, Fiering (1967) showed through a conceptual watershed representation that by using ARMA models for simulating annual precipitation, a physical basis for describing annual streamflows is produced. By considering this earlier work, Salas and Smith (1981) derived the correlation structure of annual streamflow as a function of the correlation of annual rainfall. If the latter process is assumed independent, or described by an AR(1) model, annual streamflow and groundwater storage will be described by ARMA models. This result is important, as it demonstrates an attempt to relate the stochastic behaviour of different processes of the hydrologic cycle. However, this result remains dependent upon the ARMA framework being a suitable description for precipitation variability at an annual scale, and fails to provide a physical interpretation for such an assumption.

The seemingly abstract nature of standard Markov models to describe hydrologic processes (for example why should total rainfall in a certain year depend only upon the rainfall in the preceding year?) is improved somewhat with a stationary alternative termed the *double Markov* process. Landwehr and Matalas (1986) defined such processes as

$$y_t = Au_t + (1 - A)v_t ; 0 \le A \le 1 \tag{3.18}$$

where u_t and v_t are the outcomes at time t of two independent Markov processes. Although Landwehr and Matalas (1986) interpreted these processes as being the local and regional components that contribute to tree growth, it is possible to adapt such a model to a hydrologic context. In this way, y_t can represent streamflow or rainfall derived from both local climate conditions u_t, and regional controls on atmospheric conditions v_t. Models such as this are a step closer to providing a link between at-site stochastic behaviour to *irregular* climate forcings at a range of scales (Koutsoyiannis, 2004).

Salas and Boes (1980) question the application of ARMA models from an operational rather than conceptual point of view. The inability of such models to reproduce historical drought periods, including the magnitude of *ranges* and *runs*, is considered the central deficiency for hydrologic applications. This view is shared by Bras and Rodriguez-Iturbe (1993), who state that it is quite common for simulated streamflow and precipitation series lacking droughts and floods of the magnitude present in historic sequences. Mandelbrot and Wallis (1968) introduced terms such as "Noah" and "Joseph" effects to describe low flows in historical records. These authors suggested that in order to simulate these effects accurately, either models that consider more durable after-effects (such as multiple lag models) or indeed *self-similar* models (as discussed in the following section), needed to replace ARMA models in stochastic hydrology.

The structure of ARMA models that interprets climate effects linearly, focusing only upon the reproduction of statistics may produce inaccurate estimates of persistence in hydrologic observations. Autoregression is difficult to interpret in the context of hydrologic variability, as it fails to associate climate fluctuations to hydrologic responses. By interpreting the processes of the hydrological cycle in statistical models that better reflect physical interactions, more accurate simulations of persistence can result. The following section discusses methods by which persistence in hydrology can be more explicitly modelled.

3.2.2 Modelling the Hurst phenomenon

The results of Hurst's investigations led the scientific community to time series models that could replicate this long-term phenomenon, which appeared to be the rule rather than the exception in historic time series (Klemes, 1974). Linear autoregressive models failed to produce the $H \approx 0.72$ behaviour identified by Hurst and therefore a range of alternative time series models were developed. This effort came despite the difficulties associated with obtaining accurate estimates for H in short hydrologic time series using statistics such as the rescaled adjusted range. These difficulties are illustrated in Figure 3.1, which shows realisations from an AR(1) process, which does not satisfy the definitions of mathematical persistence. These realisations show a tendency for $H \to 1$ as $\phi_1 \to 1$, although this bias reduces as sample length increases. In this figure estimates of H are obtained using the rescaled adjusted range statistics for 1000 simulations each of length 100 using a range of autocorrelation coefficients. The bias in estimates for H due to sample length is further investigated in Section 5.2, which shows that the identification of the Hurst phenomenon within hydrologic time series is obstructed by insufficient record lengths.

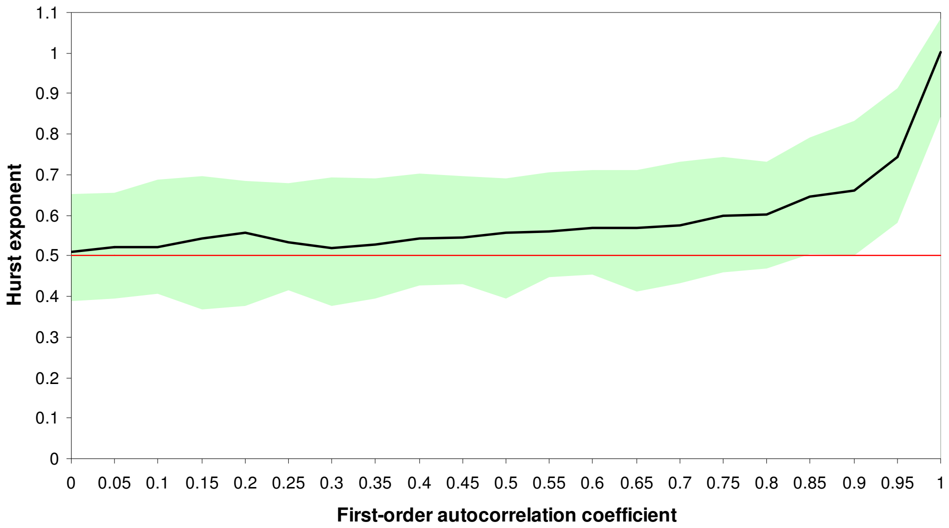

Figure 3.1 Variation in Hurst exponent with variation in autocorrelation from Monte Carlo simulations of AR(1) models showing median and 90% of samples

Notwithstanding these results, a breakthrough in efforts to devise new stochastic approaches for replicating sample estimates of *H* was achieved in the 1960s. Mandelbrot and van Ness (1968) and Mandelbrot and Wallis (1968), amongst others, proposed a model that could explicitly to preserve the Hurst phenomenon. Rather than starting with a defined model structure then testing the ability of its extrapolations to replicate observed statistical characteristics, these authors instead began with the specification of the rescaled range results, and then moved "backwards" to design an appropriate model. The Hurst phenomenon was shown to arise through a class of processes termed fractional Brownian noises (fBn's) that displayed infinite memory, a characteristic by which the value at any time step is dependent, to some degree, on all previous values. These models are stationary and display the property of self-similarity, in which short sequences have the same statistical features as rescaled long series. Infinite memory cannot arise in any physical context, yet it provides a useful device for a mathematical representation of a physical process.

Fractional noise models were developed through a desire to reproduce the geometric patterns of historic series mathematically (Klemes, 1974). This signalled a major departure from existing modelling techniques, described by Mandelbrot and van Ness (1968):

> "…we selected fractional Brownian motion so as to be able to derive the results of practical interest with a minimum of mathematical difficulty"

Since the introduction of fBn's, other models capable of reproducing the Hurst phenomenon have also been introduced, including fractional autoregressive integrated moving average (fARIMA) models (Hosking, 1984) and broken line processes (Rodriguez-Iturbe *et al.*, 1972), both of which are directly related to fBn's. Fractional ARIMA models are a novel development of the standard ARIMA process, in which the degree of differencing takes a non-integer value that is related to the Hurst exponent as $d = H - 0.5$. These models reproduce the asymptotic quality of mathematical persistence, and their representation of both short-term and long-term persistence produces a flexible modelling approach. Realisations of fractional noise processes reflect oscillatory movements of longer periodicities than Markov processes (Matalas, 1997), with the infinite sum of serial correlation coefficients for such processes being unbounded. Despite this lack of understanding and the large sampling errors of statistics used to estimate the Hurst exponent, these long-memory models produced great excitement in the scientific community. Indeed the conviction of these early authors in the mathematical precision of their stochastic developments was summarised by Wallis and O'Connell (1973), who stated that

> "to emphatically state that hydrologic records do not exhibit long-term persistence demands either a very naïve understanding of statistics or a monumentally large data base"

In order to provide some justification for the use of long memory self-similar models that can explicitly simulate the value of the Hurst coefficient in hydrologic analysis, various authors have attempted to compare their performance against more established short-memory models. Young and Jettmar (1976) sought to provide reservoir operational guidelines to assist in the use of either modelling approach for the simulation of inflows. These simulation studies preserved the covariance structures of historic observations, rather than specific statistics such as the first order autocorrelation or Hurst coefficient. A least squares analysis was used, in which the squared error for a correlogram of both Markovian and self-similar models was minimised for monthly streamflows. These results suggested minimal difference between short- or long-memory models, such that little design loss was encountered with incorrect model assumptions.

The applicability of long-memory models for hydrologic simulation was also investigated by Klemes *et al.* (1981), who undertook a comparison of Markovian and broken-line models for the simulation of annual stream flows. These authors also undertook a reservoir reliability investigation with each model using simulated monthly flows, obtained through the disaggregation of simulated annual values, as input sequences. Once again, the differences in reliability that resulted from replacing one model for the other was small when compared both to the level at which the socioeconomic impact of changes could be measured and to the accuracy of reliability estimates. In light of these results, the replacement of short-memory models with more complex long-memory models for hydrologic simulation was unjustified. A further word of caution in describing models for apparent long-range dependence is provided by Beran (1994), who noted the possible influences of aggregating various time series upon the dependence structure. In particular the conditions through which the aggregation of independent

AR(1) processes can artificially induce long-term memory are described. Although this may provide a possible explanation for mathematical persistence, it also highlights the difficulty posed in identifying genuine mathematical persistence in an observed time series.

Subsequent to the development of stochastic models designed to replicate the Hurst phenomenon during the late 1960s and early 1970s, Klemes (1974) noted that the analysis and modelling of hydrologic time series during this time lost its hydrologic context. Although approximations to fractional noise models provide simulations that exhibit the Hurst phenomenon, these models do not necessarily provide a suitable physical explanation. Klemes (1974) warns against mathematically-derived hydrologic models dominating those derived from predominantly physically-based sources:

> "(the hydrologist's) mission is to view the series in its physical context, to seek explanations of its peculiarities in the underlying physical mechanism rather than to postulate the physical mechanism from a mathematical description of these peculiarities"

This point is significant, not only in the context of the Hurst phenomenon, but also to the wider scope of hydrologic modelling. Mathematical precision of stochastic simulations is imperative; however time series models should be promoted and used within the boundaries of physical reasoning. Although it is vital that stochastic models replicate the statistical characteristics of historical series, it is somewhat more important that the relevance of such statistics is maintained. The Hurst phenomenon demonstrates the complexity and unpredictability of hydrologic processes, yet its somewhat abstract nature means that to design and promote stochastic models based on their ability to replicate its features is inappropriate.

It is well known that long-memory models such as fBn's have an impressive and flexible operational framework. However, again taking the words of Klemes (1974)

> "the ability to simulate, and even predict, a specific phenomenon does not necessarily imply an ability to explain it correctly"

From this perspective, successful operational models may, in fact, be unacceptable from a physical perspective. The aim of this thesis is to develop stochastic models that have a stronger physical basis for the identification and simulation of hydrological persistence. The Hurst phenomenon is without doubt an important characteristic of geophysical time series, however it is also clear that persistence is not the exclusive and indispensable (Klemes, 1974) feature of a process exhibiting the Hurst phenomenon. Rather it is argued here that the impact of hydrological persistence can be more reliably assessed through the direct analysis of wet and dry spells. Stochastic models designed to replicate explicitly the characteristics of observed spells are therefore preferable to models designed to reproduce solely the feature demonstrated in Hurst's double logarithmic plots of the rescaled range statistic.

3.2.3 Shifting level (SL) models

The Shifting Level (SL) model was first introduced to the field of hydrology by Boes and Salas (1978), and further developed in subsequent papers (eg Salas and Boes, 1980) as a method to simulate time series displaying sudden shifts in the mean, such as observed in some hydroclimatic processes. Salas and Boes (1980) suggest that these are capable of reproducing historical drought characteristics, which provides an advantage over linear ARMA models for operational hydrology.

A series of observations $\{y_t, t = 1,2,...,T\}$ are considered to be realisations of two independent stochastic processes, as

$$y_t = m_t + z_t \tag{3.19}$$

The m_t term corresponds to the latent mean of the observation at time *t*, whereas the z_t term is white noise with variance σ_z^2. The mean levels are normally distributed as $N(\mu, \sigma_\mu^2)$, such that $E[y_t] = \mu$ and $\text{var}[y_t] = \sigma_\mu^2 + \sigma_z^2$. The mean level remains constant for *epochs*, with the duration of such periods assumed to follow a geometric distribution, leading to the $\{m_t\}$ process shown to be a Markov chain.

Salas and Boes (1980) showed that the autocorrelation of the SL model is identical to that of an ARMA(1,1) model, for which

$$r(k) = r(1)\phi^{k-1} \tag{3.20}$$

where $r(1)$ is the first serial correlation coefficient and ϕ is the AR parameter. Even if observations are independent within each epoch, the random shifts in the mean level create dependence between observations. This feature is consistent with various hydrologic observations with Hurst *et al.* (1965) noting that the mean annual volume of the Nile River at Aswan for the period 1870-1898 was 110x10^9 m^3, while during 1899-1957 was 83x10^9 m^3. Hurst *et al.* (1965) further noted the serial correlation for the first period to be 0.11 and 0.12 for the latter period, however over the whole period this increases to 0.49, suggesting that a change in mean level can increase correlation.

Sveinsson *et al.* (2003) showed that SL models may preserve the mean, variance and the autocorrelation of a range of historical annual streamflow and climate series, as well as generating sequences with abrupt changes that are similar to historical patterns. When observing drought and storage-related statistics however, simulations from SL models failed to provide a significant improvement over simulations from the traditional ARMA(1,1) model. Fortin *et al.* (2004) extended the applications of SL models from the simulation of hydrologic data through the development of a forecasting algorithm. By fitting the SL model to annual flows in the Senegal River, Fortin *et al.* (2004) concluded that point forecasts from this model were inferior to those obtained from a linear ARMA(1,1) model. These results were explained through the fact that linear models can provide a good approximation of non-linear time series where there is excessive “noise”.

Advantages of SL models include their ability to retrospectively identify multiple shifts in a time series, and to simulate such series, thereby relating hydrologic series to changes in the regional climate. The SL model is stationary and forecasts will converge towards the mean. Fortin *et al.* (2004) suggest that a method to test whether observed streamflows are stationary could involve the comparison of the SL model with a nonstationary model that also provides for shifts in the mean. Suitable nonstationary models include segmentation or change-point models that assume single shifts in the process mean to occur in the period of a record, assuming this shift to be permanent and therefore altering conditions of the process. The application of change-point models to hydrological persistence is discussed in the following section.

3.2.4 Segmentation models

The SL model described in the previous section is related to a more general family of stochastic models that can incorporate heterogeneity into hydrologic time series. Segmentation models are designed to identify blocks of contiguous data (*segments*) within observed time series, such that each segment retains homogeneity. Time series segmentation can be considered a particular form of clustering (Kehagias, 2004), under the constraint that the linear order of the sampled data is retained.

There is extensive literature focused upon the segmentation of hydrologic time series, most work focusing upon detecting the location and magnitude of a single change in a time series of hydrologic observations. Various studies (eg Kiely *et al.*, 1998) have relied upon standard non-parametric statistical tests to detect the most significant change point in observed series. Perreault *et al.* (1999) described a Bayesian method to detect a single change in the mean level of a time series, thus dividing an observed time series into two series of random variables, operating at two different mean levels. Although the model formulation of Perreault *et al.* (1999) assumes no shifts in the variance of the time series, this aspect of model structure is considered in the later study of Perreault *et al.* (2000a). Bayesian model selection was then used by Perreault *et al.* (2000b) to investigate the applicability of a range of different models that take into account changes in mean and variances around a single change point.

The effect of having multiple abrupt changes in the statistical parameters of a time series to divide an original time series into a number of smaller segments was considered by Hubert (2000). This multiple change-point procedure can be defined in the following way:

Given a time series of length T $\{y_t, t = 1,2,...,T\}$, a series $y_i, i = i_1, i_2$ (where $i_1 \geq 1$ and $i_2 \leq T$) constitutes a segmentation of the initial series. The division of an initial series into m segments constitutes an m-order segmentation of this series. The length of segment $i_k, k = 1,2,...,m$ can be noted as $n_k = i_k - i_{k-1}$, with the local mean of the segment defined as:

$$\overline{y}_k = \frac{\left(\sum_{i=i_{k-1}+1}^{i=i_k} y_i\right)}{n_k} \qquad \textbf{(3.21)}$$

The quadratic deviation (sum of squared departures from the mean) for each segment is given as

$$d_k = \sum_{i=i_{k-1}+1}^{i=i_k} (y_i - \overline{y}_k)^2 \qquad \textbf{(3.22)}$$

And the quantity

$$D_m = \sum_{k=1}^{k=m} d_k \qquad \textbf{(3.23)}$$

is the quadratic deviation between the whole series and the considered segmentation. Hubert (2000) provides an algorithm to determine, for a given order of segmentation, the optimal segmentation of a series such that this deviation is minimised. Hubert (2000) applied this multiple-segmentation procedure to a range of annual rainfall series, although various drawbacks were identified. The "branch-and-bound" optimisation routine can only be applied to time series of up to 100 values, after which it becomes too computationally inefficient. The analysis of monthly rainfall records is therefore beyond the capability of such an algorithm. This optimisation routine was improved by Kehagias (2004), such that much longer time series containing multiple change points could be analysed efficiently.

Rasmussen (2001) outlined a procedure that combined a Bayesian approach with the generalised linear model to analyse changes in statistical parameters of hydrologic time series. By first specifying models for the processes either side of a change, this Bayesian change point analysis focuses upon locating where the change occurred and also the size of such a change. Segmentation procedures provide explicit methods for detecting spells within time series of observations, using these to infer details of persistence. Notwithstanding this, extensions to the original SL model formulation proposed by Fortin *et al.* (2004) provide a method to define wet and dry spells and also to detect abrupt changes in the local mean of a process, thereby generalising the segmentation procedure of Hubert (2000). Although such models are promising, it is argued here that hidden Markov models (HMMs) provide a better alternative, and are the focus of the following section.

3.3 Hidden Markov models (HMMs)

Hidden Markov models (HMMs), also known as Markov switching models or Markov mixture models, provide a similar modelling structure to the SL model. However rather than focusing upon shifts in the mean of a process, HMMs estimate shifts in the *state* of a process, with observations then conditional upon model state. This modelling approach therefore describes hydrologic totals in terms of fluctuations between discrete climate states. HMMs have been used successfully in a broad range of scientific applications, including speech recognition (Juang and Rabiner, 1991), image classification (Li *et al.*, 2000), modelling of biological sequences (Churchill, 1989; Le Strat and Carrat, 1999) and econometrics (Ryden *et al.*, 1998). In the field of hydrology, both discrete-valued and continuous-valued HMMs are prevalent, the former through modelling of daily precipitation data (eg Zucchini and Guttorp, 1991) and the latter with annual totals of both rainfall and streamflow observations.

One of the first applications of HMMs to modelling continuous-valued hydrologic data was by Jackson (1975a) who described a two-state HMM for annual streamflows, terming this a Markov mixture model. This model was advanced as a method for generating streamflows that produced drought lengths observed in historic records. With drought length as the model parameter of primary concern, the author hypothesised that these states represented low and normal streamflow conditions, with annual flows in each state being random Gaussian variates. Jackson (1975a) suggested extensions to this model such as incorporating correlation between flow values in successive time steps, with correlation values depending upon model state. Parameter estimation for the Markov chain transition probabilities was undertaken separately from the parameters describing conditional state distributions, and relied upon estimating values that would explicitly reproduce historical drought lengths. Little attention was paid however to defining the flow levels that constituted drought years.

The methodology introduced by Jackson (1975) was a novel approach to a hydrologic simulation study and improved simulations achieved from simpler Markov models. Bayazit (1982) extended this methodology to incorporate three-state Markov models. These were argued to preserve the run properties of flow series related to extreme periods and also the phenomenon of differential persistence, in which low flows persist longer than high flows. More recently, Thyer (2000) and Thyer and Kuczera (2000) introduced two-state HMMs to the modelling of climatic persistence within annual totals of rainfall. Using the hypothesis that climatic fluctuations led to multi-year persistence in hydrologic observations, these authors argued that the modelling structure of HMMs provided a superior description of interannual variability than could be achieved using linear models such as an AR(1) model. Thyer and Kuczera (2003a; 2003b) extended this work to consider observations at multiple sites, with Frost (2003) adopting multivariate two-state HMMs to incorporate interannual persistence into a rainfall model of six-minute resolution. In light of these recent studies however, there has been minimal use of HMMs for describing persistence in time series of monthly hydrologic totals. An annual scale may be too coarse to identify accurately the dominant modes of variability, typified by ENSO periods having an average length of 15 months. Therefore a higher frequency such as the monthly scale may more clearly reveal this climatic persistence.

Various authors have used the structure and optimisation routines of HMMs to re-formulate existing stochastic models. In particular, Fortin *et al.* (2004) recently revisited the SL model, formulating it as a HMM and developed procedures for the retrospective analysis and segmentation of streamflow time series, and also for producing forecasts. Also, Kehagias (2004) used a HMM formulation of the segmentation procedure of Hubert (2000) to detect multiple change points in hydrological and environmental time series. With these two studies showing the SL and segmentation models belong to the broader HMM family, it is appropriate to focus on the latter as a candidate modelling approach for persistent hydrological series. Furthermore, since the Markov property is a simple and mathematically tractable relaxation of the assumption of independence (MacDonald and Zucchini, 1997), discrete-time Markov chains on a finite state space are an appropriate method to represent fluctuations between dominant modes of climatic persistence. Methods for the calibration of HMMs, and their inference, are described here.

3.3.1 Fundamentals of HMMs

Discrete-time hidden Markov models (HMMs) are described in terms of a pair of processes $\{(x_t, y_t)\}$. These are models for a time series of observations $\{y_t, t = 1,2,...,T\}$, with a probability distribution determined by the state x_t of an unobserved *k*-state Markov chain. In the case of $k = 1$, the HMM degenerates to a series of mutually independent random variables. In hydrologic applications, $\{y_t\}$ can represent rainfall or streamflow totals at discrete-time intervals (for example monthly or annual totals), with $\{x_t\}$ characterising a set of climate states that influence the hydrological cycle. This provides a straightforward method for modelling the interaction between persistent climate regimes and hydrologic responses.

Suppose now that x_t can be described at any time t $(t = 1,..,T)$ as being in one of a set of k distinct states $\{s_1, s_2, ..., s_k\}$. As a result, $\{x_t, t = 1,2,...,T\}$ is the state series, and in the case of a two-state model, its value s_1 may represent a predominantly wet state and s_2 the dry state. The joint distribution of a first-order Markov chain satisfies

$$P(x_t = s_j \mid x_{t-1} = s_i, x_{t-2} = s_h, ..., x_1 = s_a) = P(x_t = s_j \mid x_{t-1} = s_i) \qquad \textbf{(3.24)}$$

If it assumed that $P(x_t = s_j \mid x_{t-1} = s_i)$ depends only on (i, j) and is thus independent of time, the model is assumed stationary, with

$$a_{ij} = P(x_t = s_j \mid x_{t-1} = s_i) \qquad \textbf{(3.25)}$$

being the set of one-step transition probabilities with $\sum_{j=1}^{k} a_{ij} = 1 \ \ (i = 1,...,k)$.

The $k \times k$ matrix $A = \{a_{ij}\}$ is termed the state transition probability matrix. The self-transition probabilities for each state s_i are then expressed as $\{a_{ii}\}$, where $a_{ii} = 1 - \sum_{j \neq i} a_{ij}$. The unobserved process x_t being in state s_i $(1 \leq i \leq k)$ at time t is expressed as $s_{i,t}$. In this model, the sequence of model states $\{x_t\}$ is not directly observed (i.e. “hidden”), rather observed through the second set of stochastic processes $\{y_t\}$. The relationship between the observed series and the hidden state sequence is formally defined by the assumption

$$P(y_t \mid X_t, Y_{t-1}) = P(y_t \mid s_{i,t}) \qquad \textbf{(3.26)}$$

where Y_{t-1} is the sequence of observations from time 1 to time $t-1$, $\{y_1, y_2, ..., y_{t-1}\}$, and similarly for X_t. The observed process may be either discrete valued or continuous, and is described as conditionally independent random variables. For continuous y_t, $P(y_t \mid s_{i,t})$ represents the height of a probability distribution function. As the state-dependent distributions of HMMs can follow any discrete- or continuous-valued distribution, a wide variety of time series can be modelled. The joint distribution of hidden and observed variables is summarised:

$$P(Y_T, X_T) = P(x_1) \prod_{t=1}^{T-1} P(x_{t+1} \mid x_t) \prod_{t=1}^{T} P(y_t \mid x_t) \qquad \textbf{(3.27)}$$

Therefore, the complete specification of a HMM requires specification of two parameters (*T* and *k*), together with three probability measures:

1. Initial state probabilities $P(x_1)$
2. Transition probabilities $P(x_t \mid x_{t-1})$ and
3. Output (observation) probabilities $P(y_t \mid x_t)$

If initial state probabilities are defined as

$$\pi^i = P(x_1 = s_i) = P(s_{i,1}) \tag{3.28}$$

then $\pi = \{\pi_i\}$ is the initial state distribution.

Although given a particular state, the observed sequence is not assumed to share the Markov property, and consequently HMMs are a special case of state-space models (eg Kitigawa, 1987). HMMs are also a generalisation of mixture models, and if the model states are assumed to be independent on time instead of Markovian, a mixture model is obtained exactly (Hughes and Guttorp, 1994). Markov chains are known as "memory-less" models, as change in their model structure is only dependent upon their current state. An inherent characteristic of Markov chains is that the durations in each hidden state are exponentially-distributed. The discrete-time analogue of this, the geometric distribution, is demonstrated by showing that the probability of observing *d* consecutive periods of a model remaining in state s_i, having self-transition probabilities a_{ii}, is

$$p_i(d) = (a_{ii})^{d-1}(1 - a_{ii}) \tag{3.29}$$

A probabilistic framework for movement between the HMM states produces an explicit model to account for persistence. Higher values of self-transition probability a_{ii} produce a model with a tendency to remain in state s^i for longer periods.

Transitions between states can reflect fluctuations in the broader climate, in effect reproducing the inherent persistence of climatic phenomena. The strength of this persistence is measured by the sum of self-transition probabilities. This view is supported by MacDonald and Zucchini (1997), who note that in some applications of HMMs, the states of the Markov chain may have a useful substantive interpretation. In applications of HMMs to hydrologic variables (eg Zucchini and Guttorp, 1991; Thyer and Kuczera, 2000), climate states may correspond to meteorologically defined weather states. Furthermore, even if HMMs are not substantive models, they may still provide a useful empirical model. This is similar to ARMA models, which while having little association to subject-matter, provide valuable empirical resources.

3.3.2 Moments of HMMs

In order to derive succinctly the general moments of a HMM, assume that observations in each model state are drawn from Gaussian distributions, such that

$$y_t = \mu_{j,t} + \sigma_{j,t}\varepsilon_t \tag{3.30}$$

where ε_t is a standard normal variate, and $\mu_{j,t}$ and $\sigma_{j,t}$ are the mean and standard deviation of state *j* assuming the model is in state *j* at time *t*.

Using the results of Timmermann (2000), the centred moments of the observation sequence are derived from the following formula, assuming $\boldsymbol{\mu}$ to be the unconditional mean of the sequence:

$$\begin{aligned} E[(y_t - \mu)^n] &= E[E[(y_t - \boldsymbol{\mu})^n \mid y_t]] \\ &= E[E[\{(\mu_{j,t} - \boldsymbol{\mu}) + \sigma_{j,t}\varepsilon_t\}^n \mid y_t]] \\ &= \sum_{j=1}^{k} \pi_j \sum_{d=0}^{n} \binom{n}{d} (\mu_{j,t} - \boldsymbol{\mu})^{n-d} (\sigma_{j,t})^d \, E[(\varepsilon_t)^d] \end{aligned} \tag{3.31}$$

where the last expression is obtained from the binomial formula (after Rasmussen and Akintug, 2004). An expression for the expected value (unconditional mean) of the observations is then derived as:

$$\begin{aligned} E[(y_t - \boldsymbol{\mu})] &= \sum_{j=1}^{k} \pi_j \sum_{d=0}^{1} \binom{1}{d} (\mu_{j,t} - \boldsymbol{\mu})^{1-d} (\sigma_{j,t})^d \, E[(\varepsilon_t)^d] \\ \therefore E[y_t] - \boldsymbol{\mu} &= \sum_{j=1}^{k} \pi_j (\mu_{j,t} - \boldsymbol{\mu}) \\ \therefore E[y_t] &= \sum_{j=1}^{k} \pi_j \mu_{j,t} - \boldsymbol{\mu} \sum_{j=1}^{k} \pi_j + \boldsymbol{\mu} \\ &= \sum_{j=1}^{k} \pi_j \mu_j \end{aligned} \tag{3.32}$$

since $\sum_j \pi_j = 1$ by definition. The notation for the state means in the last line of this expression is modified to take into account the independence of the mean value from the time order of observations. Likewise the variance of the observation sequence $\boldsymbol{\sigma^2}$, which is the second centred moment, is shown by

$$\boldsymbol{\sigma^2} = \sum_{j=1}^{k} \pi_j [(\mu_j - \boldsymbol{\mu})^2 + (\sigma_j)^2] \tag{3.33}$$

In addition to the variance, an expression for the autocovariance of the observations needs to also be specified to provide necessary information about the second moments of the observations sequence. The autocovariance at lag τ , $\gamma(\tau)$, can be specified as

$$\begin{aligned}\gamma(\tau) &= E[(y_{t+\tau}-\boldsymbol{\mu})(y_t - \boldsymbol{\mu})] \\ &= E[(\mu_{i,t+\tau} + \sigma_{i,t+\tau}\varepsilon_{t+\tau} - \boldsymbol{\mu})(\mu_{j,t} + \sigma_{j,t}\varepsilon_t - \boldsymbol{\mu})] \\ &= E[\mu_{i,t+\tau}\mu_{j,t}] - \boldsymbol{\mu}^2\end{aligned} \tag{3.34}$$

assuming that the model is in state j at time t and state i at time $(t+\tau)$. The last line of this expression is attained by noting that all product terms involving ε_t or $\varepsilon_{t+\tau}$ become zero after taking expectations (Rasmussen and Akintug, 2004). The expectation term in the last line is

$$\begin{aligned}E[\mu_{i,t+\tau}\mu_{j,t}] &= \sum_{j=1}^{k}\sum_{i=1}^{k}\mu_{i,t+\tau}\mu_{j,t}P(s_{i,t+\tau}, s_{j,t}) \\ &= \sum_{j=1}^{k}\sum_{i=1}^{k}\mu_{i,t+\tau}\mu_{j,t}P(s_{i,t+\tau} \mid s_{j,t})P(s_{j,t}) \\ &= \sum_{j=1}^{k}\sum_{i=1}^{k}\mu_{i,t+\tau}\mu_{j,t}\pi_j P(s_{i,t+\tau} \mid s_{j,t})\end{aligned} \tag{3.35}$$

using the result that the stationary distribution of model states equals initial probabilities. The probability of a transition between model states in τ steps is derived from the one-step transition probabilities. For the case of $\tau = 2$, a transition between model states is established

$$P(x_{t+2} \mid x_t) = \int P(x_{t+2} \mid x_{t+1})P(x_{t+1} \mid x_t)dx_{t+1} \tag{3.36}$$

And likewise for $\tau = 3$

$$P(x_{t+3} \mid x_t) = \int P(x_{t+3} \mid x_{t+2}) \int P(x_{t+2} \mid x_{t+1})P(x_{t+1} \mid x_t)dx_{t+1}dx_{t+2} \tag{3.37}$$

This formula can then be generalised for all τ by multiplying the transition probabilities and integrating τ times. Rasmussen and Akintug (2004) provide a concise representation of Eq. 3.37 that makes use of this basic property of Markov chains, thereby removing the double summation. By defining $\mathrm{M} = (\mu_1\mu_2...\mu_k)$ as the vector of state conditional means and Π as the diagonal matrix containing stationary probabilities $\pi_1, \pi_2, ..., \pi_k$, Eq. 3.37 is expressed for a lag of τ as

$$E[\mu_i\mu_j] = \mathrm{M}\Pi\mathrm{A}^{\tau}\mathrm{M}^T \tag{3.38}$$

using the transition probability matrix A defined earlier. The HMM autocovariance function for the HMM can then be expressed as

$$\gamma_\tau = \mathrm{M}\Pi\mathrm{A}^\tau\mathrm{M}^T - \boldsymbol{\mu}^2 \tag{3.39}$$

The autocovariance function can be standardised by the variance of the process, $\rho_\tau = \gamma_\tau / \sigma^2$, in order to produce the autocorrelation function (acf) at lag τ. The ARMA(1,1) model was shown in Section 3.2.1 to have an acf of the form

$$\rho_\tau = \rho(1)\phi^{\tau-1} \tag{3.40}$$

with $\rho(1)$ the first serial correlation coefficient and ϕ the autoregressive parameter. This is an example of an exponentially-decaying acf, which is characteristic of *short-memory* models. In the case of the HMM, it is clear that the ratio of $\rho_\tau / \rho_{\tau-1}$ will be constant for all $\tau > 2$, and it follows that the acf will decay at the exponential rate of $\rho_\tau = \alpha\beta^\tau$ similar to that of an ARMA(1,1) model. The importance of this result is clear when observing the context into which these models are presented. As discussed in the previous chapter, these stochastic models are being developed for time series displaying hydrological persistence, which is unrelated to other interpretations of long-memory characterised by an acf decaying at a rate approximating $\rho_\tau = \alpha\tau^{-\beta}$ where $\alpha, \beta \in (0,1)$. Such models need to show autocorrelations that decay exponentially. The HMM is therefore inappropriate for modelling long-range persistence as defined by Beran (1994), yet is consistent with the representation of hydrological persistence.

3.3.3 HMM modelling assumptions

The two main assumptions to be made in the calibration of a HMM are the number of model states and the form of state conditional distributions. The simplest implementations are two-state models, which in the context of hydrological persistence reflect a tendency for the climate to produce predominantly wet (W) and predominantly dry (D) conditions. This assumption is justified in light of the evidence presented in Section 2.2.3 that suggests that broad-scale atmospheric conditions fluctuate between two stable regimes. The two-state (binary) modelling structure is defined by two transition probabilities, $P(x_t = D \mid x_{t-1} = W)$ and $P(x_t = W \mid x_{t-1} = D)$, abbreviated to P_{WD} and P_{DW} respectively. These probabilities are complementary to the two self-transition probabilities, such that $P_{DW} + P_{DD} = 1$ and $P_{WD} + P_{WW} = 1$. An example of a two-state (binary) HMM is shown in Figure 3.2, adapted from Elliott *et al.* (1995), which demonstrates that the state fluctuations of a two-state Markov chain tends to be concealed when observed through noise.

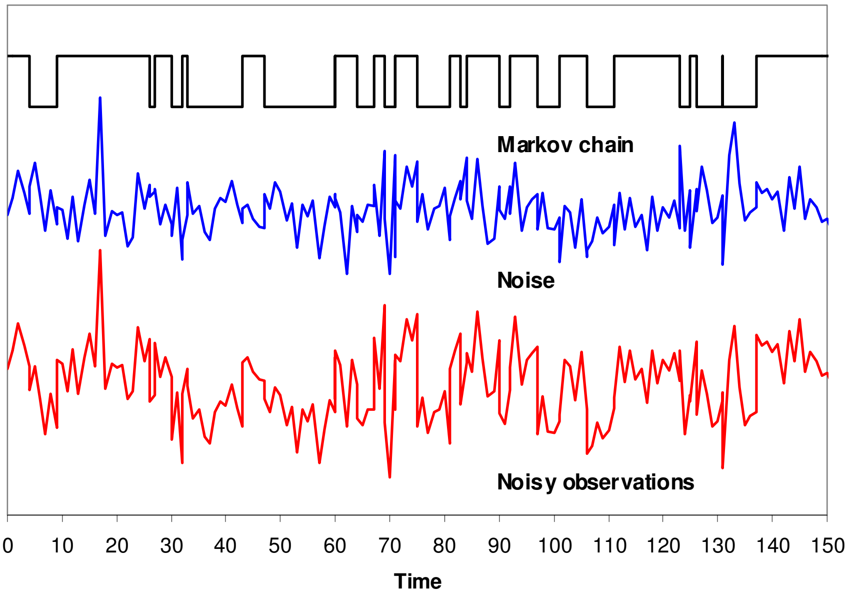

Figure 3.2 Schematic diagram of a binary HMM (adapted from Elliott *et al.*, 1995)

In the standard representation of a HMM, the observed series is a random output dependent on the value of the current state, through an *emission distribution* function that usually follows an assumed probability distribution. In the application of two-state HMMs by Thyer (2000) and Thyer and Kuczera (2000), the state conditional distributions for annual rainfall were assumed to be Gaussian. The marginal distribution of hydrologic totals aggregated over shorter time scales (eg monthly or seasonal totals) generally have higher positive skew than annual totals. The modelling of such data may therefore demand alternative modelling assumptions.

Given the form of HMMs presented in the previous section, Rabiner (1989) outlines two main problems of interest that need to be solved in real-world applications of these models. The first problem is that when presented with a sequence of observations $\{y_t\}$ and a modelling framework, how is the probability of this sequence given the model $P(Y_T \mid \theta)$ computed efficiently? This can be termed an *evaluation* problem, which calculates the probability that a specific sequence was generated by the model. Associated with this is the requirement to maximise this probability to some optimality criteria. The second problem is one in which details of the hidden part of the model is estimated. Given the observation sequence and model, how is a sequence of model states that best “explains” the observations identified? Methods to solve each of these modelling problems are presented in the following section.

3.3.4 Evaluating HMM likelihood

Fitting a HMM to an observed sequence requires evaluating the probability that the sequence was produced by the specific model, $P(Y_T \mid \theta)$. More specifically, as HMM model-fitting reduces to the task of estimating a set of unknown model parameters θ, the evaluation problem can be re-formulated in terms of these parameters. The likelihood function is the probability of the observed data given the parameter vector θ, $P(Y_T \mid \theta)$. This probability is defined without regard to the state sequence such that X_T can be incorporated in the following manner

$$\begin{aligned} L_T(\theta) &= P(Y_T \mid \theta) \\ &= \sum_{x_1,\dots,x_T} P(Y_T, X_T \mid \theta) \\ &= \sum_{x_1,\dots,x_T} P(x_1 \mid \theta) \prod_{t=2}^{T} P(x_t \mid x_{t-1}, \theta) P(y_t \mid x_t, \theta) \end{aligned} \tag{3.41}$$

The maximum likelihood estimates (MLEs, $\hat{\theta}$) of the parameters are the values of θ that maximise $L_T(\theta)$. The simplest method for estimating unknown model parameters is by the direct numerical maximisation of the likelihood function (Zucchini and MacDonald, 2002), however this may be computationally intractable. As an alternative, efficient algorithms such as the Baum-Welch algorithm (Baum *et al.*, 1970) facilitate the calibration of a HMM to observation sequences. This algorithm is also known as the *Forward-Backward procedure* and has been used in many HMM applications (eg Juang and Rabiner, 1991). In fact, this method can be interpreted as an early example of an algorithm of EM type (see Dempster *et al.*, 1977). In this algorithm, the joint probability of a partial observation sequence Y_t and the value of the model state at the end of such a sequence $P(Y_t, s_{j,t} \mid \theta)$ is evaluated. This probability, referred to as a *forward variable*, is further defined in terms of the recursion:

$$\begin{aligned} \alpha_t(j) &= P(Y_t, s_{j,t}) \\ &= P(y_t \mid s_{j,t}) \sum_{i=1}^{k} P(s_{j,t} \mid s_{i,t-1}) P(Y_{t-1}, s_{i,t-1}) \\ &= P(y_t \mid s_{j,t}) \sum_{i=1}^{k} P(s_{j,t} \mid s_{i,t-1}) \alpha_{t-1}(i) \\ &= P(y_t \mid s_{j,t}) \sum_{i=1}^{k} a_{ij} \alpha_{t-1}(i) \end{aligned} \tag{3.42}$$

This recursion is initialised with $\alpha_1(i) = P(y_1, s_{i,1}) = P(y_1 \mid s_{j,1}) \pi_i$. The basic idea of this algorithm is to successively pass each of the multiple summations in the likelihood as far to the right as possible (Hughes and Guttorp, 1994). It follows that the likelihood of the sequence is readily computed as

$$L_T(\theta) = \sum_{i=1}^{k} P(Y_T, s_{i,T} \mid \theta) = \sum_{i=1}^{k} \alpha_T(i) \tag{3.43}$$

using only the forward variables at the terminal position in the observation series.

In a similar manner, a *backward variable* can be considered as the probability of a partial observation sequence from time $(t+1)$ to the end, conditioned on the model state at time t. A recursive procedure for this calculation, again omitting conditional dependence on unknown model parameters, is as follows:

$$\begin{aligned}\beta_t(i) &= P(y_{t+1}, y_{t+2}, ..., y_T \mid s_{i,t}) \\ &= \sum_{j=1}^{k} P(s_{j,t+1} \mid s_{i,t}) P(y_{t+1} \mid s_{j,t+1}) P(y_{t+2}, .., y_T \mid s_{j,t+1}) \\ &= \sum_{j=1}^{k} P(s_{j,t+1} \mid s_{i,t}) P(y_{t+1} \mid s_{j,t+1}) \beta_{t+1}(j) \\ &= \sum_{j=1}^{k} a_{ij} P(y_{t+1} \mid s_{j,t+1}) \beta_{t+1}(j)\end{aligned} \tag{3.44}$$

The initialisation step to this recursion subjectively defines $\beta_T(i) = 1$ for all *i*.

Results from Eq. 3.43 and Eq. 3.44 are used to evaluate the "optimal" state sequence associated with the observation sequence and the model assumptions. This is a useful result from HMM calibration, as it provides important information about the role of model states. Given the observation sequence $\{y_t, t = 1,2,...,T\}$, the probability of the model in state s^j at time *t* is

$$\begin{aligned}\gamma_t(j) &= P(x_t = s_j \mid Y_T) \\ &= \frac{P(Y_t, s_{j,t}) \times P(y_{t+1}, y_{t+2}, ..., y_T \mid s_{j,t})}{P(Y_T)} \\ &= \frac{\alpha_t(j)\beta_t(j)}{\sum_j \alpha_t(j)\beta_t(j)}\end{aligned} \tag{3.45}$$

The normalisation factor $P(Y_T)$ has the result of $\sum_i \gamma_t(i) = 1$. The most likely state at each time, q_t, will then be the state s_j that maximises the term $\gamma_t(j)$,

$$q_t = \underset{1<i<k}{\arg\max}[\gamma_t(i)] \tag{3.46}$$

Although this provides the most likely state at each time step *t* through maximising the expected number of correct states, problems may arise when the system has more than two states. When some HMM transition probabilities are equal to zero, the most favourable state sequence may not even be valid. One solution to this problem is to find the *best* state sequence by maximising the probability of a sequence of states. A formal method for achieving this is the Viterbi algorithm (Forney, 1973), which is similar to the forward recursion except that the summation procedure in the latter is replaced by a maximisation over previous states.

3.3.5 Global optimisation routines

The calibration of HMMs relies upon identifying parameter values that maximise the HMM likelihood, with parameter estimates $\hat{\theta}$ that maximise the likelihood taken as optimal values. This is an important aspect in the application of HMMs, and it is useful to discuss methods to obtain the most favourable point estimates of HMM parameters.

The calibration of HMMs to hydrologic time series may present a complex optimisation problem, with optimal parameter values being difficult to identify if the multi-parameter space contains many local minima. Global optimisation routines present a method of function minimisation that rarely suffers from convergence problems in the presence of multiple optima that can affect local-type direct search optimisation methods such as the simplex method (Nelder and Mead, 1965).

The shuffled complex evolution (SCE) algorithm introduced by Duan *et al.* (1992; 1993) is a robust, effective and efficient strategy for function maximisation/ minimisation. This algorithm combines four concepts that have each proved successful for global optimisation: a combination of probabilistic and deterministic approaches, the concept of clustering, systematic evolution of a *complex* of points across the parameter space, and the utilisation of competitive evolution. Gan and Biftu (1996) noted that these four features represented the best features of several optimisation methods. This algorithm was constructed around the controlled random search (CRS) method described by Price (1983), using its best features such as global sampling and complex evolution, and incorporated the powerful concepts of competitive evolution and complex shuffling. These latter features allow the sample information to be thoroughly exploited in order to find global solutions.

The SCE algorithm is initiated by sampling a random set of parameter values from the feasible parameter space. This set of points, termed a *population*, is then partitioned into a number of smaller groups (*complexes*), each of which can evolve independently according to a competitive complex evolution strategy, adapting the deterministic simplex method of Nelder and Mead (1965). This method of evolution allows each complex to search the parameter space in different directions. The complexes are periodically shuffled to enable information sharing, and at random locations new parameter values are introduced to the complexes to ensure the process of evolution does not get trapped by unpromising regions (Duan *et al.*, 1992). This method is repeated until sufficient convergence is achieved, with the population moving towards globally optimal values. By combining competitive evolution and complex shuffling, the SCE algorithm ensures that information about the parameter space obtained by each complex of samples is shared across the entire population, which allows an efficient search of the feasible parameter space.

Duan *et al.* (1992) demonstrated that the SCE method efficiently and effectively identifies the global optima for parameters of a conceptual rainfall-runoff model. Thyer *et al.* (1999a) presented a comparison of the SCE algorithm with another global optimisation routine, the three-phase simulated annealing (SA) algorithm. Both algorithms were used to calibrate an identical conceptual rainfall-runoff model, using the same data sets and objective function, with performance measured through robustness and efficiency. Although the two algorithms had a similar level of robustness, interpreted as the probability of finding the same optima from a series of independent trials, the efficiency of the SCE was at least six times that of the SA algorithm. Thyer *et al.* (1999a) attributed this superiority to the use of multiple complexes in the SCE, which provides this algorithm with more information about the response surface. The SCE algorithm is used in this work to evaluate maximum likelihood estimates for HMM parameters.

3.4 Bayesian modelling framework

The calibration of HMMs in this thesis is developed in order to produce a statistical description of the uncertainty in model output. Although the SCE method provides an effective means to find an optimal set of parameters according to the criterion of the HMM likelihood function, the uncertainty associated with these estimates is not addressed. This is a limitation in the calibration of many hydrologic models, as large uncertainty around estimated maximum likelihood values relates directly to uncertainty in the model itself.

A persuasive method to assess parameter uncertainty is to adopt a Bayesian approach. Bayesian inference considers the vector of unknown model parameters θ as random variables rather than fixed values (Perreault *et al.*, 1999), using a statistical distribution to expresses the uncertainty about θ. Prior to collecting data, the knowledge of these parameter values given the assumed model M is summarised by a density known as the prior distribution $P(\theta \mid M)$. Prior knowledge of the model system can then be integrated through this prior. Bayes' theorem is used to combine the model being considered with the observed data to update the prior information. This produces a posterior distribution $P(\theta \mid Y, M)$ from which statistical inference of unknown model parameters can be made. Bayes' theorem is summarised by:

$$P(\theta \mid Y, M) = \frac{P(Y \mid \theta, M)P(\theta \mid M)}{P(Y \mid M)} \tag{3.47}$$

The generalised probability distribution $P(Y \mid \theta, M)$ is the likelihood function of the data sequence, with the denominator being the marginal likelihood for this model, defined as

$$P(Y \mid M) = \int P(Y \mid \theta, M) P(\theta \mid M) d\theta \qquad \textbf{(3.48)}$$

Since the marginal likelihood of the model is independent of θ, Bayes' theorem is often simplified to

$$P(\theta \mid Y) \propto L(\theta) P(\theta) \qquad \textbf{(3.49)}$$

The prior distribution reflects beliefs about model parameters prior to model fitting, with the posterior indicating the updated beliefs after observing sample data, thereby containing all of the available information about the parameter vector θ. Whereas other calibration methods may focus upon maximisation procedures to generate a single vector of parameters, the focus of Bayesian modelling is the entire posterior distribution, and Bayesian statistical inference therefore reduces to summarising this distribution (Campbell *et al.*, 1999). Bayesian inference has been a controversial aspect of statistical modelling (Raftery, 1995), due to its reliance upon prior distributions subjectively determined by the user. In large samples however, the prior has very little influence upon the posterior.

3.4.1 Markov chain Monte Carlo (MCMC) methods

In HMMs, as with many complex models, it is not possible to derive an analytical expression for the posterior distribution. When it is not possible to evaluate this explicitly, numerical integration or analytical approximation techniques are often required (Brooks, 1998). The Markov chain Monte Carlo (MCMC) method provides an alternative, through the construction of aperiodic and irreducible Markov chains that have stationary distributions approximating the posterior distribution of interest. The key idea of MCMC procedures is to start with arbitrary values of θ and then conduct a random walk through the state space of parameters (Campbell *et al.*, 1999) by generating a sequence of dependent values from a Markov chain. Under certain conditions, these samples will converge to the stationary posterior distribution.

In recent years, a number of studies have applied MCMC methods to the calibration of hydrologic models, such as conceptual rainfall-runoff models. The Gibbs sampler (GS) method (see Casella and George, 1992) has been used in a range of hydrologic studies, including Adamson *et al.* (1999) and Barreto and de Andrade (2000), who used this MCMC algorithm for flood analysis and monthly streamflow forecasting respectively. Combining MCMC with the HMM framework, Thyer and Kuczera (2000; 2003b) published two studies that used GS in the calibration of a two-state HMM to annual rainfall at a single site (2000) and at multiple sites (2003b). Lu and Berliner (1999) also applied GS to estimate the parameters of a HMM used to simulate daily streamflow forecasts. The Metropolis algorithm (Metropolis *et al.*, 1953) is perhaps the broadest implementation of MCMC methods, used by Kuczera and Parent (1998) for the parameterisation of watershed rainfall-runoff models. Frost (2003) utilised the more general Metropolis-Hastings (MH) algorithm in the calibration of a multi-site HMM to annual rainfall in order to condition the parameters of a high frequency rainfall model. Marshall *et al.* (2004) presented a comparative study of the Metropolis-Hastings and Adaptive Metropolis algorithms in their application to conceptual rainfall-runoff models. This study followed the work of Bates and Campbell (2001), who had earlier demonstrated the efficacy of the MH approach to this calibration problem.

The posterior distribution of the model parameter vector $\boldsymbol{\theta}$ of size p is sampled with the Metropolis algorithm in the following manner:

1. An initial estimate of $\boldsymbol{\theta}^0$ is made from the parameter space, with maximum likelihood estimates often chosen

2. For i=1,2,…

   a. A proposed value for $\boldsymbol{\theta}^*$ is made from a proposal density $\pi(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{\mathbf{i-1}})$. The form, location and covariance of this density need to be estimated, with parameters generally depending on previous samples. The original application of Metropolis *et al.* (1953) used a multivariate Gaussian for the jump density, although this was generalised by Hastings (1970) to allow for non-symmetric distributions.

   b. The proposed value $\boldsymbol{\theta}^*$ is accepted as the new value $\boldsymbol{\theta}^{\mathrm{i}}$ with a probability α calculated as

   $$\alpha = \min\left\{1, \frac{p(Y \mid \theta^*)p(\theta^*)\pi(\theta^{i-1} \mid \theta^*)}{p(Y \mid \theta^{i-1})p(\theta^{i-1})\pi(\theta^* \mid \theta^{i-1})}\right\} \qquad \mathbf{(3.50)}$$

   Here, $p(Y|\boldsymbol{\theta}^*)$ is the likelihood function associated with the new samples, with $p(\boldsymbol{\theta}^*)$ being their prior distribution.

   c. This acceptance/rejection step is repeated multiple times, until the sequence of samples obtained converges to a stationary distribution that is the posterior.

An important aspect of MCMC algorithms is the specification of starting points for sampling. A range of rigorous methods have been proposed in the literature for selecting such points (eg Gelman and Rubin, 1992), although more straightforward methods such as using maximum likelihood estimates are also suitable. As a consequence, optimisation routines such as the SCE algorithm may provide useful starting points. One of the contentious issues (Brooks, 1998) of MCMC algorithms is whether to use a single Markov chain for sampling the posterior, or to use numerous shorter chains in parallel. Although using a single long chain may produce samples that are closer to the target distribution, taking a number of parallel chains can guard against portions of the sample space remaining unexplored. Such chains can be initiated in different regions of the sample space, or alternatively all can start at the maximum likelihood estimates $\hat{\theta}$. In order to use Metropolis samples as being representative of the posterior, it is important that for a given starting value $\boldsymbol{\theta}^0$ and proposal density, the sequence of $\boldsymbol{\theta}^i$ samples converges to a stationary distribution. It is important that chains are run over a "warm-up" period, the samples during which are then discarded so that the influence of the starting values is not too high.

As noted earlier, a multivariate Gaussian distribution is a suitable choice for the proposal density. The notation $\pi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})$ describes a distribution centred at the current sample location, with an initial estimate of the covariance being based on the Hessian around the sample mode (Frost, 2003). Using previous samples, this covariance can also be updated throughout the sampling process. The size of the covariance is an important parameter of this algorithm, as it directly affects the acceptance rate of the sampling procedure and its coverage of the posterior distribution. Recently, Haario *et al.* (2001) proposed the Adaptive Metropolis (AM) algorithm, a variation of the standard Metropolis algorithm in which the covariance of the proposal density is updated at each iteration based on all information obtained up to that point. By using all of the previous states of the Markov chain to calculate the covariance, the AM algorithm loses the Markovian nature of the standard Metropolis algorithm, although Haario *et al.* (2001) showed that it retains the correct ergodicity properties allowing its samples to converge to the stationary posterior distribution. The advantage of the AM method lies in the fact that the size and location of the proposal distribution need not be chosen, and that it begins to accumulate information from the beginning of the simulation. The proposal distribution for the candidate point $\boldsymbol{\theta}^i$ is a multivariate Gaussian distribution with mean at the current point $\boldsymbol{\theta}^{i-1}$ and covariance C_i given by $s_p R$, where R is the covariance matrix based on all samples up to state $\boldsymbol{\theta}^{i-1}$ and s_p is a scaling parameter depending only on the dimension p of the parameter vector. Haario *et al.* (2001) suggest a value for s_p as $(2.4)^2 / p$, as this will optimise the mixing properties of the Metropolis search in the case of Gaussian target distributions and Gaussian proposal distributions. The computational cost of the covariance calculation is quite small, as it satisfies the following recursion formula

$$C_{i+1} = \frac{i-1}{i} C_i + \frac{s_p}{i}\left(i\bar{\theta}_{i-1}\bar{\theta}_{i-1}^{\,T} - (i+1)\bar{\theta}_i\bar{\theta}_i^T + \theta_i\theta_i^T + \varepsilon I_p \right) \qquad \textbf{(3.51)}$$

with $\bar{\theta}_i$ the sample mean up to iteration i and ε a small-valued constant. The AM algorithm is straightforward to use and its rapid start allows the search to be more effective at early stages of simulation than the traditional Metropolis algorithm. This method is used throughout the present work to evaluate the posterior distributions for unknown model parameters.

3.4.2 Convergence of MCMC algorithms

Monte Carlo Markov Chain (MCMC) algorithms have become extremely popular in the field of Bayesian analysis, offering a method to iteratively simulate posterior distributions of model parameters (Cowles and Carlin, 1996). The essential idea of iterative simulation is to draw values of a random variable x from a sequence of values that converge to the desired *target distribution* of x (Gelman and Rubin, 1992). However, one of the most important implementation problems of MCMC algorithms remains the difficulty in deciding when such simulations have indeed reached this *target distribution*. Due to the transient phase of Markov chains, which is the time taken until the chain settles down to stationary behaviour (Brooks and Roberts, 1998), there is need for a method to determine at what point it is reasonable to assume that samples are representative of the underlying stationary distribution of the Markov chain, which is the posterior distributions of the model parameters. Samples taken during the transient phase, also termed the "burn-in" period, are discarded.

Various specialised techniques exist to *a priori* determine required burn-in lengths for Markov chains (eg Meyn and Tweedie, 1994), however as Brooks and Roberts (1998) note, in general it is difficult to apply such results to MCMC algorithms. Consequently, it is necessary to perform statistical analysis on MCMC output in order to assess convergence. Brooks and Roberts (1998) provide a comparative review of various approaches to assess MCMC convergence, which can be broadly categorised as being either methods that are based on monitoring selected output from the Markov chains, or methods that exploit aspects of the theoretical properties of the algorithm. Techniques of the former category are the simplest to implement, requiring only the output from MCMC simulations and little or no knowledge of the mechanism used to generate the output.

The convergence assessment technique of Gelman and Rubin (1992) is based upon normal theory approximations to exact Bayesian posterior inference (Cowles and Carlin, 1996), and requires the analysis of m independent sequences to form a distributional estimate of the variance of the random variable (Brooks and Roberts, 1998). This analysis provides a basis for an estimate of how close the Markov chains are to stationarity. The method of Gelman and Rubin (1992) is based upon estimating the variance of the target distribution from the model output. The estimator that is used, $\hat{V}$, is constructed from a weighted average of the between-chain variance and within-chain variance, such that similar variances indicate that all chains have escaped the influence of their starting points and have traversed all the target distribution (Cowles and Carlin, 1996). Brooks and Roberts (1998) summarise this method by first defining B/n, the variance between the m sequence means, where $2n$ is the desired number of MCMC iterations, as:

$$\frac{B}{n} = \frac{1}{(m-1)} \sum_{i=1}^{m} \left(\bar{\theta}_i - \bar{\theta}\right)^2 \qquad \textbf{(3.52)}$$

where $\bar{\theta}_i = \frac{1}{n} \sum_{t=n+1}^{2n} \theta_i^t$ and $\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \bar{\theta}_i$, and θ_i^t is the t^{th} sample from chain *i*. The next step in this method is to calculate *W*, which is the mean of the *m* within-sequence variances s_i^2:

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2 \qquad \textbf{(3.53)}$$

where $s_i^2 = \frac{1}{(n-1)} \sum_{t=n+1}^{2n} \left(\bar{\theta}_i^t - \bar{\theta}_i\right)^2$. Thus the estimator $\hat{V}$ is defined as:

$$\hat{V} = \frac{(n-1)}{n} W + \frac{(m+1)}{m} \frac{B}{n} \qquad \textbf{(3.54)}$$

Once this estimator is obtained for a given number of iterations and sequences, it would be beneficial to monitor the ratio of $\hat{V}$ to the variance of the true posterior distribution to determine how close to convergence the iterations have reached. Unfortunately the variance of the underlying target distribution is unknown, so convergence must be monitored with the factor by which the estimator might shrink if sampling were continued indefinitely (Cowles and Carlin, 1996). This is achieved through:

$$R_c = \frac{n}{(n-2)} \frac{\hat{V}}{W} \qquad \textbf{(3.55)}$$

where the correction term is needed to produce an estimator that approximates a Student *t* distribution. Large values of R_c suggest either that the estimate of the variance $\hat{V}$ can be decreased through more simulations, or that more samples will increase *W*, as the simulated sequences have not explored all of the target distribution. If R_c is close to 1, Gelman and Rubin (1992) suggest that each of the *m* sequences of *n* values are close to the target distribution.

The variance ratio method described here provides a straightforward convergence diagnostic that requires only the output from MCMC iterations, although it is clear that for some higher dimensional problems or models that have highly correlated parameter values, alternative approaches may be required to assess convergence. Diagnostic methods based on the spectral density of MCMC samples, or indeed methods that also use the form of the transition distribution that drives the Markov chain are amongst the alternative approaches. The latter approach may provide a more reliable diagnostic (Brooks and Roberts, 1998) due to the additional information provided by knowledge of the mechanism generating the output. For results presented throughout this thesis, the variance ratio method described is utilised to assess the convergence of posterior estimates for HMM parameters that are generated from the Adaptive Metropolis algorithm.

3.4.3 Bayesian model selection

In later chapters, the standard HMM formulation is extended in various ways to provide different models for persistent time series. When presented with statistical models that represent different statistical theories, it is necessary to identify the most appropriate model for each specific data set. Bayesian model selection provides a means by which the performance of different models can be compared, and its main characteristics summarised here.

The theory of model selection describes the problem of using data Y to select one model M_i from a list of candidate models $(M_1,...,M_k)$. A Bayesian solution to this problem is to compute the posterior probability $P(M_i \mid Y)$, which provides a numerical summary of the evidence in favour of model M_i (Wasserman, 2000) and to then select the model that maximises this probability. Prior knowledge about two models M_i and M_j is expressed through the ratio of prior probabilities in favour of M_i, $P(M_i) / P(M_j)$. Evidence in favour of one model over another is also measured by the *posterior odds* of one model versus the alternative, evaluated as the ratio of the respective posterior probabilities: $P(M_i \mid Y) / P(M_j \mid Y)$. The ratio of posterior odds in favour of M_i to prior odds in favour of M_i produces a more commonly used measure of evidence known as the Bayes Factor (BF_{ij}):

$$BF_{ij} = \frac{P(M_i \mid Y)}{P(M_j \mid Y)} \bigg/ \frac{P(M_i)}{P(M_j)} \qquad \textbf{(3.56)}$$

By using Bayes' theorem, it is clear that

$$BF_{ij} = \frac{P(Y \mid M_i)}{P(Y \mid M_j)} \qquad \textbf{(3.57)}$$

where $P(Y \mid M_i)$ is the termed the marginal likelihood for model M_i.

As a consequence, the value of BF_{ij} describes the extent to which the data changes the evidence in favour of one model over an alternative (Wasserman, 2000). In many cases the prior evidence for each model is assumed to be equal, under which circumstance the Bayes Factor reduces to the posterior odds. The Bayes Factor summarises evidence provided by the data in favour of one scientific theory, such that a value of $BF_{ij} = 10$ indicates that model M_i is ten times more likely to have generated the data than model M_j. Values of $BF_{ij} < 1$ are interpreted as evidence for model M_j over M_i with $BF_{ji} = 1/BF_{ij}$.

In order to evaluate BF_{ij} it is necessary to estimate the respective marginal likelihoods for each model. These densities are obtained by integrating over the parameter space, such that

$$P(Y \mid M_i) = \int P(Y \mid \theta_i, M_i) P(\theta_i \mid M_i) d\theta_i \tag{3.58}$$

where θ_i are parameters drawn from model M_i, $P(Y \mid \theta_i, M_i)$ is the likelihood function and $P(\theta_i \mid M_i)$ the prior density. The marginal likelihood is also termed the predictive probability of the data, which Kass and Raftery (1995) describes as being the probability of seeing the data that actually were observed, calculated *before* any data became available.

The integral may be evaluated analytically, although for most cases it is intractable and needs to be computed through numerical methods. Gelfand and Dey (1994) introduced an unbiased and consistent estimator of the marginal likelihood that also exploits MCMC routines. By defining $\tau(\theta)$ as any proper density such that $\int \tau(\theta) d\theta = 1$, Bayes' theorem is adapted to produce the relationship

$$\int \tau(\theta) \frac{p(\theta \mid Y, M_i) p(Y \mid M_i)}{P(Y \mid \theta, M_i) p(\theta \mid M_i)} d\theta = 1 \tag{3.59}$$

A rearrangement then produces

$$\frac{1}{p(Y \mid M_i)} = \int \frac{\tau(\theta)}{P(Y \mid \theta, M_i) p(\theta \mid M_i)} p(\theta \mid Y, M_i) d\theta \tag{3.60}$$

By taking m posterior samples from the Metropolis output, a sample estimate becomes

$$p(Y \mid M_i) = \left\{ \frac{1}{m} \sum_{k=1}^{m} \frac{\tau(\theta_k)}{P(Y \mid \theta_k, M_i) P(\theta_k \mid M_i)} \right\}^{-1} \tag{3.61}$$

Gelfand and Dey (1994) note that a natural choice for τ is a multivariate normal with means and covariance computed from the posterior samples of each parameter. In this computation, it is necessary to specify prior distributions associated with the parameters of model M_i. The prior probability is a marginal probability, and describes the knowledge of a variable before evidence is considered. Prior information about parameters may come from other data or from the subjective knowledge of experts (Kass and Raftery, 1995) however the formulation of priors is always problem-specific. If specific details about a variable are known *a priori*, such as fairly precise scientific information, this can be expressed in a Bayesian framework through the use of *informative* priors. However in many cases, little knowledge of the behaviour of model parameters may be known *a priori*, or only general information is desired to be included in the modelling process. Under these conditions it may be suitable to choose neutral or *reference* priors, interpreted as representing the views of a modeller lacking strong beliefs about the nature of model parameters, such that the posterior distribution is dominated by the likelihood.

Kass and Wasserman (1996) address the concept of selecting prior distributions by convention, as a *standard of reference*, describing the theory behind formal rules to select these priors that was developed by Jeffreys (1967). Under the algebraic convenience known as conjugacy, certain prior distributions (known as conjugates) have the same form as the posteriors. For example when a binomial distribution describes a series of samples, the Beta distribution model is a conjugate prior for the unknown proportion of *successes*. Throughout this work, the method of Gelfand and Dey (1994) is used as a method for Bayesian model selection. In response to large numbers produced, the natural logarithms of Bayes Factors are generally provided. The posterior distributions for model parameters obtained from the AM algorithm are used to generate the multivariate normal density τ that is centred on the posterior means, using the covariance of posterior samples. Conjugate priors for unknown parameters are developed for the different models, and it is ensured that each prior is proper. This provides a technique to assess the suitability of various stochastic models for describing hydrological persistence.

3.5 Summary of chapter

This chapter has discussed the various issues that underline the stochastic modelling of hydrological persistence. This phenomenon results from complex interactions between climate processes on a range of time scales and different aspects of the hydrological cycle. Persistence is an important factor in water resource planning and management, and it is vital that the characteristics of wet and dry spells are reproduced in both simulations and forecasts of hydrologic observations.

Attention is focused upon hidden Markov models (HMMs) as an appropriate method to represent persistence in hydrologic data. These models are parsimonious, and by being dominated by regimes with different mean values, can adequately describe temporal persistence within generalised climate conditions. This modelling framework has been applied to a range of scientific problems, however with regard to the representation of hydroclimatic persistence previous uses have been limited to modelling time series of annual rainfall and streamflow totals. Hydrological persistence is intimately linked with quasi-periodic climate systems such as the El Niño Southern Oscillation (ENSO), which fluctuates at an average frequency of 4 years, having an average duration of 15 months. Given these durations, it is more appropriate for persistence within hydrologic observations to be analysed using sub-annual time scales such as monthly totals, which also provide the benefit of additional data from which persistence can be identified.

In light of persistence within rainfall and streamflow time series, HMMs provide a conceptually superior model structure than alternative stationary time series models such as autoregressive moving average (ARMA) models that enjoy wide use in hydrologic simulation studies. The Bayesian methodology of the Markov chain Monte Carlo (MCMC) approach provides a useful method to evaluate parameter uncertainty in model estimation, and the Adaptive Metropolis (AM) algorithm was chosen for this work due to its computational simplicity. Posterior distributions of model parameters are utilised further through Bayesian model selection, with methods for evaluating Bayes Factors through the Gelfand-Dey estimator described. Other measures of interest from the calibration of HMMs, such as the time series of hidden state probabilities have been presented, and these are used extensively in later chapters to demonstrate the close relationship between climate states and the phases of global circulation phenomena.