Chapter 6 Problems with the calibration of Gaussian HMMs to annual rainfall

Hidden Markov models (HMMs) were introduced in Section 3.3 as a method to incorporate climatic persistence into stochastic models for hydrologic data. The spells analyses of Chapter 4 demonstrated statistically significant persistence within monthly rainfall and streamflow data from across Australia. In this chapter however, HMMs are calibrated to time series of annual rainfall totals. Annual rainfall from various capital cities have been previously analysed in the studies of Thyer (2000), Srikanthan *et al* (2002c) and Frost (2003), each of which concluded that these series demonstrated multi-decadal persistence. The accuracy of these previous studies is analysed here, with a reassessment made of the inherent persistence in these data. The modelling approach is illustrated with time series of rainfall observations from six mainland capital cities in Australia; Adelaide, Brisbane, Darwin, Melbourne, Perth, and Sydney, and also from Alice Springs. The calibration of HMMs to data from a single rainfall gauge in Sydney (Observatory Hill, BOM identification 66062) is then compared to the calibration to a series of spatially-averaged rainfall totals that correspond to the meteorological district in which this gauge is located (BOM District 66). The inclusion of this latter series clarifies persistence in the Sydney rainfall data; for the two-state Gaussian HMMs to identify true climatic persistence, evidence should be available in each of these series.

6.1 Statistics of annual rainfall series

In this chapter HMMs are calibrated to eight time series of annual rainfall totals, aggregated over different water years. Monthly data from Adelaide and Perth, which have winter-dominated rainfall regimes, are aggregated over January-December periods, whereas the summer-dominated rainfall of Darwin is aggregated over July-June periods. Annual aggregations of the monthly rainfall for Alice Springs commence in September, which has the lowest monthly average. The monthly data from Brisbane, Melbourne and the two Sydney gauges are aggregated using April-March water years, consistent with the period over which annual scale variability in ENSO is generally observed (eg Chiew *et al.*, 1998). These eight series, which represent the various rainfall regimes across this country, show considerable variability as illustrated by the statistics of Table 6.1.

Table 6.1 Statistics for the eight annual rainfall series

Rain gauge	BOM number	Period of record	Mean (mm)	Median (mm)	Standard Deviation	Skew
Adelaide	23034	1956-2001	454.0	467.4	103.2	0.15
Alice Springs	15590	1942-2001	285.2	272.7	149.4	1.83
Brisbane	40214	1860-1993	1155.6	1164.6	366.5	0.72
Darwin	14015	1870-1941	1538.8	1552.4	281.8	-0.15
Melbourne	86071	1856-1999	657.1	659.1	126.1	0.27
Perth	09034	1876-1991	868.4	854.2	162.4	0.10
Sydney	66062	1859-2001	1222.8	1203.4	330.5	0.56
District 66	66	1913-2002	1146.4	1092.2	308.0	0.87

The spatially-averaged data for District 66 is investigated alongside these various point-rainfall series, as it allows further analysis of persistence in the Sydney region. If the calibration of HMMs identifies persistence that is similar to that revealed within the point-rainfall data, evidence for such persistence representing true underlying climatic features will become more persuasive. The time series of the Sydney and District 66 samples are shown in Figure 6.1. The district-averaged series closely resembles data from the single gauge, with this relationship reinforced by the significant linear correlation of $r = 0.89$ for the 88 values over the common period of 1913-2000.

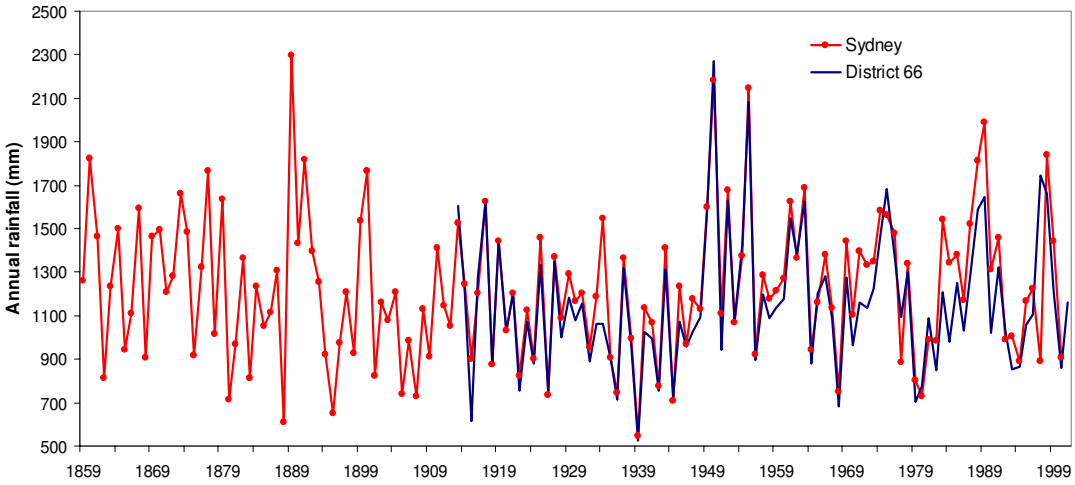

Figure 6.1 Time series of annual rainfall for the Sydney and District 66 data

An interesting feature of Figure 6.1 is the tendency for lower annual totals (with reduced variability) during a 45-50 year period ending in 1948 than periods either prior to or subsequent to this. The significance of this period of lower rainfall is investigated is investigated by first dividing the Sydney record into three sub-series: (1859-1900), (1901-1948) and (1949-2000), and using 2-sample t-tests to determine whether the means of each adjoining sub-series are statistically different. Although the positions of these breakpoints have been chosen prior to

testing, these tests show the middle period to have a sample mean (1092) that is lower at significance levels of $p = 0.010$ and $p < 0.001$ to the sample means of the first (1268) and third (1307) periods respectively. Furthermore, Levene's test (Brown and Forsythe, 1974) demonstrates that the variance of the middle period is significantly lower $(p = 0.025)$ than the preceding period. These hypothesis tests suggest that climate conditions around Sydney may have underwent a transition at the turn of last century, and again in the mid-1940s, a feature that would significantly impact upon the hydrological regime. Although there is a possibility that a feature such as this is reflective of changes in data recording methods, an apparent change in climate state of the mid-1940s is consistent with observations of previous studies as noted in Section 2.1.2. A multi-decadal dry period in the time series of annual rainfall suggests a period of hydrological persistence, precisely the feature represented by the HMM framework.

The annual rainfall series for Sydney and District 66 are shown on Gaussian probability plots in Figure 6.2. The former series plots as an approximate straight line, suggesting consistency with random draws from a Gaussian distribution. The higher skew of the latter series suggests that this might be too heavy in its upper tail to be consistent Gaussian variates, which is reinforced through the calculation of Anderson-Darling goodness-of-fit statistics. Samples of length 100 are rejected as Gaussian at a significance level of $\alpha = 0.05$ for AD statistics less than 0.754. The annual data from Sydney show a goodness-of-fit statistic of 0.632, although annual rainfall for District 66 produces a value of 1.059.

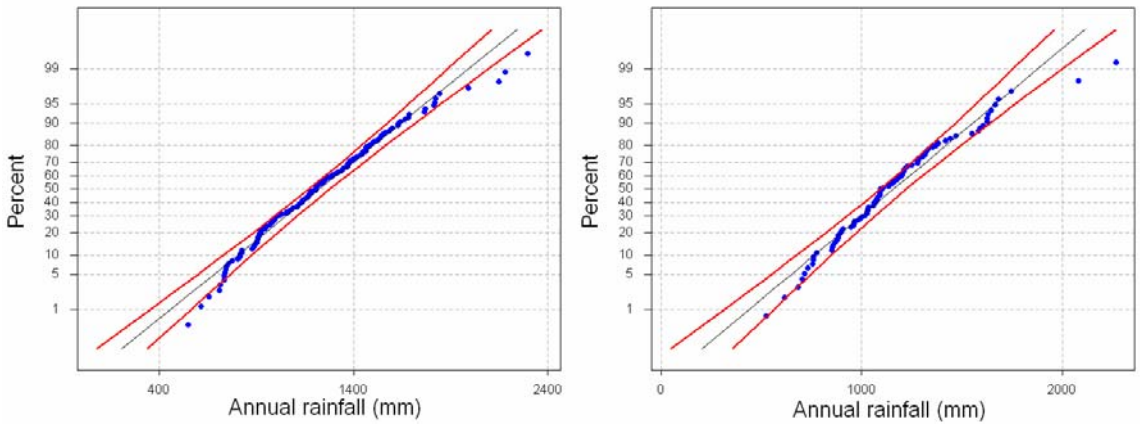

Figure 6.2 Gaussian probability plots showing the distribution of annual totals of Sydney (on left) and District 66 (on right)

The Anderson-Darling goodness-of-fit statistics from fitting Gaussian distributions to each annual rainfall series are presented in Table 6.2, with series that are not rejected as Gaussian at a 5% level shown in bold. The two series that are inconsistent with random Gaussian variates are District 66 and Alice Springs, which also have the highest skewness of these eight data series. The suitability of HMMs to describe these data is analysed in the following section.

Table 6.2 Anderson-Darling goodness-of-fit statistics from calibrating Gaussian distributions to the various annual rainfall records, with values significant at a 5% level shown in bold

	AD statistic
Adelaide	**0.411**
Alice Springs	2.764
Brisbane	**0.753**
Darwin	**0.377**
Melbourne	**0.595**
Perth	**0.376**
Sydney	**0.632**
District 66	1.059

6.2 Calibration of two-state Gaussian HMMs to annual rainfall

In the calibration of two-state HMMs to these annual rainfall series, state conditional distributions are assumed to be Gaussian, an assumption that is consistent with the approximate form of the marginal distributions. This modelling assumption was also used by Thyer and Kuczera (2000) and Frost (2003) to define long-term hydroclimatic persistence in annual Australian rainfall, and advocated by Srikanthan *et al.* (2002b) for the stochastic generation of data at this time scale. Furthermore, the assumption of Gaussian distributions for annual rainfall totals, aggregated from high frequency data, is justified in terms of the central limit theorem.

Two-state Gaussian HMMs are described by six model parameters. Maximum likelihood estimates (MLEs, $\hat{\theta}$) for these parameters are obtained from the SCE algorithm. Using $\hat{\theta}$ as the starting estimates for each of 6 independent Markov chains in the Adaptive Metropolis algorithm, posterior distributions $p(\theta | Y)$ are estimated using 10,000 samples for each chain. The convergence of the HMM simulations from calibrating this model to the annual Sydney rainfall data is analysed using the variance ratio method as described in Section 3.4.2. Table 6.3 shows the convergence diagnostic for each HMM parameter that is estimated (subscripts denote climate states), for a selection of aggregation lengths up to a maximum of 3000. By taking the average of convergence diagnostics at each aggregation length, it is apparent that this statistic approaches unity as the number of samples increase, indicating that each of the 10 sets of 6,000 simulated values are close to the target (posterior) distribution. Therefore it can be assumed that the combination of 6,000 samples taken from 10 independent Markov chains produces estimates that are representative of the stationary posterior distributions of the model parameters.

Table 6.3 Convergence diagnostics from using the variance ratio method to analyse 6,000 samples from 10 Adaptive Metropolis chains to obtain estimates of posterior distributions for a two-state Gaussian HMM calibrated to annual Sydney rainfall

	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$	$n = 2000$	$n = 2500$	$n = 3000$
P_{WD}	2.210	1.117	1.183	1.082	1.020	1.018	1.029
P_{DW}	1.523	1.635	1.261	1.228	1.143	1.080	1.062
μ_W	1.313	1.398	1.349	1.248	1.124	1.089	1.053
μ_D	1.638	1.310	1.217	1.121	1.073	1.060	1.042
σ_W	2.009	1.553	1.261	1.138	1.048	1.059	1.043
σ_D	1.894	1.258	1.149	1.031	1.048	1.060	1.066
Average	**1.764**	**1.378**	**1.236**	**1.141**	**1.076**	**1.061**	**1.049**

In order to investigate the influence of a higher number of samples being taken from a smaller number of independent Markov chains, posterior distributions for the Sydney data are now estimated from 10,000 samples from 6 chains. Table 6.4 shows convergence diagnostics for a range of aggregation lengths up to 5000. As with Table 6.3, these results indicate that by the time 10,000 samples have been obtained from each chain, the within-chain variance dominates between-chain variance, which suggests that each chain is close to the target distribution. It is clear that the Adaptive Metropolis algorithm identifies the true posterior distributions associated with the calibration of a two-state Gaussian HMM to the annual rainfall data for Sydney. Although not shown, similar convergence results to the Sydney data were obtained for each of the other annual rainfall series, using both combinations of numbers of Markov chains and samples, demonstrating the suitability of this modelling approach for investigating posterior distributions of HMM parameters.

Table 6.4 Convergence diagnostics from using the variance ratio method to analyse 10,000 samples from 6 Adaptive Metropolis chains to obtain estimates of posterior distributions for a two-state Gaussian HMM calibrated to annual Sydney rainfall

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 3000$	$n = 4000$	$n = 5000$
P_{WD}	3.555	1.374	1.036	1.014	1.059	1.023	1.011
P_{DW}	2.313	1.205	1.292	1.115	1.096	1.056	1.028
μ_W	1.620	1.912	1.030	1.156	1.096	1.042	1.022
μ_D	1.243	1.282	1.135	1.055	1.092	1.058	1.030
σ_W	2.461	1.196	1.082	1.119	1.027	1.072	1.066
σ_D	2.577	1.210	1.173	1.075	1.047	1.047	1.028
Average	**2.295**	**1.363**	**1.125**	**1.089**	**1.069**	**1.049**	**1.031**

The posterior distributions for the HMM transition probabilities estimated from each rainfall series are summarised in Table 6.5 with posterior medians and 90% credibility limits. These posteriors are obtained from 6,000 samples taken from 10 chains. The posterior median

provides a superior point estimate for posterior distributions than the mean, as it remains unaffected by extreme values in the posteriors. In a two-state Markov chain, the expected residence time in each model state is calculated as the inverse of the probability of moving out of that state. Using the posterior medians as estimates for the two transition probabilities, the Sydney data have an expected residence time of 5 years in a dry state but less than 2 years for the wet state. Such results are inconsistent with long-term persistence, and suggest that this model fails to simulate the multi-decadal dry epoch suggested by the time series of Figure 6.1.

Table 6.5 Medians of posterior distributions for HMM transition probabilities, with 90% credibility intervals from the calibration of two-state Gaussian HMMs

	P_{WD}	P_{DW}
Adelaide	**0.605** (0.068, 0.967)	**0.035** (0.003, 0.806)
Alice Springs	**0.350** (0.097, 0.727)	**0.088** (0.030, 0.208)
Brisbane	**0.663** (0.117, 0.975)	**0.090** (0.019, 0.441)
Darwin	**0.390** (0.022, 0.939)	**0.538** (0.033, 0.956)
Melbourne	**0.668** (0.114, 0.963)	**0.313** (0.008, 0.911)
Perth	**0.376** (0.008, 0.944)	**0.504** (0.010, 0.942)
Sydney	**0.530** (0.139, 0.922)	**0.200** (0.030, 0.880)
District 66	**0.516** (0.194, 0.908)	**0.150** (0.038, 0.521)

The calibration results for the other annual rainfall series also fail to show evidence of long-term persistence. Results for the Adelaide, Alice Springs and Brisbane data show estimates for P_{DW} that are much lower than P_{WD}, which suggests that a majority of years are in a dry climate state. This result does not reflect strong two-state persistence in these annual rainfall series, especially with the Adelaide and Brisbane data, which show an average wet state duration of less than 2 years. Furthermore the P_{WD} estimates for these series are associated with wide credibility intervals, indicating that the length of available data in these series may be too short for accurate parameter estimation. Although transition probability estimates for the other series are closer in magnitude, these are too high to reflect long-term persistence and have 90% credibility intervals that are spread over much of the (0, 1) interval.

The posteriors for parameters of the HMM conditional distributions from each annual rainfall series are summarised in Table 6.6. These results tend to reflect transition probability estimates, such as the parameters of the wet state distributions for the Adelaide and Brisbane data being

poorly defined. For the other series, the upper bound of the 90% credibility interval for the dry state means are generally lower than the lower bound for the wet state mean, indicating that these posteriors are well-separated. This suggests that the Gaussian HMM estimates two distinct conditional distributions in these annual rainfall series, although transition probability estimates show little evidence for multi-decadal persistence. Estimates for the HMM parameters are similar between the Sydney and District 66 series, which highlights a similar pattern of persistence between these two series.

Table 6.6 Medians of posterior distributions for parameters of HMM conditional distributions, with 90% credibility intervals

	μ_W	μ_D	σ_W	σ_D
Adelaide	**753.3** (454, 2630)	**446.1** (402, 475)	**274.9** (52, 898)	**103.3** (77, 134)
Alice Springs	**520.4** (364, 700)	**239.5** (219, 262)	**219.8** (140, 368)	**79.9** (65, 101)
Brisbane	**1490.6** (1200, 2127)	**1098.8** (1028, 1159)	**496.1** (331, 801)	**303.3** (258, 358)
Darwin	**1589.8** (1511, 1778)	**1469.1** (1105, 1565)	**270.2** (127, 442)	**285.7** (78, 512)
Melbourne	**715.5** (656, 953)	**631.9** (590, 664)	**131.5** (68, 216)	**108.3** (67, 138)
Perth	**895.5** (853, 1811)	**828.5** (596, 879)	**162.4** (109, 327)	**156.3** (88, 410)
Sydney	**1425.7** (1238, 1848)	**1143.7** (979, 1218)	**362.8** (274, 520)	**271.2** (189, 338)
District 66	**1442.3** (1208, 1860)	**1066** (997, 1148)	**366.6** (251, 533)	**223.9** (172, 287)

Posterior distributions for each of the HMM parameters estimated from the Sydney rainfall series are shown as histograms in Figure 6.3. The large uncertainty around each parameter is evident, particularly for the two transition probabilities, which make it difficult to interpret significant two-state persistence.

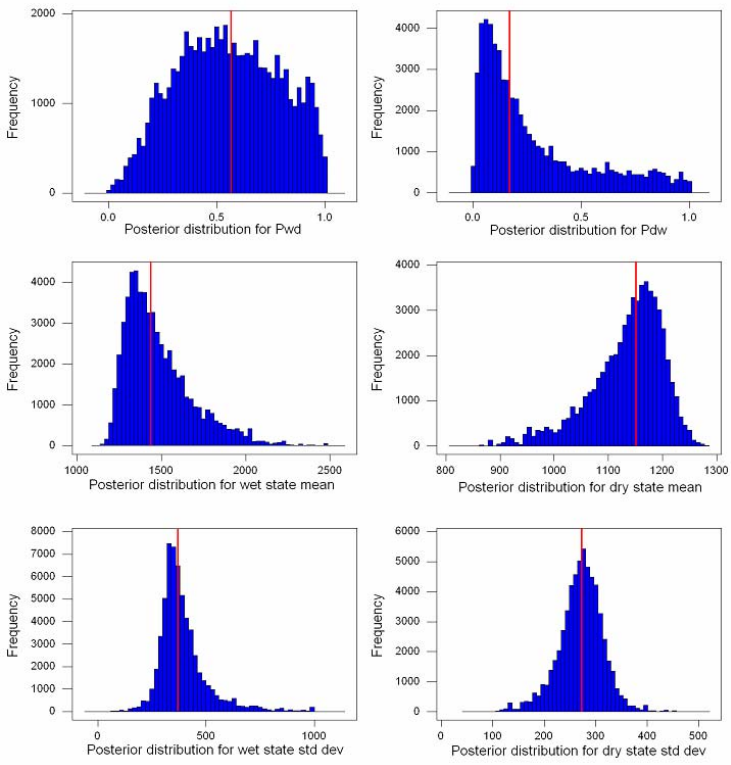

Figure 6.3 Posterior distributions from the calibration of a two-state Gaussian HMM to the annual rainfall data for Sydney

The MLEs of the HMM parameters are used to generate the probability of being in a wet state for each year in the Sydney data, shown alongside the time series of annual rainfall totals in Figure 6.4. These two series are strongly correlated $(r = 0.80, p < 0.001)$ and display the intuitive result of years with high annual totals also showing high probabilities of having been generated from a wet model state. An interesting feature of the state series is the high probability that the model remains within an extended dry state from 1901 until a sharp transition to a wet state in the late 1940s, corresponding to the sharp increase in annual rainfall recorded at the gauge during 1948/9. Although transition probability estimates suggest that a multi-decadal dry epoch is unlikely, this feature is still apparent in the posterior state series. This result was exploited by Thyer and Kuczera (2000) as support for the hypothesis of the HMM being suitable for modelling long-term persistence in annual rainfall time series. The strongest correlation between annual rainfall data and HMM state series is from Alice Springs $(r = 0.82, p < 0.001)$.

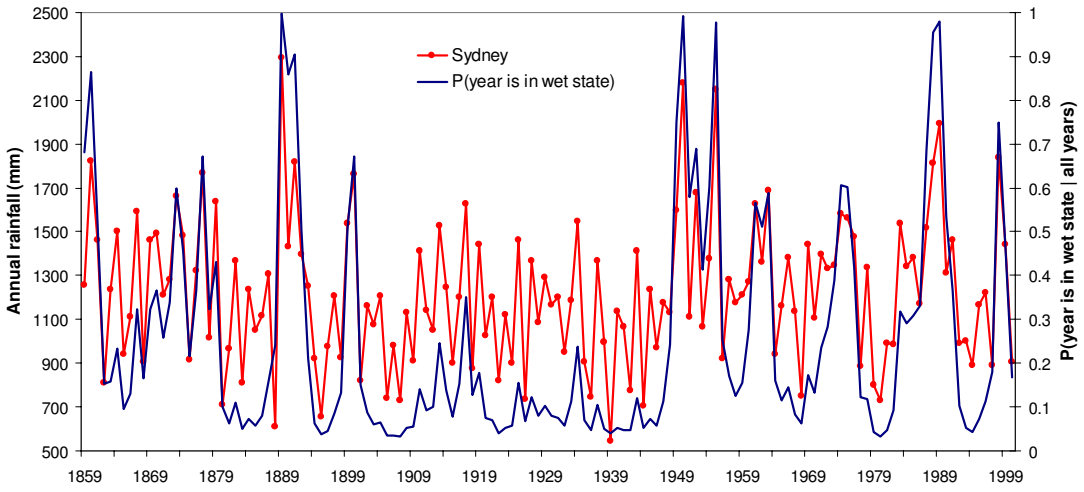

Figure 6.4 Time series of annual rainfall for Sydney shown alongside the posterior state series associated with parameter MLEs after the calibration of a two-state Gaussian HMM

The relationship between persistence in the Sydney and District 66 series is shown clearly through the posterior state series derived from parameter MLEs for each series in Figure 6.5. These state series have a linear correlation of $r = 0.84$ ($p < 0.001$), which shows that over the period 1913-2001, two-state Gaussian HMMs identify similar wet and dry persistence. The two state series shown a sharp transition to a wet state in 1948/9, and identify comparable climate state features after this time. The district-averaged rainfall series is derived from various point rainfall records, so the fact that posterior state probabilities from this series are similar to those from the single Sydney gauge record suggests that the HMM identifies true climatic features.

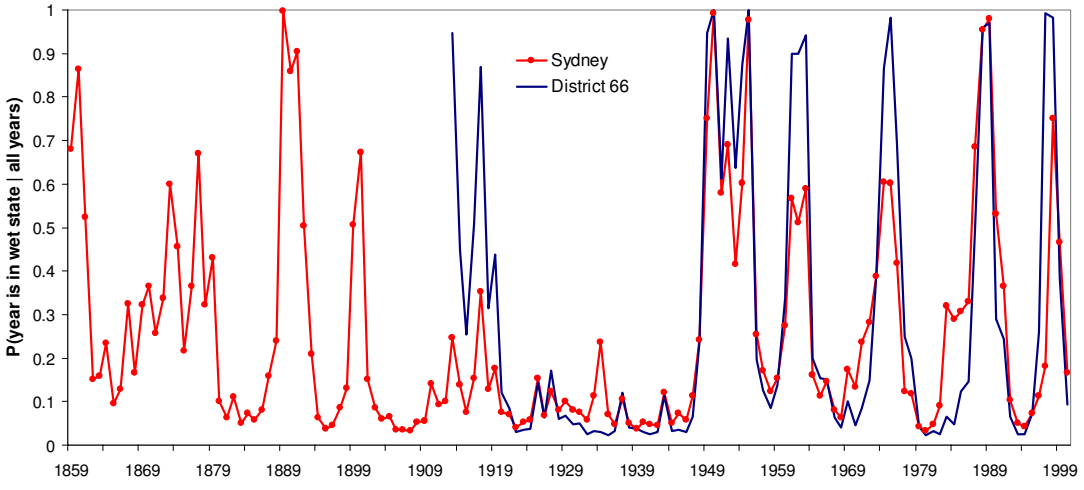

Figure 6.5 Posterior state series associated with parameter MLEs after the calibration of a two-state Gaussian HMM to the annual rainfall for both Sydney and District 66

The posterior state series is a useful output from the HMM calibration, as it provides a direct representation of persistence in each model state. However by only generating this series with

parameter MLEs, an evaluation of the variability around posterior state probabilities is prevented. It is possible though to take advantage of the Bayesian framework that was adopted for parameter estimation and evaluate the posterior distribution for the state probability series. This development of the Bayesian paradigm was not utilised by Thyer (2000) or Frost (2003), or indeed in other publications, yet it is vital for the accurate interpretation of the calibrated HMM. This presents a new approach to analysing the efficacy of HMM calibration, and provides a means to revisit the results of previous studies into possible persistence within annual rainfall data across Australia.

In order to generate the posterior distribution for the state series, each sample of model parameters from the Metropolis output is used to generate a series of state probabilities. The 60,000 state probability values at each time step are ranked to produce a posterior median state probability (with 90% credibility interval) for the entire series. The posterior state series is a manifestation of HMM parameters, and it is evident that narrower posterior distributions will realise narrower credibility intervals around the posterior state series. The series of posterior median state probabilities, termed the *median state series*, is interpreted here to be the best representation of these state probabilities. In light of this new development of analysing HMM state series, median state series are used in preference to the state series derived from MLEs of HMM parameters throughout this work.

For the annual rainfall series from Sydney, the median state series is highly correlated to the MLE state series that was shown in Figure 6.4 $(r = 0.97)$. However by showing the median state series for this series alongside its 90% credibility interval in Figure 6.6, it is clear that the HMM fails to identify a clear sequence of wet and dry states. The credibility interval tends to encompass most of the (0, 1) interval regardless of the median state probabilities. The wide credibility interval around the median state series in Figure 6.6 shows that by accounting for parameter uncertainty when evaluating posterior state probabilities, there is little certainty as to which model state generated the annual totals. The median state series suggests a high probability for an extended dry state remaining until 1948/9, which was interpreted as a persistent dry epoch. By observing parameter uncertainty, it is clear that little information can be inferred from the calibration of a two-state Gaussian HMM to the annual rainfall series of Sydney. Furthermore, the credibility bounds around posterior state probabilities for the district-averaged data show a similar width to those shown in Figure 6.6. As inference based on the calibration of two-state HMMs is compromised in periods of increased uncertainty in state probabilities, conclusions about long-term persistence in this data series based solely upon the MLE state probabilities are inappropriate. It is apparent that poorly estimated transition probabilities such as those described in Table 6.5 lead to poorly estimated posterior state probabilities.

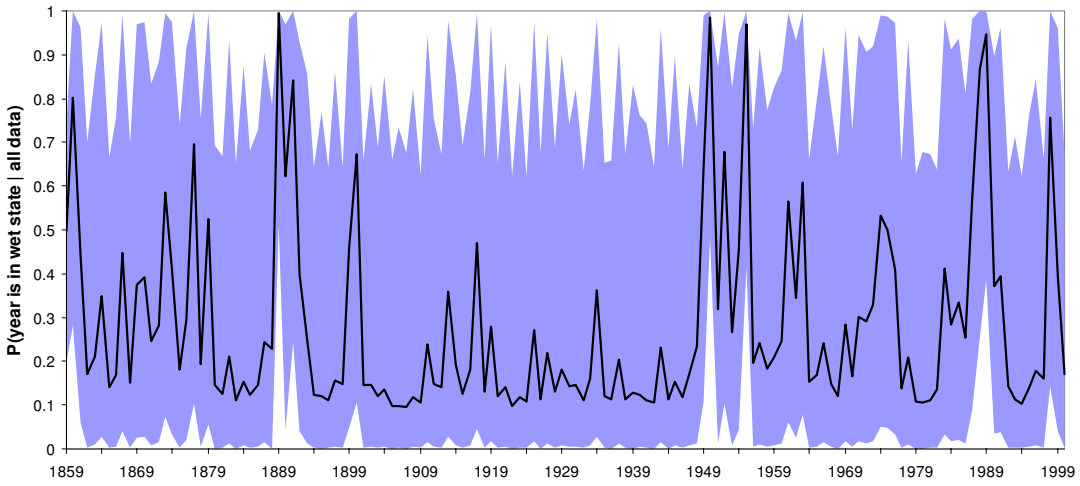

Figure 6.6 Median state series and 90% credibility interval from the calibration of a two-state Gaussian HMM to the annual rainfall for Sydney

The calibration of two-state Gaussian HMMs to annual rainfall data from the various capitals produces state series that fail to reflect significant two-state persistence. However with the calibration of a two-state Gaussian HMM to the annual rainfall data for Alice Springs producing a median state series that was strongly correlated to the time series of annual rainfall, it is interesting to see in Figure 6.7 that this series also has tighter credibility bounds than the Sydney data. It is possible that this annual rainfall series, being recorded in an arid zone, more clearly illustrates fluctuations between persistent climate states than the Sydney record can.

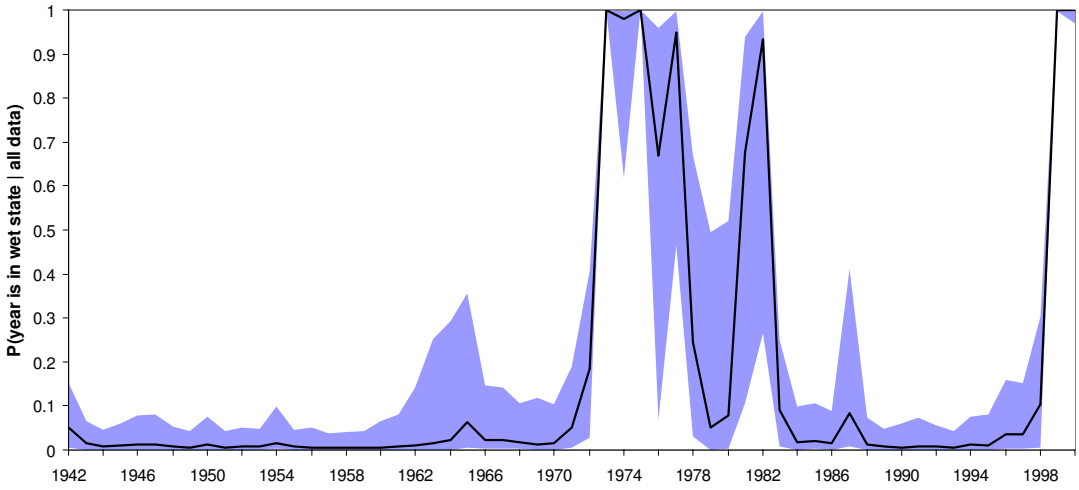

Figure 6.7 Median state series and 90% credibility interval from the calibration of a two-state Gaussian HMM to the annual rainfall for Alice Springs

The posterior state series shown in Figure 6.6 indicates that there is little evidence for two-state persistence in the annual rainfall data of Sydney following the calibration of a two-state Gaussian HMM. This result can be attributed to either the absence of hydroclimatic persistence

at an annual time scale in the Sydney data, or indeed that the available data is of insufficient length to provide a statistically significant description of persistence that is present. In order to investigate the possibility of the latter, two-state Gaussian HMMs are calibrated to simulated data in the following section.

6.3 Calibration of two-state Gaussian HMMs to simulated data

The wide credibility bounds around HMM state probabilities for the time series of annual Sydney rainfall suggests at first that the Gaussian HMM is unsuitable for these data. However with a length of only 142 years, there is the danger that the observed series is too short for the two-state Gaussian HMM to estimate adequately the model parameters. In order to determine whether this model identifies correct model parameters in a series of this length, a time series is first simulated from a two-state Gaussian HMM with parameters equal to the posterior medians shown in Table 6.5. After simulating this series of 142 values, a two-state Gaussian HMM is then calibrated using the same method as described earlier. Model parameters are estimated and then compared with the values that were used in simulation. This process is then repeated with a series length of 1000, with posterior state series generated from both simulations.

The posterior estimates from these two simulations are summarised in Table 6.7, and it is evident that true parameter values are contained within 90% credibility intervals for both simulations. The widths of these credibility intervals however are large considering the correct model was fitted. This is particularly apparent for the two transition probabilities of the shorter simulation. It is important to note though that the accuracy of transition probability estimates increase dramatically for the longer simulations, with a contraction in credibility intervals.

Table 6.7 Parameter estimates for simulated series of lengths 142 and 1000, with posterior median and 90% credibility intervals shown

Parameter	Simulated value	Estimates for simulation length 142	Estimates for simulation length 1000
P_{WD}	0.513	**0.528** (0.068, 0.947)	**0.519** (0.254, 0.788)
P_{DW}	0.200	**0.457** (0.014, 0.951)	**0.198** (0.077, 0.477)
μ_W	1426	**1248** (1156, 1597)	**1437** (1308, 1696)
μ_D	1144	**1090** (947, 1178)	**1171** (1124, 1209)
σ_W	363	**294** (200, 444)	**366** (295, 421)
σ_D	271	**253** (136, 329)	**266** (220, 292)

The effect of wide posteriors on the estimation of posterior state probabilities for the shorter simulation is clearly displayed in Figure 6.8. This figure shows the median posterior state series, together with 90% credibility limits, with circles representing the state sequence of the original simulation. Posterior medians are centred on a value of 0.5 over the entire series, which demonstrates an inability to identify wet and dry states. A superior HMM calibration would be characterised by median state probabilities that are closer to either 0 or 1, hence Figure 6.8 suggests a very poor fit to the simulated series. The credibility interval around the median state series for the shorter simulation has a similar width to the interval obtained for the original Sydney series. Although this series was simulated with known state transitions, the calibration results suggest that a length of only 142 values is possibly too short to obtain meaningful inference.

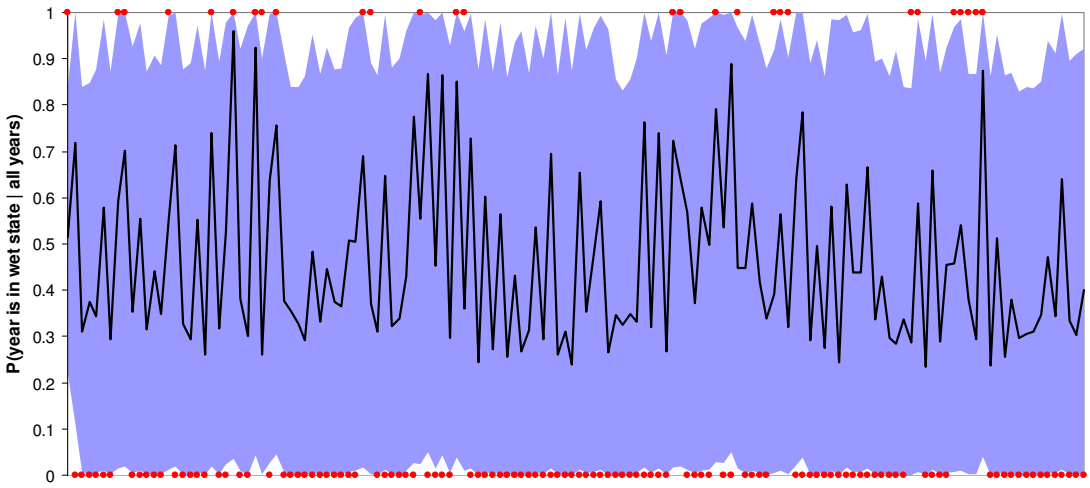

Figure 6.8 Median state series and 90% credibility interval from the calibration of a two-state Gaussian HMM to series of length 142 simulated from a two-state Gaussian HMM

The improvement in model calibration that is achieved with a longer data series is now demonstrated through the posterior state series associated with the longer simulation, shown in Figure 6.9. Posterior state probabilities for a sequence of only 142 values are shown here for the purpose of comparison to Figure 6.8. This median state series is characterised by credibility intervals that are significantly narrower than obtained for the shorter series. The average width of the 90% credibility limits over this sample length is reduced from 0.922 in the shorter series to 0.516 in this series, indicating a more precise HMM calibration.

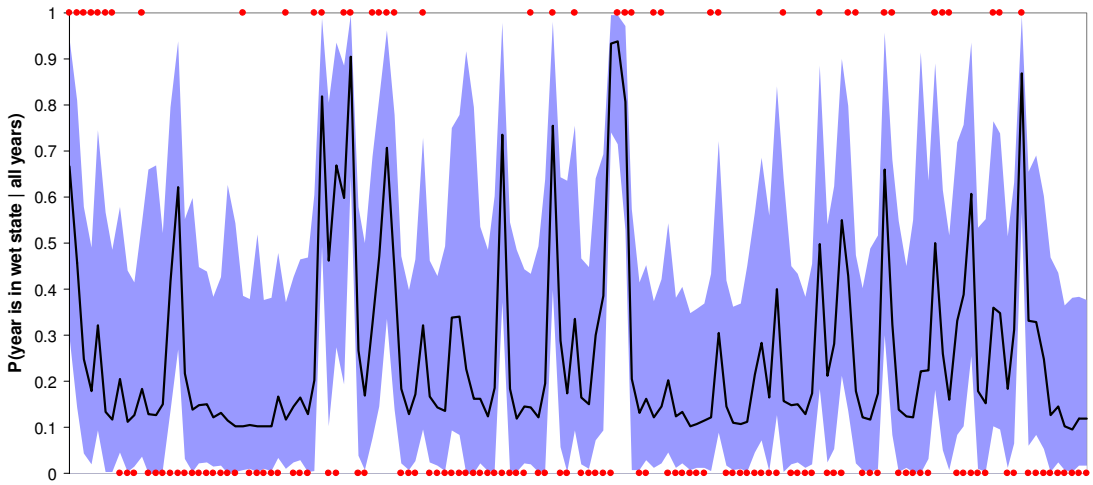

Figure 6.9 Median state series for 142 values and 90% credibility interval from the calibration of a two-state Gaussian HMM to series of length 1000 simulated from a two-state Gaussian HMM

These two simulations show the inability of a two-state Gaussian HMM to identify clear transitions between model states within time series of similar length to annual rainfall series. Although these series have been generated with explicit persistence within states, the available information is inadequate to allow for accurate parameter estimation. As a result, it is clear that any conclusions made about the persistence structure of this observed series may therefore be misleading. Consequently monthly records may offer a more appropriate source of information for identifying hydrological persistence with a HMM, providing that a suitable parametric form of the model can be established. It is worth noting here that hidden states can be identified in the absence of two-state persistence, where such "states" represent the components of a mixture distribution. In the following section, the relationships between HMM transition probabilities under these conditions are investigated.

6.4 HMMs degenerating to mixture distributions

A mixture of two normal distributions can often closely approximate a skewed marginal distribution. A HMM includes a mixture as the special case of state transitions being independent of the current state, which in the two-state case is represented as

$$P_{WD} = P_{DD} \text{ and } P_{DW} = P_{WW} \quad \textbf{(6.1)}$$

The stationary distributions of the Markov chain, P_{wet} and P_{dry}, which are equivalent to the proportion of time that the HMM spends in the wet and dry states respectively, satisfy

$$\begin{bmatrix} P_{wet} & P_{dry} \end{bmatrix} \times \begin{bmatrix} P_{WW} & P_{WD} \\ P_{DW} & P_{DD} \end{bmatrix} = \begin{bmatrix} P_{wet} & P_{dry} \end{bmatrix} \quad \textbf{(6.2)}$$

Such that

$$P_{wet} = \frac{P_{DW}}{P_{WD} + P_{DW}} \text{ and } P_{dry} = \frac{P_{WD}}{P_{WD} + P_{DW}} \quad \textbf{(6.3)}$$

and $P_{wet} + P_{dry} = 1$. If $P_{DW} = P_{WW}$, this becomes

$$P_{wet} = P_{DW} = P_{WW} \quad \textbf{(6.4)}$$

Therefore, this condition is equivalent to

$$P_{WD} + P_{DW} = 1 \quad \textbf{(6.5)}$$

If state transitions are independent of the current state, the construct of distinct climate states is redundant, and the model can be interpreted as merely a mixture distribution. If there is a tendency for the model to persist in either state, the sum of P_{WD} and P_{DW} will be less than unity. A sum of transition probabilities greater than 1 corresponds to a tendency for weak persistence within states. A Bayesian credibility interval for the sum of transition probabilities can be used as an indicator for persistence. If the upper limit of the 90% credibility interval is less than 1, evidence for hydrological persistence is claimed, yet if the interval includes 1, there is no convincing evidence to dismiss the possibility of the HMM being merely a mixture. Alternatively, a tendency towards persistence can be expressed as the sum of self-transition probabilities ($P_{WW} + P_{DD}$) being greater than unity; a relationship that can be generalised to more than two model states. This is discussed further in Section 7.3.

The model calibration for the various annual rainfall series is now examined for evidence of significant two-state persistence. The posterior distributions for the two transition probabilities are used to obtain a posterior distribution for $P_{WD} + P_{DW}$ in each case. Figure 6.10 is a histogram of the posterior distribution for the sum of transition probabilities from the calibration of a two-state Gaussian HMM to the annual rainfall data for Sydney. This posterior has a median of 0.830, but with a wide 90% credibility interval (0.336, 1.401) that includes 1, the possibility of this model degenerating to a normal mixture cannot be rejected. This reinforces the suggestion from the posterior state series that a two-state Gaussian HMM fails to identify persistence in this annual rainfall series.

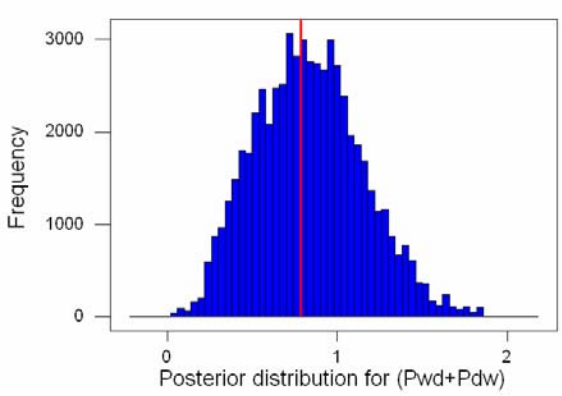

Figure 6.10 Posterior distribution for the sum of transition probabilities from the calibration of a two-state Gaussian HMM to the annual rainfall data for Sydney with median shown

The posterior distributions for the sums of transition probabilities from the calibration of two-state Gaussian HMMs to the various annual rainfall series are summarised in Table 6.8. These results demonstrate that the Alice Springs data is the only series that shows statistically significant evidence for two-state persistence. These conclusions are consistent with inference gained from the posterior state series associated with each rainfall series that suggested little evidence for two-state persistence in any of these series apart from Alice Springs. In light of these results, it is unlikely that HMMs with more than two conditional model states would identify significant persistence in annual rainfall data, and subsequently these are not calibrated.

Table 6.8 Posterior medians with 90% credibility intervals for the sums of transition probabilities from the calibration of two-state Gaussian HMMs to the various annual rainfall series

	$P_{WD} + P_{DW}$
Adelaide	**0.749** (0.127, 1.425)
Alice Springs	**0.449** (0.162, 0.862)
Brisbane	**0.806** (0.169, 1.201)
Darwin	**0.919** (0.354, 1.588)
Melbourne	**0.995** (0.391, 1.590)
Perth	**0.912** (0.217, 1.548)
Sydney	**0.830** (0.336, 1.401)
District 66	**0.666** (0.320, 1.176)

6.5 Summary of chapter

Throughout a number of previous studies, HMMs have been described as suitable models for persistence in time series of annual rainfall in Australia. In this chapter, a more thorough statistical analysis of two-state Gaussian HMMs that are calibrated to annual rainfall across Australia has presented little evidence for two-state persistence in these data. A simple statistical test, based on estimates of HMM transition probabilities, has been developed to estimate the significance of persistence. This test reinforces previous results that indicate either an absence of hydrological persistence at an annual scale, or that annual data is of insufficient length to identify this feature accurately. Through simulation studies, it was shown that fluctuations between states are difficult to identify within time series of equivalent length to annual rainfall series.

This chapter has shown that a two-state Gaussian HMM is unable to identify two-state climatic persistence within the time series of annual rainfall from cities across Australia. In their investigation of long-term persistence in Australian point rainfall, Thyer and Kuczera (2000) suggested that the calibration of a two-state Gaussian HMM to these data supported their modelling assumptions and identified strong persistence. The time series of annual rainfall for Sydney suggests a multi-decadal period of below-average values. However the calibration of a two-state Gaussian HMM to this series was unable to provide credible evidence for a continuous dry state existing over this period. Although this result may be due to the short duration of the annual record, the modelling assumption of the two-state Gaussian HMM may also be inappropriate for observing persistence at an annual scale. As a result it is pertinent to investigate whether the calibration of two-state HMMs to time series of higher frequency data, such as monthly values, is an acceptable alternative. Persistence in monthly rainfall is investigated in the following chapter.

Chapter 7 Using parametric HMMs to identify persistence in monthly rainfall

The previous chapter showed that calibrations of Gaussian HMMs are unable to provide statistically significant evidence for persistence in the annual rainfall records of Australia. Probable causes for such observations are either the lack of significant climatic persistence at this scale, or indeed that there is insufficient data. An alternative approach to the investigation of annual persistence that examines both these possibilities is to calibrate HMMs to time series of monthly rainfall data. Sampling hydrologic data at monthly time scales not only provides more information concerning interactions between climate processes and rainfall observations, but was shown in Chapter 5 to identify significant persistence in spatially-averaged rainfall. In previous studies such as Thyer and Kuczera (2000) and Frost (2003), the calibration of HMMs focused solely upon rainfall observations aggregated over annual periods. HMMs have not previously been described as appropriate time series models for monthly rainfall, and their suitability is investigated in this chapter. Through monthly data from the sites introduced in Chapter 6, a clear comparison of persistence observed at both frequencies is achieved.

7.1 Statistics of monthly rainfall data

In this chapter, two-state HMMs are calibrated to time series of monthly totals from the seven point rainfall and one spatially-averaged rainfall series. Statistics for these series of monthly observations are shown in Table 7.1. The monthly record from each gauge has higher skew than the corresponding series of annual totals, a characteristic that is shared by many monthly rainfall records. These statistics demonstrate large variability in the rainfall regime of Australia.

Table 7.1 Statistics for the eight series of monthly rainfall totals

Rain gauge	BOM number	Period of record	Mean (mm)	Median (mm)	Standard Deviation	Skew
Adelaide	23034	1956-2001	37.8	33.1	28.4	0.84
Alice Springs	15590	1942-2001	24.0	9.7	40.3	3.48
Brisbane	40214	1860-1993	96.1	66.8	99.4	2.78
Darwin	14015	1870-1941	128.3	53.4	162.1	1.31
Melbourne	86071	1856-1999	54.8	49.3	33.6	1.20
Perth	09034	1876-1991	72.4	46.5	74.8	1.23
Sydney	66062	1859-2001	102.0	74.4	93.3	1.92
District 66	66	1913-2002	95.0	69.1	87.6	2.13

The monthly time series from Sydney and District 66 are highly correlated over the period 1913-2001 ($r = 0.96$), demonstrated by the 10-year time series plot shown in Figure 7.1 that is indicative of the entire series. The close relationship of these time series reflects the association between the series of annual rainfall totals from these two gauges described in Figure 6.1.

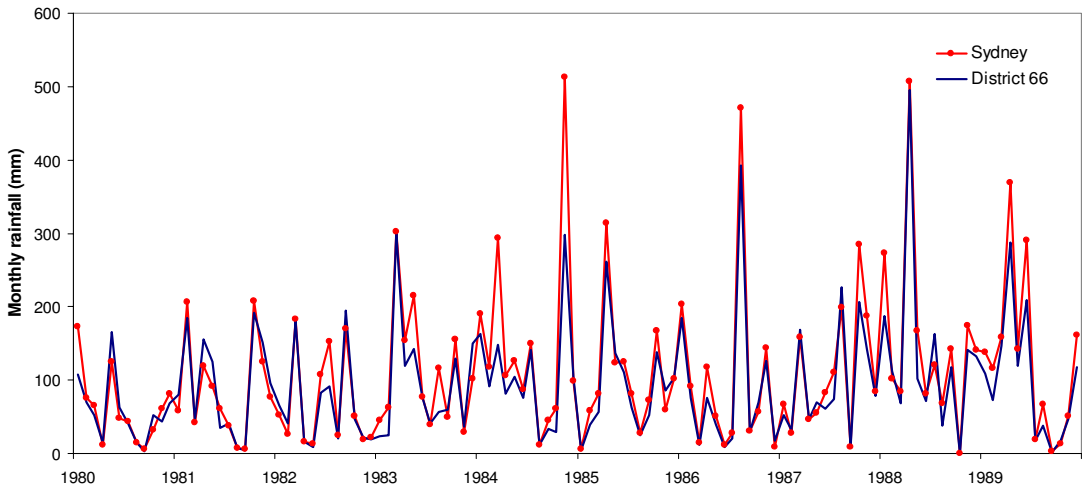

Figure 7.1 Time series of monthly rainfall observations from Sydney and District 66 over the period 1980-1989

The seasonal nonstationarity in these two data sets is demonstrated in Figure 7.2, with mean monthly totals alongside historical 90% limits around these means. The two time series display a minimum monthly average rainfall in September, with maxima recorded in March. The large positive skews of the long-term statistics in Table 7.1 are evident in the 90% limits that tend towards values higher than the mean.

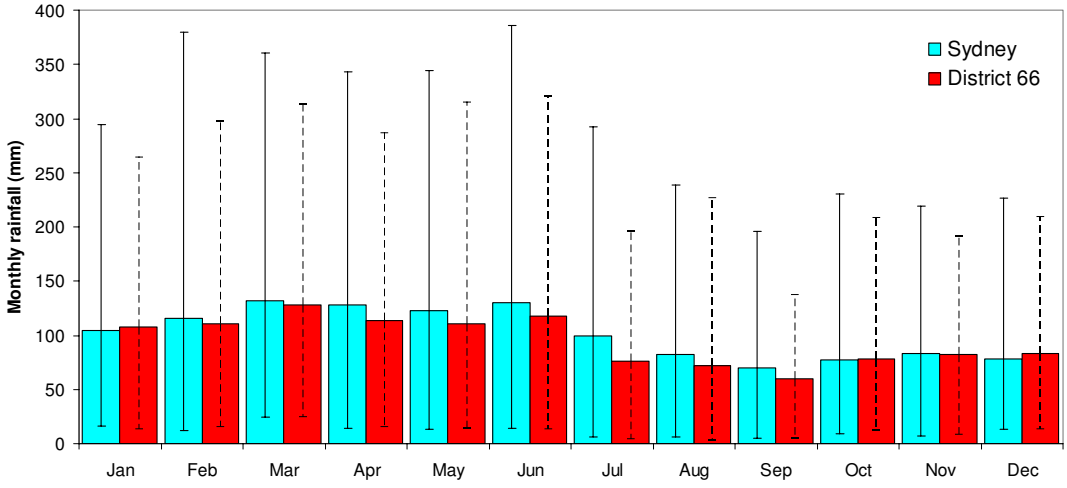

Figure 7.2 Mean monthly rainfall totals with bars showing 90% of monthly values for Sydney and District 66

In order to analyse persistence within these monthly rainfall records, it is necessary to first remove the seasonal nonstationarity. In a similar approach to that used with the district-averaged data in Chapter 4, this is achieved through first subtracting the arithmetic mean for samples in each month from the original values. Through dividing by the sample standard deviation for that month, this process produces a time series of zero mean and unit variance.

The eight time series of deseasonalised monthly values are then scaled to produce series of scaled positive anomalies. A convenient method for this scaling is to multiply deseasonalised values by the observed standard deviation from a particular month and then add the mean value from this same month. By plotting the scaled positive anomalies on lognormal probability plots, it is apparent that deseasonalised values from many of these both series approximate random draws from lognormal distributions. This is demonstrated by the Anderson-Darling goodness-of-fit statistics shown in Table 7.2, which show that lognormal distributions are more appropriate (providing lower AD statistic) than either Gaussian or gamma distributions for each monthly rainfall series.

Table 7.2 Anderson-Darling goodness-of-fit statistics from the calibration of various distributions to the scaled positive anomalies of each monthly rainfall record

	Gamma distribution	Lognormal distribution	Gaussian distribution
Adelaide	1.16	0.79	7.48
Alice Springs	24.74	13.40	57.91
Brisbane	15.45	2.92	64.89
Darwin	20.86	14.49	33.85
Melbourne	1.46	1.17	24.99
Perth	9.02	2.43	26.63
Sydney	9.06	0.90	67.31
District 66	10.22	0.43	45.91

As an example of this scaling procedure, the Sydney and District 66 data are scaled so that they have a mean (104.31) and standard deviation (77.44) equal to those of the observed January data for Sydney. The distributions of these scaled variates are displayed on lognormal probability plots in Figure 7.3. Although these distributions deviate slightly from straight lines at the tails, such deviation is insufficient to suggest that lognormal distributions are inadequate.

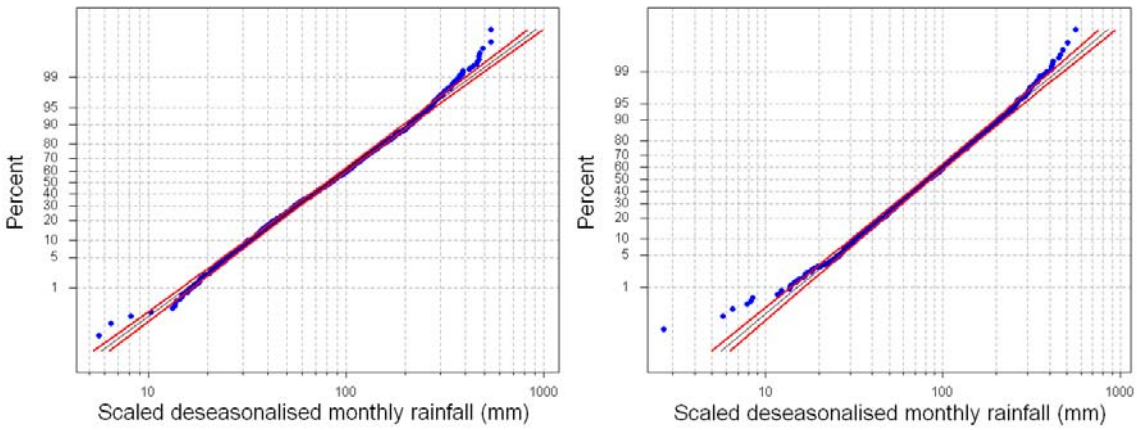

Figure 7.3 Lognormal probability plots showing scaled deseasonalised monthly rainfall for Sydney (left) and District 66 (right)

The scaled deseasonalised series for Sydney and District 66 are now compared in Figure 7.4 over the same 10-year period that was used in Figure 7.1, and once again their close relationship is apparent. Persistence within the eight deseasonalised monthly rainfall series is investigated in the following section using parametric HMMs.

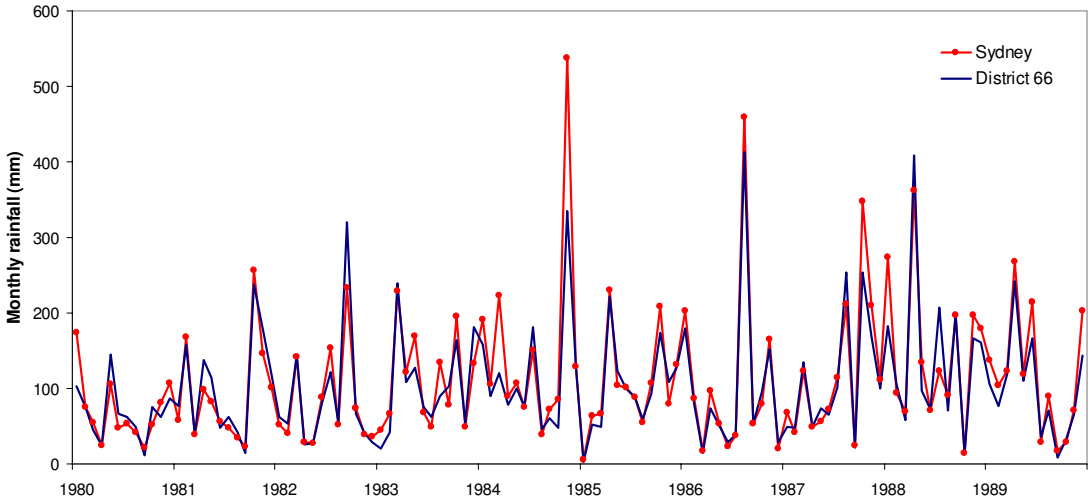

Figure 7.4 Time series of scaled deseasonalised monthly rainfall for Sydney and District 66 over the period 1980-1989

7.2 Calibration of two-state HMMs to monthly rainfall

7.2.1 Calibration of two-state lognormal HMMs

Results in the previous section indicated that the conditional state distributions for deseasonalised monthly rainfall series approximate lognormals. These observations can be exploited in the initial calibration of HMMs. By assuming that the physical processes that

distribute monthly rainfall anomalies lognormally are conditioned upon underlying climate states, the HMM state conditional distributions are also assumed to comprise random lognormal variates. Furthermore as shown in Section 6.4, an assumption of Gaussian conditional distributions may result in the calibration of a Gaussian mixture, which weakens the supposition of hydroclimatic persistence. The calibration of a two-state lognormal HMM requires six model parameters to be estimated; $P(y_t \mid s_t^j)$ is lognormal such that $P(\ln(y_t) \mid s_t^j) \approx N(\mu_j, \sigma_j^2)$. Once again the SCE algorithm is used to obtain MLEs for these parameters, with posteriors inferred through 6,000 samples from 10 Markov chains in the Adaptive Metropolis algorithm. Variance ratio diagnostics indicate that these numbers of samples are sufficient to obtain accurate estimates of the posterior distributions for each of these monthly rainfall series.

The posterior distributions for the HMM transition probabilities estimated from each rainfall series are summarised in Table 7.3 by posterior medians and 90% credibility limits. These results show that Adelaide, Darwin and Sydney are the only series that have transition probability estimates that are approximately equal.

Table 7.3 Medians of posterior distributions for HMM transition probabilities, with 90% credibility intervals from the calibration of two-state lognormal HMMs

	P_{WD}	P_{DW}
Adelaide	**0.198** (0.041, 0.588)	**0.309** (0.049, 0.656)
Alice Springs	**0.915** (0.731, 0.992)	**0.079** (0.047, 0.136)
Brisbane	**0.028** (0.015, 0.051)	**0.660** (0.445, 0.911)
Darwin	**0.208** (0.167, 0.256)	**0.403** (0.336, 0.475)
Melbourne	**0.090** (0.021, 0.299)	**0.575** (0.228, 0.914)
Perth	**0.049** (0.033, 0.075)	**0.824** (0.684, 0.923)
Sydney	**0.343** (0.214, 0.490)	**0.379** (0.215, 0.506)
District 66	**0.031** (0.010, 0.093)	**0.654** (0.413, 0.954)

As suggested in Section 6.2, HMM calibrations identifying large differences in estimates of P_{WD} and P_{DW} diminish the possibility of these data showing long-term hydroclimatic persistence. Intuitively with monthly data, residence times in the two model states should be similar, as dominant climate modes such as ENSO persist in each of their phases for approximately equal durations. The posterior medians of transition probabilities from the monthly Brisbane data suggest residence times of approximately 35 months in a wet state, yet

only 1.5 months in a dry state. Similar features are apparent in the calibrations of two-state lognormal HMMs to the deseasonalised monthly data for Alice Springs, Melbourne, Perth and District 66. These results suggest that two-state lognormal HMMs are inappropriate models for monthly-scale persistence in these data.

The full posteriors from the calibration of a two-state lognormal HMM to the deseasonalised monthly rainfall series for Sydney are shown in Figure 7.5 as histograms. The estimates of the two monthly transition probabilities have much tighter credibility intervals than was achieved through the calibration of a two-state Gaussian HMM to the annual data. Posterior medians of transition probabilities suggest similar residence times in the two model states, which are on average between 2.5–3 months. The two-state persistence identified by the lognormal HMM is clearly of a frequency that cannot be identified with annual rainfall totals. This calibration identifies conditional distributions with well-separated means, suggesting that the monthly time scale appropriately describes hydroclimatic persistence.

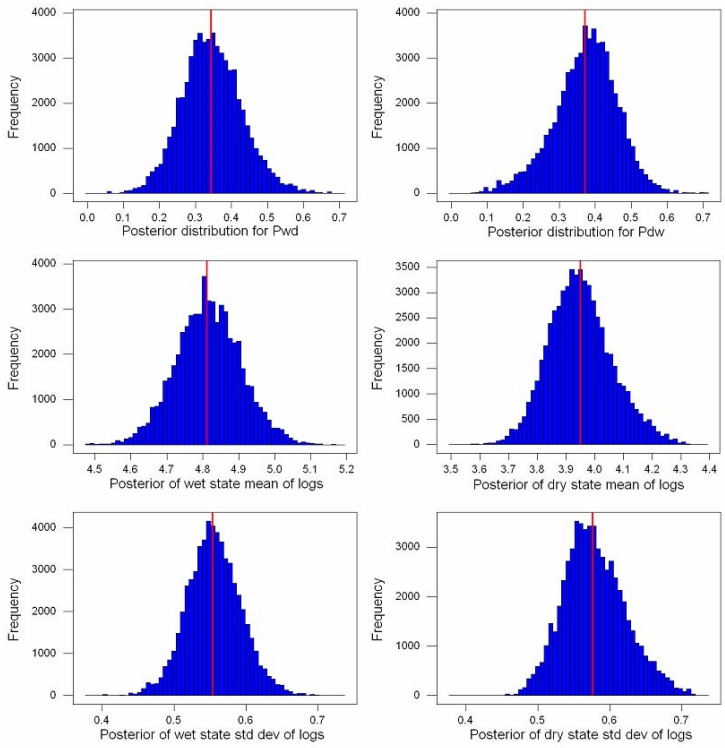

Figure 7.5 Posterior distributions, with median values, from the calibration of a two-state lognormal HMM to deseasonalised monthly rainfall for Sydney, with the state conditional means and standard deviations of the natural logarithms of monthly data shown

The series of posterior state probabilities provides a method for comparing the calibration of a two-state lognormal HMM to the deseasonalised monthly rainfall data with HMM calibrations at an annual scale. Figure 7.6 shows the posterior state series for the former together with its 90% credibility interval from calibration to the deseasonalised monthly rainfall of Sydney. A

10-year period has been chosen here to allow clearer presentation of the entire series, although the tight credibility intervals through this period are indicative of the full record. The uncertainty bounds of this state series are much narrower than for the median series derived from the annual rainfall totals. A line at probability of 0.5 is used to divide the posterior state series between likely wet months and likely dry months. This 10-year period has regular intervals during which the entire 90% credibility bounds are either side of the dividing line at a probability of 0.5, showing a clear two-state persistence pattern. The improved identification of persistence at this scale over annual totals is shown by the credibility intervals in Figure 7.6 having average width of 0.393 as opposed to 0.753 at the annual scale.

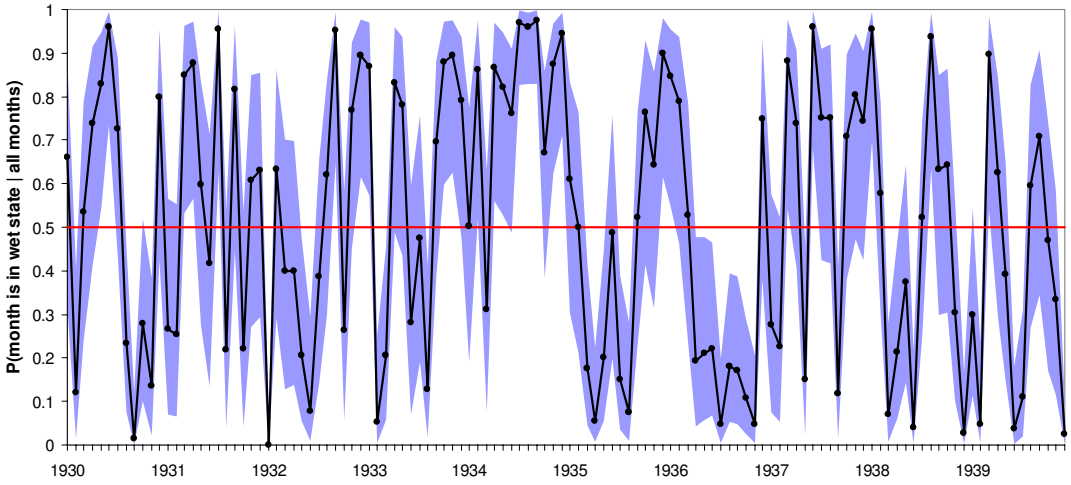

Figure 7.6 Median state series and 90% credibility interval from the calibration of a two-state lognormal HMM to the deseasonalised monthly rainfall for Sydney

The median state series for the monthly rainfall is now shown in Figure 7.7 alongside the time series of scaled monthly anomalies over the same 10-year period. Over these 120 months, most large positive anomalies are shown to exist in a wet state, which is an expected result. For the entire series (1859-2000), deseasonalised monthly values from Sydney are significantly correlated with the median posterior state probabilities ($r = 0.82$). This relationship is better expressed through a non-parametric rank correlation, which reduces the influence of the median state probabilities being constrained over the (0, 1) interval, having magnitude $r = 0.97$.

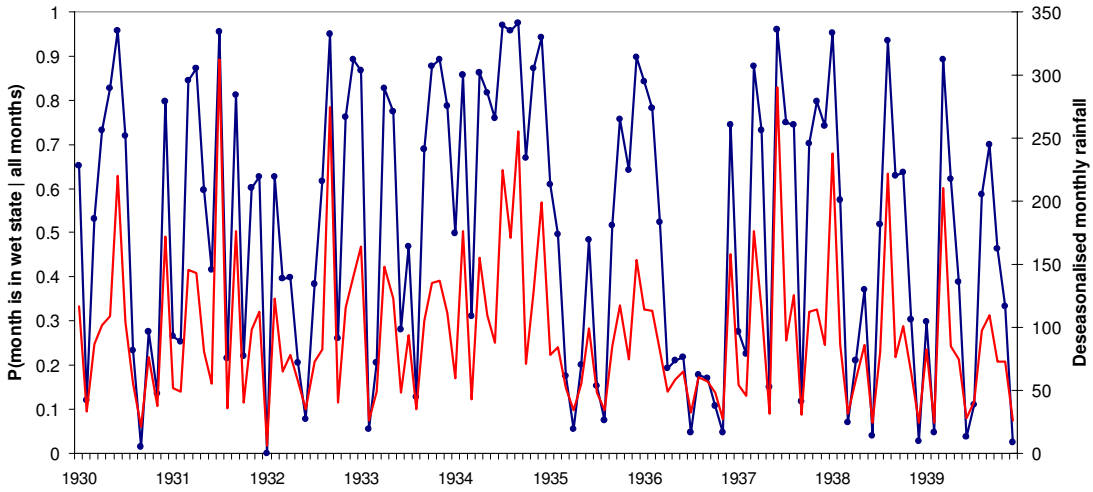

Figure 7.7 Time series of deseasonalised monthly rainfall for Sydney (red line) shown alongside the median state series following the calibration of a two-state lognormal HMM

The calibration of a two-state Gaussian HMM to the time series of annual rainfall was shown to degenerate to a mixture of Gaussians. Therefore it is pertinent to investigate whether the two-state lognormal HMM identifies significant two-state persistence in the series of deseasonalised monthly totals. In a similar manner to that described in Chapter 6, this investigation relies upon an analysis of the posterior distribution for the sum of HMM transition probabilities, summarised by the histogram in Figure 7.8. This distribution has a median value of 0.715, with a 90% credibility interval (0.588, 0.826) that rejects the possibility of this model degenerating to a mixture of lognormal distributions.

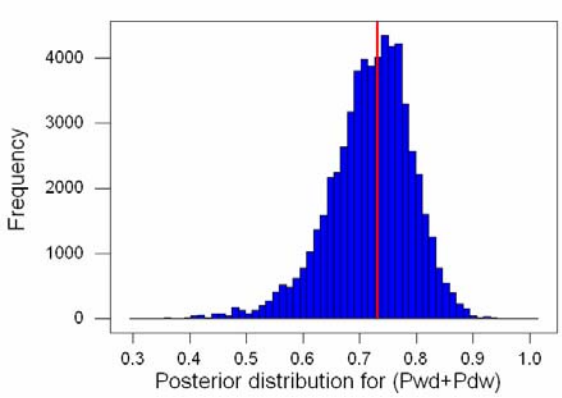

Figure 7.8 Posterior distribution of the sum of transition probabilities from the calibration of a two-state lognormal HMM to deseasonalised monthly rainfall for Sydney with median shown

This result demonstrates that HMMs identify underlying persistence in the Sydney monthly rainfall record, independent of annual seasonality, with an assumption that state conditional distributions are lognormal. It is important to now investigate whether this same model can identify significant persistence in other time series of deseasonalised monthly rainfall.

The posterior distributions for the sums of transition probabilities from calibrating two-state Gaussian HMMs to the various monthly rainfall series are summarised in Table 7.4. The three series that had approximately equal transition probability estimates, Adelaide, Darwin and Sydney, each have 90% credibility intervals around the sum of transition probabilities that are contained within (0, 1) intervals. This indicates that the two-state lognormal HMMs also identify significant monthly-scale persistence in the Adelaide and Darwin rainfall data. For the other monthly series, posterior medians of transition probabilities showed that a single state was favoured through the calibration of two-state lognormal HMMs. This characteristic reflects an absence of long-term persistence in these data. Although Table 7.4 shows that Alice Springs is the only series with a 90% credibility interval of $P_{WD} + P_{DW}$ that exceeds unity, the upper bounds of these intervals for the other series are close to 1. This suggests that the possibility of these HMMs degenerating to mixtures of lognormal distributions cannot be rejected at an approximate 10% level.

Table 7.4 Posterior medians with 90% credibility intervals for the sums of transition probabilities from the calibration of two-state lognormal HMMs to the deseasonalised monthly rainfall series

	$P_{WD} + P_{DW}$
Adelaide	**0.559** (0.306, 0.905)
Alice Springs	**0.996** (0.811, 1.088)
Brisbane	**0.691** (0.473, 0.939)
Darwin	**0.612** (0.523, 0.707)
Melbourne	**0.685** (0.392, 0.977)
Perth	**0.874** (0.736, 0.976)
Sydney	**0.726** (0.588, 0.826)
District 66	**0.695** (0.448, 0.985)

The calibration results summarised in Table 7.4 show little evidence for two-state persistence in the monthly rainfall of various locations across Australia. However these results were obtained through assuming that the deseasonalised monthly rainfall data are lognormally-distributed in each model state. Before assumptions about the influence of climatic persistence at this time-scale are made, it is pertinent to investigate the role that assumptions concerning the parametric form of these conditional distributions may have upon the ability of this model to identify true underlying persistence.

The monthly rainfall data from Sydney demonstrate significant two-state persistence when assuming lognormal conditional distributions. In order to analyse the influence of modelling assumptions, these data are now calibrated with a two-state Gaussian HMM. This produces posterior medians of 0.618 and 0.315 for P_{WD} and P_{DW} respectively, with 90% credibility intervals of (0.548, 0.688) and (0.276, 0.357). The posterior for $P_{WD} + P_{DW}$ is shown as a histogram in Figure 7.9, and this distribution has a median value of 0.934 and a 90% credibility interval of (0.860, 1.006). With this interval marginally including 1, the model calibration just fails to reject the notion of the two-state Gaussian HMM degenerating to a mixture of Gaussians at a significance level of 10%. As a consequence, the assumption of lognormally-distributed conditional distributions appears more appropriate for identifying stable climate states in this particular data set.

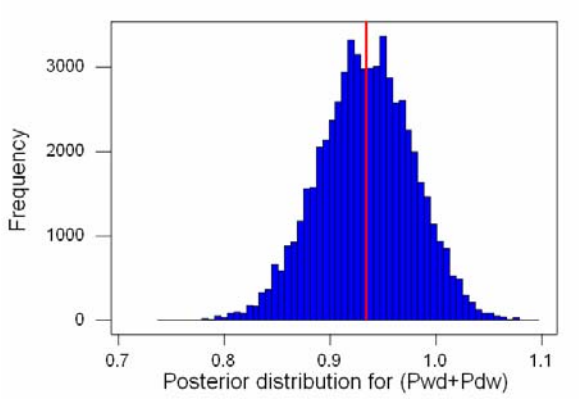

Figure 7.9 Posterior distribution of the sum of transition probabilities for the calibration of a two-state Gaussian HMM to deseasonalised monthly rainfall for Sydney with median shown

This section has clearly shown that the form of conditional distributions is a centrally important aspect in the calibration of HMMs to persistent data. These results have demonstrated that HMMs with skewed conditional distributions are adequate descriptors of persistence in various monthly rainfall series. It is therefore crucial that alternative forms of conditional distributions are now investigated; as these may provide evidence for persistence in monthly rainfall data that were not apparent with model calibrations in this section.

7.2.2 Calibration of two-state gamma HMMs

The gamma distribution is a suitable alternative to lognormal conditional distributions for the calibration of HMMs to deseasonalised monthly rainfall data. This distribution is defined by two parameters and can take a positive skew similar to the lognormal. The calibration of a two-state gamma HMM therefore requires six model parameters to be estimated, which is similar to the lognormal and Gaussian models previously described. The probability density function for this distribution is as follows:

$$f(X) = \frac{1}{b^a \Gamma(a)} X^{a-1} e^{-\frac{X}{b}} \quad \textbf{(7.1)}$$

From the calibration of a two-state gamma HMM to the various deseasonalised monthly rainfall series, posterior distributions for transition probabilities and the sum of these probabilities are summarised in Table 7.5. The 90% credibility intervals for $P_{WD} + P_{DW}$ from the HMM calibration to monthly rainfall for Brisbane, Perth and District 66 all include a value of 1, indicating that these data fail to show significant persistence from assuming gamma state conditional distributions. Although the monthly data for Melbourne shows 90% of the posterior for the sum of transition probabilities having a value less than unity, which suggests evidence for two-state persistence at a 10% level, median estimates for the two transition probabilities show expected dry state durations that are 11 times the expected length of wet state periods. The tendency to remain in a single state for a vast majority of the data series is inconsistent with the hypothesis of multiple persistent climate states. Consequently the assumption that the two state conditional distributions are derived from gamma rather than lognormal distributions fails to provide an improved description of persistence in these four monthly rainfall series.

Table 7.5 Medians of posterior distributions for HMM transition probabilities and their sum, with 90% credibility intervals from the calibration of two-state gamma HMMs

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Adelaide	**0.334** (0.145, 0.615)	**0.228** (0.075, 0.454)	**0.588** (0.350, 0.847)
Alice Springs	**0.476** (0.403, 0.552)	**0.330** (0.268, 0.400)	**0.808** (0.696, 0.921)
Brisbane	**0.449** (0.318, 0.609)	**0.330** (0.219, 0.462)	**0.781** (0.576, 1.016)
Darwin	**0.354** (0.293, 0.422)	**0.240** (0.191, 0.297)	**0.595** (0.505, 0.696)
Melbourne	**0.517** (0.280, 0.807)	**0.047** (0.010, 0.152)	**0.590** (0.304, 0.857)
Perth	**0.610** (0.399, 0.786)	**0.189** (0.123, 0.311)	**0.802** (0.597, 1.015)
Sydney	**0.426** (0.351, 0.515)	**0.285** (0.195, 0.401)	**0.718** (0.606, 0.830)
District 66	**0.444** (0.276, 0.686)	**0.423** (0.231, 0.673)	**0.882** (0.618, 1.191)

An important result from Table 7.5 is that the three series that show significant persistence from the calibration of two-state lognormal HMMs, Adelaide, Darwin and Sydney, also show significant persistence from the calibration of two-state gamma HMMs. Furthermore these calibrations estimate similar values for the two transition probabilities, demonstrating a tendency for climate states to persist over similar durations. As a result the two-state gamma

HMM provides a suitable description of persistence in these monthly rainfall series. The final data series, from Alice Springs, failed to show persistence when calibrated with a two-state lognormal HMM. However results in Table 7.5 show that a two-state gamma HMM can identify two-state persistence, suggesting this latter model offers a superior description of these data.

With the two-state gamma HMM identifying significant persistence in the monthly rainfall from Sydney, it is important to compare the calibrations of this model with the two-state lognormal HMM. Posteriors for parameters of the two conditional distributions from the former calibration are summarised in Table 7.6.

Table 7.6 Medians of posterior distributions for parameters of gamma distributions, with 90% credibility intervals, from the calibration of a two-state gamma HMM to the deseasonalised monthly rainfall for Sydney

a_W	a_D	b_W	b_D
3.875	**3.632**	**41.34**	**17.47**
(3.46, 4.54)	(3.30, 3.99)	(36.8, 46.2)	(15.2, 20.4)

The shape parameters of a gamma distribution are related to the mean and variance of the distribution through $E(X) = ab$ and $Var(X) = ab^2$. Using the posterior medians shown in Table 7.6 as estimates of the model parameters, the moments for lognormal HMM conditional distributions are compared to those for the gamma HMM conditional distributions in Table 7.7.

Table 7.7 Estimated means (with standard deviations) for state conditional distributions from the calibrations of HMMs to the deseasonalised monthly rainfall for Sydney

	Wet state	Dry state
Lognormal HMM	143.74 (86.33)	60.97 (38.16)
Gamma HMM	160.19 (81.38)	63.45 (33.29)

The moments of conditional distributions shown in Table 7.7 indicate that two-state gamma and two-state lognormal HMMs identify similar underlying distributions in the Sydney monthly rainfall. The probability distribution functions for these models that are generated from posterior median parameter estimates are shown alongside each other in Figure 7.10, with a higher wet state mean for the gamma HMM clearly apparent.

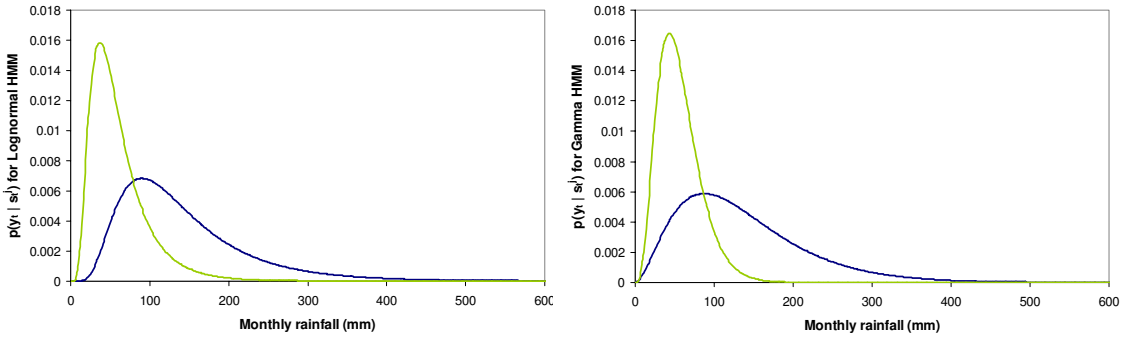

Figure 7.10 State conditional distributions estimated from the calibration of a two-state lognormal HMM (left) and a two-state gamma HMM (right) to the deseasonalised monthly rainfall for Sydney

With the gamma HMM identifying conditional distributions in the Sydney monthly rainfall that are similar to the lognormal HMM, it is important to determine whether these two models estimate a similar sequence of monthly states within these data. In fact the linear correlation between the median state series for the two models is very strong ($r = 0.951, p < 0.001$), with these series also having a rank correlation of $r = 0.993$. It is clear therefore that these models identify similar persistence in the Sydney rainfall series.

In order to generate Bayes Factors between each HMM variant, suitable prior distributions are required for each model parameter. As outlined in Section 3.4.3 conjugate priors are desirable, and the conjugate prior for the location parameter of a lognormal distribution is Gaussian, defined with the mean and standard deviation of the natural logarithms of monthly variates. Conjugate priors for the scale parameter of lognormal distributions are inverse chi-squared distributions, the shapes of which are defined with the variance of log samples. Uniform priors over the interval (0, 1) are also used for HMM transition probabilities. For gamma conditional distribution, explicit forms for conjugate priors are not available, so uniform priors defined with statistics of the marginal distribution of rainfall observations are used in this analysis for both parameters. By fitting a single gamma distribution to the marginal, maximum likelihood estimates for the "a" and 'b" parameters are 1.83 and 57.13 respectively. Uniform priors are subsequently defined over intervals (0, 10) for the "a" parameter and (0, 100) for the "b" parameter. Upper bounds for these distributions were varied to test sensitivity; however it was assumed that the "a" parameter would always take smaller values than the "b" parameter. Uniform priors form vague distributions, and allow the data to dominate the posterior. Bayesian model selection produces an estimate of 4.0 for $\ln BF_{L,G}$, which is marginally in favour of the lognormal distribution.

Table 7.8 Model selection results from the calibrations of two-state lognormal (LN) and two-state gamma HMMs to deseasonalised monthly rainfall series

		Maximum Likelihood	$\ln BF_{L,G}$
Adelaide	Lognormal	-2535.4	
	Gamma	-2531.2	-140.4
Alice Springs	Lognormal	-2832.7	
	Gamma	-2777.8	-52.4
Darwin	Lognormal	-5390.0	
	Gamma	-5375.6	-42.4
Sydney	Lognormal	-9389.4	
	Gamma	-9391.9	4.0

The two-state gamma HMM describes monthly persistence in Sydney rainfall in a manner similar to a two-state lognormal HMM. Table 7.8 shows model selection results for the four deseasonalised monthly series that showed significant persistence through the calibrations of two-state gamma and lognormal HMMs. Negative Bayes Factors in this table favour the two-state gamma HMM over the two-state lognormal HMM. Consequently for the Adelaide, Alice Springs and Darwin monthly data, Bayes Factor analyses favour the two-state gamma HMM, suggesting that for these data, the latter model is superior in a Bayesian context. With the two parametric HMMs show here identifying significant persistence in certain deseasonalised monthly rainfall series, it is now appropriate to investigate the relationship between model states and other measures of physical processes that underlie hydrologic variability.

7.2.3 Relationships between HMM states and measures of climate variability

The El Niño Southern Oscillation has a major influence upon rainfall variability across much of eastern Australia, and it is likely that hydrological persistence in these regions would reflect changes in ENSO indices. As the posterior state series from the calibration of a two-state lognormal HMM to the Sydney data is highly correlated to the state series from a two-state gamma HMM, only the state series from the former model is analysed here. For regions of Australia influenced by ENSO variability, El Niño periods are generally associated with below-average rainfall, with La Niña episodes generally producing above-average rainfall. It follows that two-state persistence in such regions would display higher wet state probabilities in La Niña months and lower probabilities in El Niño months.

The efficacy of the calibration of a lognormal HMM to the Sydney monthly rainfall is first examined through relationships between posterior state probabilities and the time series of NINO3 5-mrm values. By classifying each month as El Niño, La Niña or ENSO neutral, using the method described in Section 2.2.1, 2-sample t-tests show La Niña periods being associated with a mean wet state probability of 0.587, significantly higher ($p < 0.001$) than the mean wet state probability in El Niño months (0.486).

The relationships between hidden climate states and alternative measures of persistence in the climate, such as ENSO indices, are further analysed in Table 7.9, with months segregated on the basis of ENSO phase and HMM state. The 5-mrm values of the NINO3 series are again used to classify ENSO months. Using the median state series from the calibration of the two-state lognormal HMM, a probability of 0.5 is then used to separate months into having been generated from a wet or a dry climate state. These results show a tendency for La Niña episodes to coincide with wet climate states, which is expected given the influence of ENSO upon rainfall along the east coast of Australia. The influence of El Niño episodes however are not revealed as clearly as La Niñas. These phases have almost equal chance of coinciding with wet or dry climate states, thus lacking the strong bias towards dry states that would be expected if the HMM was isolating the influence of ENSO in these rainfall data. This demonstrates that the persistence identified in the Sydney rainfall data originates from the hydroclimatic influences of various global climate phenomena.

Table 7.9 Numbers of months in which most probable HMM states from the calibration of a two-state lognormal HMM to the deseasonalised monthly rainfall for Sydney coincide with different ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	243	380	304
Dry state	261	342	186

This section has shown that two-state parametric HMMs identify persistence within deseasonalised monthly rainfall data, particularly from Sydney, that is revealed through alternative measures. It is therefore pertinent to investigate methods that extend this modelling approach, in order to further analyse the nature of this hydrological persistence. The following section investigates the possibility for multiple-state persistence in these data.

7.3 Using HMMs to identify possible three-state monthly persistence

A pragmatic approach to modelling hydroclimatic persistence with HMMs is to define two broad climate states as previously demonstrated. Although two-state models that describe predominantly wet and predominantly dry conditions provide a useful mechanism for downscaling climatic variability to hydrological persistence, HMMs are not constrained to this basic structure. As an example, applications of HMMs to continuous speech recognition can incorporate different states for individual acoustic-phonetic units, thus requiring 40-50 states to describe all sounds in the English language (Rabiner, 1989). An increase in the number of states

increases model complexity, hence the number of parameters to be estimated. This complexity however may allow persistence to be more clearly observed.

The number of model states is an important aspect of model development. One approach is to fix this number prior to calibration in order to reflect a range of system characteristics, and attention then focused upon other model characteristics such as the form of conditional distributions. An alternative approach is to choose the number of states through model selection procedures in order to deliver an optimal model structure. Robertson *et al.* (2004) used a HMM to describe daily rainfall occurrence across 10 rain gauges in northeast Brazil, applying cross-validation methods to identify the most suitable number of model states. Following the calibration of a 4-state model to these data, these authors used anomalies in outgoing long-wave radiation and surface winds to observe the meteorological characteristics associated with each rainfall state. Intraseasonal and interannual teleconnections were revealed in different states, and provided a physical basis for the predictability of daily rainfall.

The influence of El Niño and La Niña episodes upon Australian hydrology has been previously noted, however between these ENSO extremes the climatic conditions of the eastern equatorial Pacific tend to exist in a "neutral" phase. ENSO may therefore be interpreted as a tendency for the Pacific climate to fluctuate between three distinct states rather than two, and it is possible a similar characteristic is evident in rainfall observations. Under these conditions, three-state persistence within monthly rainfall is justified; however the combined influence of non-ENSO climate modes such as the IPO and the IOD may realise an even greater number of discrete climate states. In the calibration of HMMs to monthly rainfall series, Bayesian model selection methods such as Bayes Factors may determine the most appropriate number of hidden states.

Three-state HMMs are now calibrated to the deseasonalised monthly rainfall series introduced earlier, using assumptions that each state conditional distribution is either lognormal or gamma. In the calibration of these models, six transition probabilities require estimation, together with two parameters for each of the three state conditional distributions. The calibration of a three-state lognormal HMM to the time series of deseasonalised monthly rainfall for Sydney produces posterior medians that approximate the state conditional distributions $P(\ln(y_t) \mid s_t^j)$ as $N(5.01,0.49^2)$, $N(4.50,0.57^2)$ and $N(3.93,0.58^2)$. These distributions are well-separated, with wet and dry state distributions closely matching estimates for a two-state lognormal HMM. For a three-state gamma HMM, the wet, neutral and dry state distributions are estimated respectively as $Gamma(5.34,36.47)$, $Gamma(7.75,12.99)$ and $Gamma(4.75,10.46)$. The relationships between the distributions for each model are described in Figure 7.11, and it is clear that the gamma HMM shows a longer-tailed wet state.

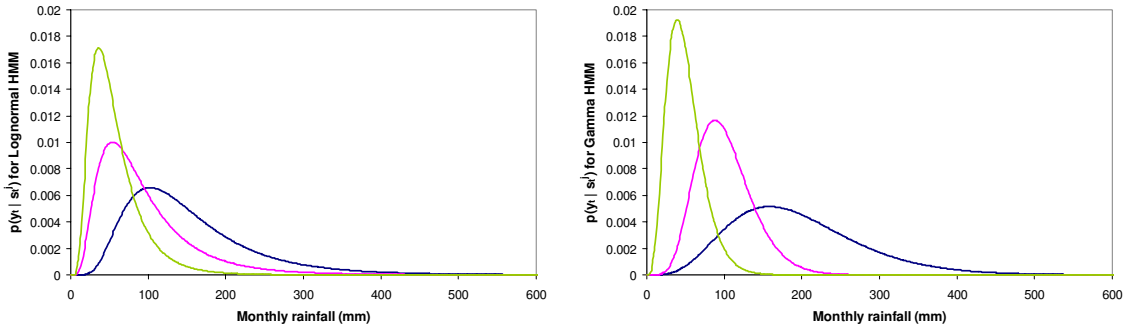

Figure 7.11 State conditional distributions estimated from the calibration of a three-state lognormal HMM (left) and three-state gamma HMM (right) to the deseasonalised monthly rainfall for Sydney

Figure 7.11 shows that both three-state lognormal and gamma HMMs identify well-separated conditional distributions within the Sydney monthly rainfall series, suggesting significant three-state persistence. To further analyse these calibrations, the posteriors for the six transition probabilities for the lognormal and gamma HMMs are shown in Figure 7.12 and Figure 7.13 respectively. Vertical lines indicate the median values of these distributions, and the subscript N denotes the "neutral" climate state.

Posterior distributions for the lognormal HMM transition probabilities have greater spread than those from the gamma HMM. Furthermore, these distributions have wider bounds than the posteriors of either transition probability from the calibration of a two-state lognormal HMM to these data. This indicates that the three-state lognormal HMM produces inferior estimates of transitions between the model states. Importantly the dry state self-transition probability for both the three-state lognormal HMM and the three-state gamma HMM, which is the complement of $P_{DW} + P_{DN}$, is higher than self-transition probabilities for the other two states. This implies that these two models identify a majority of months in a dry climate state.

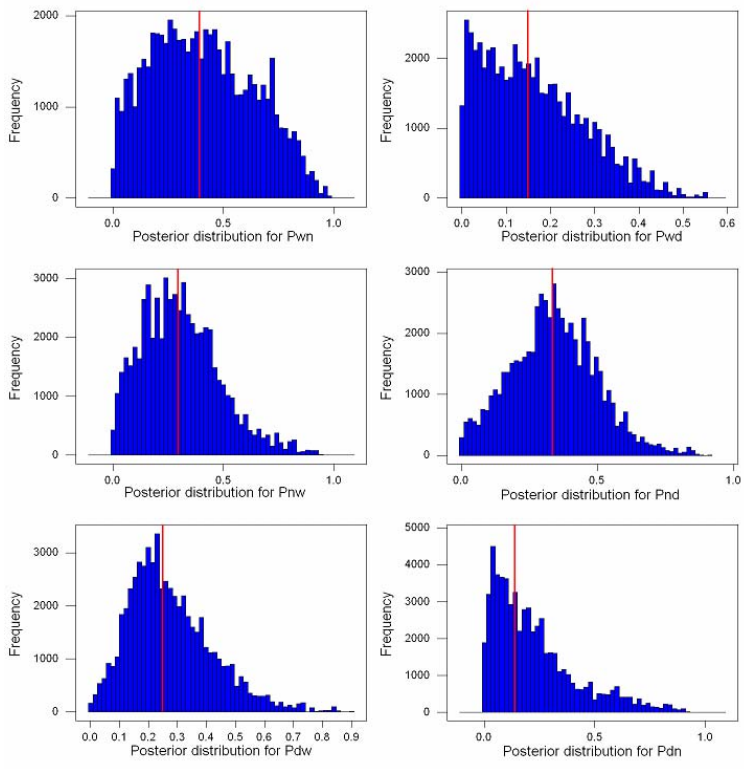

Figure 7.12 Posterior distributions of transition probabilities from the calibration of a three-state lognormal HMM to deseasonalised monthly rainfall for Sydney with medians shown

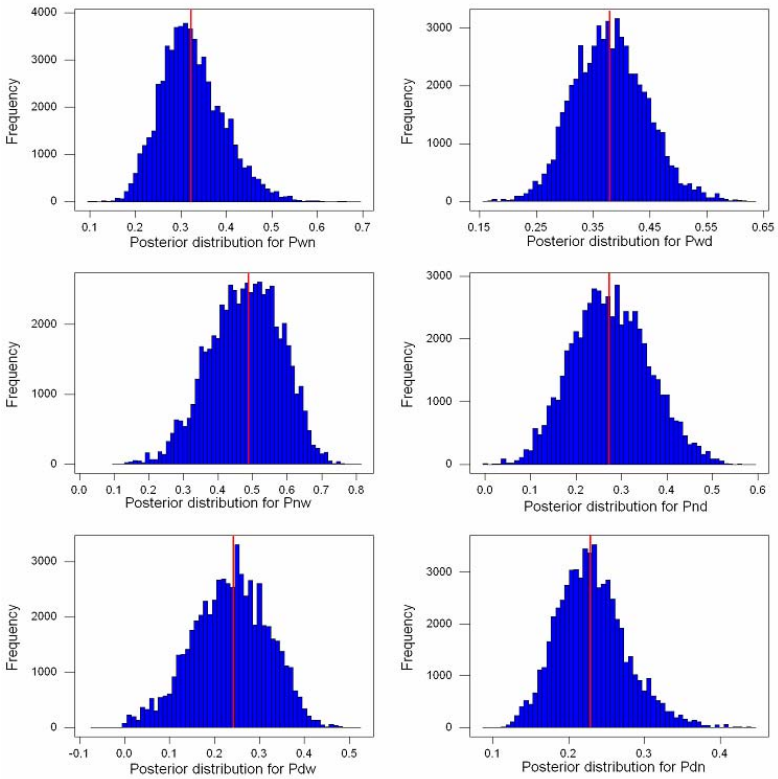

Figure 7.13 Posterior distributions of transition probabilities from the calibration of a three-state gamma HMM to deseasonalised monthly rainfall for Sydney with medians shown

In three-state HMMs, posterior state probabilities need to be calculated separately for each model state. As a result it is difficult to compare the credibility intervals of the state series as shown previously with the calibration of two-state HMMs. However by estimating credibility bounds around posterior state series for wet, neutral and dry states separately, the probability for month t existing in state j is estimated by the posterior mean probability. Using this approach, the sequence of state probabilities from the calibration of a three-state lognormal HMM to the deseasonalised monthly rainfall for Sydney is shown in Figure 7.14 over a 10-year period (Jan 1930 to Dec 1939). A majority of months in this sequence show higher dry state probabilities than either neutral state or wet state probabilities. The state that shows the highest posterior probability at each month is the most likely model state. Figure 7.15 then shows the most likely state sequence over this 10-year period, with a bias towards dry states clearly demonstrated.

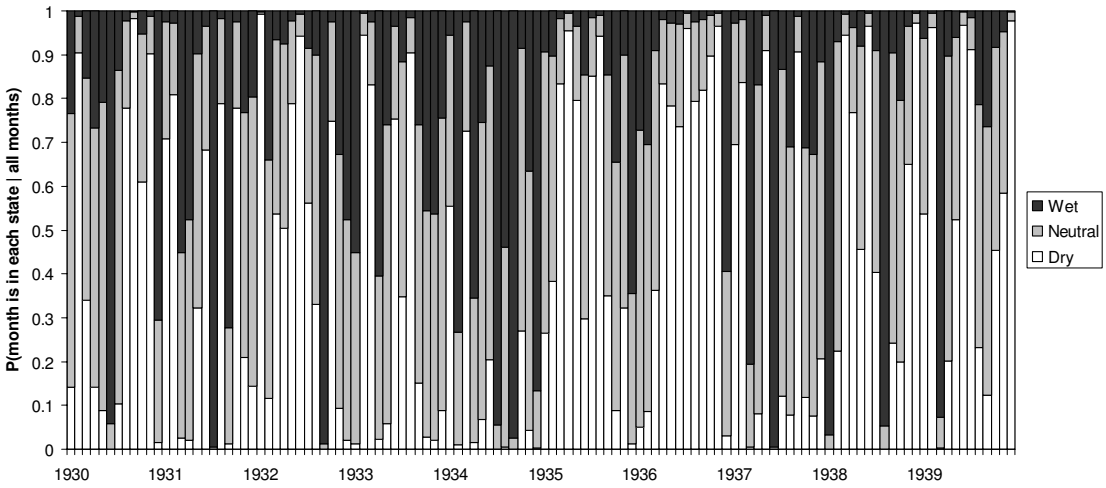

Figure 7.14 Posterior state probabilities over a 10-year period from the calibration of a three-state gamma HMM to the deseasonalised monthly rainfall for Sydney

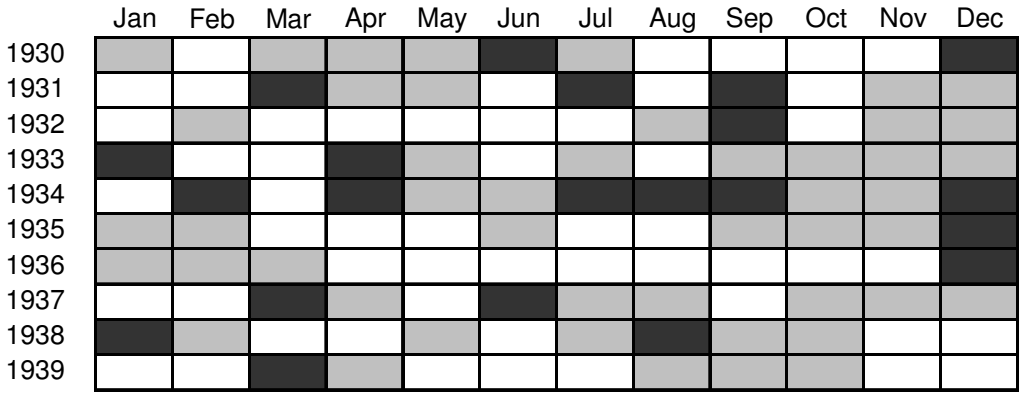

Figure 7.15 Sequence of most likely states for a 10-year period from the calibration of a three-state lognormal HMM to the deseasonalised monthly rainfall for Sydney, colours as used in Figure 7.14

Figure 7.15 indicates that 52 of the 120 months are most likely to have been generated from a dry state distribution. Over the entire time series, 743 months are most likely to be in a dry state

(43.3%), as opposed to 375 months (21.9%) in a wet state and 598 months (34.8%) in a dry state. The ratio of the number of dry months to wet months in this three-state model (1.98) is much higher than for the two-state model (1.17).

The significance of state transitions identified by the three-state lognormal HMM is now analysed in terms of observed hydroclimatic variability. By categorising each month in the Sydney record by likely model state, monthly wet states are associated with an average deseasonalised anomaly of 222.4, which is significantly higher than the average total for neutral states (102.8) and dry states (45.9). This suggests that three-state persistence identified by the gamma HMM identifies observed modes of rainfall variability.

By using the series of most likely states to separate monthly NINO3 values into those associated with wet, neutral and dry months, 2-sample t-tests show HMM wet states to have an average NINO3 value of -0.014, compared with a mean value in dry months of 0.165. This highlights a tendency towards more La Niña months coinciding with wet states and more El Niños occurring during dry months. A 2-sample t-test shows that the difference between average NINO3 values in these two states is statistically significant $(p < 0.001)$. Similarly, neutral months are coupled to significantly lower NINO3 values than during dry months $(p = 0.002)$.

The relationship between persistence and measures of ENSO variability is further analysed in Table 7.10, which shows the numbers of months in which most likely climate state coincides with ENSO phase. This table shows a bias towards more wet state months occurring during La Niña episodes, and months during El Niño phases are more likely to be modelled as dry climate states. Both of these results are consistent with monthly persistence in Sydney rainfall reflecting ENSO variability. The neutral climate state fails to correspond with ENSO neutral phases however, which suggests either that the three-state HMM is too complex a model for this persistence, or that climate phenomena other than ENSO have significant roles.

Table 7.10 Numbers of months in which most probable HMM states from the calibration of a three-state gamma HMM to the deseasonalised monthly rainfall for Sydney coincide with different ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	101	149	125
Neutral state	161	246	191
Dry state	242	327	174

In order to investigate the significance of three-state persistence in the time series of the deseasonalised monthly rainfall for Sydney, it is necessary to analyse the calibration of the three-state gamma HMM. In Section 6.4, the degeneration of a two-state HMM to a mixture distribution was shown to occur with $P_{WD} + P_{DW} = 1$. A 90% credibility interval for this sum including a value of 1 provides evidence for a mixture distribution, whereas $P_{WD} + P_{DW} < 1$ provides evidence against this mixture. In the two-state model this inequality is equivalent to $P_{WW} + P_{DD} > 1$, such that a credibility interval around the sum of self-transition probabilities being entirely greater than 1 indicates two-state persistence. This analysis can be extended to the case of three or more states, in which the sum of self-transition probabilities is used as the test statistic. In the three-model case, the posterior distribution for $P_{WW} + P_{NN} + P_{DD}$ is therefore analysed. By simplifying the sum of six transition probabilities of a three-state HMM as ΣTP, the sum of self-transition probabilities is calculated as $(3 - \Sigma TP)$. A 90% credibility interval around $P_{WW} + P_{NN} + P_{DD}$ being above unity is evidence against three-state persistence degenerating to a mixture of three distributions.

To investigate whether three-state HMMs maintain persistence that is identified with two-state models, the distributions of the sums of self-transition probabilities for the lognormal and gamma models are shown in Figure 7.16 with vertical lines at posterior medians.

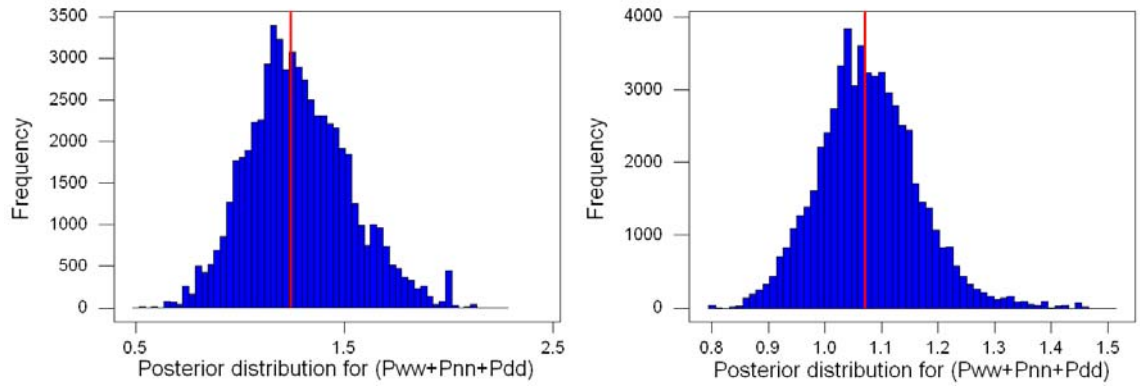

Figure 7.16 Posterior distributions for the sums of self-transition probabilities from the calibration of a three-state lognormal HMM (left) and a three-state gamma HMM (right) to the deseasonalised monthly rainfall for Sydney with medians shown

With the three-state lognormal HMM having wider posteriors for transition probabilities than the three-state gamma HMM, the former model also has a wider distribution for the sum of self-transition probabilities. This distribution has a median of 1.238, with 95% of posterior samples greater than 0.895, contrasting to a median estimate from the three-state gamma HMM of 1.072 with 95% of its samples lying above 0.939. Even though each three-state HMM shows values for the sum of self-transition probabilities to be similar to values obtained from the two-state

HMMs, the 90% credibility intervals for the three-state HMMs include a value of 1. Consequently, the possibility of these three-state models degenerating to mixture distributions cannot be rejected at the 10% significance level.

The calibrations of three-state lognormal and three-state gamma HMMs to the various deseasonalised monthly rainfall series are summarised in Table 7.11. These results show significant evidence for three-state persistence in the monthly rainfall of Adelaide, Alice Springs, Brisbane and Melbourne from calibrating both lognormal and gamma HMMs. Monthly rainfall for Darwin and Perth show persistence using only one of these three-state HMMs. Furthermore, deseasonalised monthly rainfall for Sydney and District 66 fail to reject the notion of three-state persistence degenerating to either a mixture of three lognormals or a mixture of three gammas.

Table 7.11 Medians of posterior distributions for the sums of self-transition probabilities, with 90% credibility intervals, from the calibrations of three-state lognormal HMMs and three-state gamma HMMs to deseasonalised monthly rainfall series

	$P_{WW} + P_{NN} + P_{DD}$ 3-state lognormal HMM	$P_{WW} + P_{NN} + P_{DD}$ 3-state gamma HMM
Adelaide	**1.415** (1.075, 1.745)	**1.494** (1.030, 1.876)
Alice Springs	**1.287** (1.117, 1.503)	**1.798** (1.465, 2.340)
Brisbane	**1.620** (1.357, 1.894)	**1.402** (1.036, 1.735)
Darwin	**1.222** (1.119, 1.348)	**1.264** (0.850, 1.745)
Melbourne	**1.486** (1.132, 1.910)	**1.712** (1.407, 2.051)
Perth	**1.300** (0.849, 1.766)	**1.347** (1.093, 1.656)
Sydney	**1.260** (0.915, 1.714)	**1.072** (0.939, 1.225)
District 66	**1.340** (0.975, 1.793)	**1.207** (0.906, 1.522)

With estimates of transition probabilities showing significant three-state persistence in various deseasonalised monthly rainfall series, it is important to investigate whether three-state HMMs offer superior descriptions of persistence within these monthly data than two-state HMMs. Bayes Factors are evaluated between two-state lognormal HMMs, three-state lognormal HMMs and two-state and three-state gamma HMMs using the same prior distributions as those described previously. This model selection process for the eight monthly rainfall series are summarised in Table 7.12, with the optimal models shown for each monthly series.

Table 7.12 Bayes factors from comparing calibrations to deseasonalised monthly rainfall series of two-state lognormal HMMs, to three-state lognormal HMMs and gamma HMMs

	$\ln BF_{2L,3L}$	$\ln BF_{2L,2G}$	$\ln BF_{2L,3G}$	Optimal model
Adelaide	-68.4	-70.1	-79.2	3-state gamma HMM
Alice Springs	-23.4	-26.2	16.6	2-state gamma HMM
Brisbane	-5.6	1.9	-11.2	3-state gamma HMM
Darwin	-51.9	-21.2	-28.9	3-state lognormal HMM
Melbourne	0.4	0.5	4.8	2-state lognormal HMM
Perth	2.9	8.7	-1.9	3-state gamma HMM
Sydney	1.2	2.0	2.8	2-state lognormal HMM
District 66	10.1	1.0	-3.4	3-state gamma HMM

After reviewing the results of both Table 7.11 and Table 7.12, there is evidence to show that three-state gamma HMMs are the most appropriate models for the monthly rainfall in Adelaide, Brisbane and Perth, whereas a three-state lognormal HMM is appropriate for the Darwin data. Transition probability estimates suggested significant three-state persistence in the monthly rainfall series of Alice Springs and Melbourne, however Bayesian model selection indicates these data are better described with two-state persistence. Although the three-state gamma HMM is the optimal model for the District 66 series, Table 7.11 shows that the possibility of three-state persistence degenerating to a mixture of three gamma distributions cannot be rejected at a 10% level. Furthermore, model selection results for this series are different to those from the Sydney monthly series, to which the District 66 data is closely related. This indicates that assumptions used in the calibrations of these models to the deseasonalised monthly rainfall for District 66 could be unsuitable, and this is investigated more thoroughly in the following section.

7.4 Problems in the calibration of incorrect parametric HMMs

The parametric forms of state conditional distributions are vitally important aspects in HMM design, and unrealistic assumptions about the form of these distributions can bias the estimation of transition probabilities, leading to incorrect model fitting. Difficulties posed by the need to define these conditional distributions are illustrated in this section, using the time series of deseasonalised monthly anomalies from District 66 as an example.

7.4.1 Using monthly rainfall for District 66 to illustrate calibration problems

In order to demonstrate the influence of parametric assumptions upon model estimation, it is useful to return to the results from the calibration of a two-state lognormal HMM to the time series of deseasonalised monthly rainfall recorded in District 66. As shown in Table 7.2, these data are consistent with random lognormal draws with $P(\ln(y_t)) \approx N(4.40, 0.72^2)$. Although

the calibration of a two-state lognormal HMM to the deseasonalised monthly rainfall for Sydney identifies strong two-state persistence in these data, the calibration of this model to the District 66 data produces some contrasting results. Table 7.13 summarises the posteriors from these calibrations; a large difference between transition probability estimates suggests a tendency for the HMM to remain within the wet state for the entire District 66 series. Broad posteriors for the mean and standard deviation of dry state observations show that this model cannot accurately identify the dry state, further emphasising the degeneration into a single distribution.

Table 7.13 Comparison of posterior distributions showing median and 90% credibility intervals, from the calibration of two-state lognormal HMMs to deseasonalised monthly rainfall in Sydney and District 66

	P_{WD}	P_{DW}	Wet state location	Dry state location	Wet state scale	Dry state scale
66	**0.031** (0.01, 0.09)	**0.654** (0.41, 0.95)	**4.461** (4.42, 4.52)	**3.437** (2.61, 4.00)	**0.649** (0.62, 0.68)	**0.924** (0.69, 1.31)
Syd	**0.343** (0.29, 0.49)	**0.379** (0.22, 0.51)	**4.814** (4.67, 4.97)	**3.945** (3.80, 4.15)	**0.555** (0.50, 0.62)	**0.575** (0.51, 0.66)

The differences in the calibrations of two-state lognormal HMMs to the deseasonalised monthly anomalies from Sydney and District 66 are striking. Although these series are strongly correlated, the Sydney data demonstrate clear two-state monthly persistence that is absent in the spatially-averaged series. The two-state lognormal HMM tends to approximate the marginal monthly distribution for District 66 with a single "wet state" distribution. Taking exponentials of a series generated from $N(4.461, 0.649^2)$, the parameters of which equal the posterior medians of wet state parameters, produces a series of lognormal variates that has an expected value of 106.87 and theoretical standard deviation 77.34 using the standard relationships $E[X] = \exp(\lambda + 0.5\zeta^2)$ and $Var[X] = E^2[X](\exp(\zeta^2) - 1)$ where $\lambda = 4.461$ and $\zeta = 0.649$. The moments of this exponential series are very similar to the moments of the deseasonalised monthly series for District 66, for which the sample mean and standard deviation are 104.31 and 77.05 respectively.

The behaviour of the two-state lognormal HMM is further demonstrated in Figure 7.17, the posterior state series for the deseasonalised monthly rainfall for District 66 shown over the same 10-year period as used for the Sydney data in Figure 7.6. The general pattern of posterior state probabilities remaining close to 1 during this period is indicative of state probabilities for the entire 90-year series. Indeed over this period, only 18 months (1.7% of all values) have a median wet state probability less than 0.5. A visual comparison of this figure with the corresponding series from the monthly Sydney data shows two vastly different HMM state series, and these have a linear correlation of $r = 0.40$ that is dramatically lower than the

correlation between the two series of deseasonalised monthly rainfall $(r = 0.87)$. Furthermore a dry state distribution describes only the extreme low values in the series, which is shown by the 18 months that have a median wet state probability less than 1 (hence more likely to have been generated from a dry state distribution) corresponding to an average deseasonalised monthly value of 16.4, as opposed to an average value over the marginal distribution of 104.3.

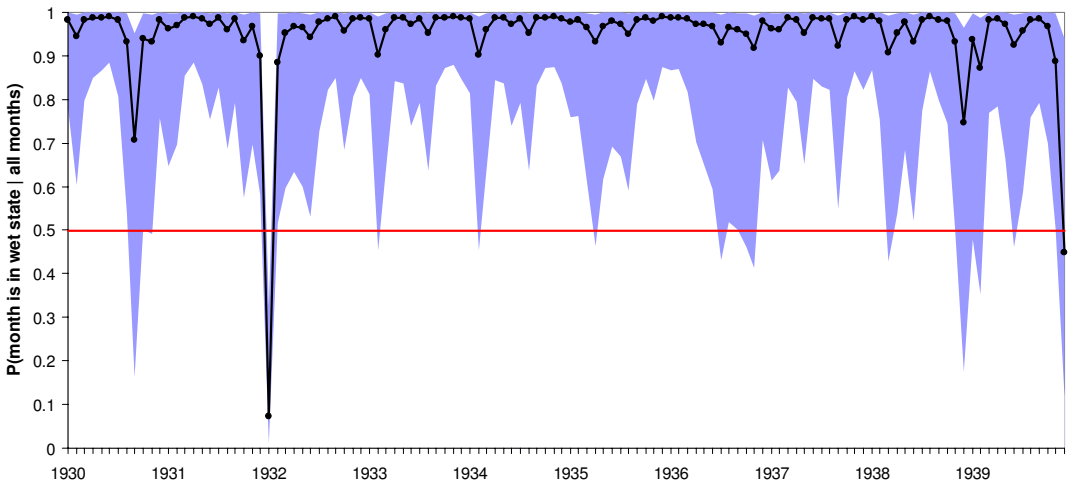

Figure 7.17 Median state series and 90% credibility interval over a 10-year period from the calibration of a two-state lognormal HMM to deseasonalised monthly rainfall for District 66

Together with transition probability estimates that failed to reject the possibility of the two-state model fitting only a mixture of two lognormals, Figure 7.17 indicates that the two-state lognormal HMM inappropriately describes persistence in the monthly rainfall over District 66. Figure 6.5 showed that the posterior state series from calibrating two-state Gaussian HMMs to the annual rainfall series from Sydney and District 66 are closely related, which suggests that a similar sequence of hidden states are estimated in these records. For the two monthly time series however, two-state lognormal HMMs fail to identify a common sequence of hidden states. If the outputs from these models are consistent with broad-scale climatic influence in the Sydney region, monthly-scale persistence in point rainfall would likely reflect persistence in the time series of spatially-averaged rainfall.

The discrepancy apparent in these results can be explained in two ways: either that the apparent model states are unrelated to hydroclimatic interactions or that the parametric assumptions of the two-state lognormal HMM are unrealistic. However since the monthly state series for Sydney appeared to reflect ENSO changes described by the NINO3 series, it is clear that the HMM is consistent with some of the hydroclimatic influence on the Sydney monthly rainfall. Therefore it is important that the manner by which incorrect parametric assumptions can lead to inaccurate identifications of hidden climate states is now investigated, through a range of simulation studies in the following section.

7.4.2 Calibration of incorrect HMMs to simulated data

A time series is simulated over length 1080 from a two-state Gaussian HMM, using parameter values that are shown in Table 7.14. This series has an arithmetic mean of 112.9, with a sample standard deviation of 53.0 and skew of 1.00. Taking natural logarithms of these values produces a series with mean 4.61 and standard deviation 0.52. A two-state Gaussian HMM is calibrated to this series to examine its ability to recover known parameter values. Table 7.14 summarises these results and indicates that the Gaussian HMM correctly identifies each model parameter within 90% credibility intervals. This result was expected given that the correct model framework was being calibrated.

Table 7.14 Comparison of posterior distributions from the calibration of a two-state Gaussian HMM to a time series simulated from an identical model

	P_{WD}	P_{DW}	μ_W	μ_D	σ_W	σ_D
Simulated values	0.4	0.15	170	90	60	30
Gaussian HMM	**0.438** (0.35, 0.52)	**0.133** (0.11, 0.17)	**182.1** (166, 194)	**91.4** (89, 94)	**51.7** (44, 62)	**30.6** (29, 32)

The median state series from the calibration of the Gaussian HMM to these simulated data is shown together with a 90% credibility interval in Figure 7.18 over a period of 120 values. The state series used in simulation is shown with circles. The narrow credibility intervals around the parameter estimates from this model produce tight bounds around the median state series.

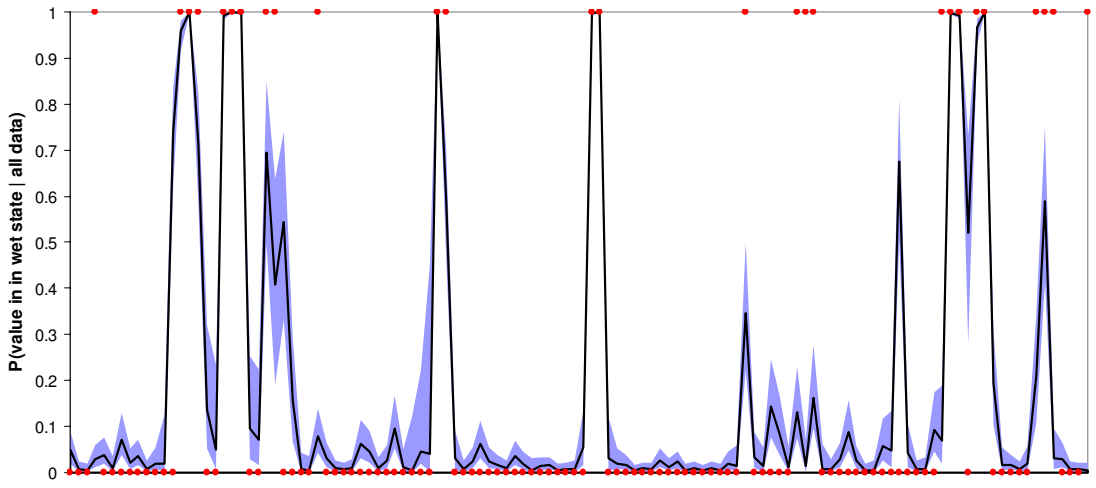

Figure 7.18 Median state series and 90% credibility interval for 120 values from the calibration of a two-state Gaussian HMM to the simulated time series of length 1080

Now if correct forms of state conditional distributions are deemed unknown, and incorrectly assumed to follow random draws from two lognormal distributions, a two-state lognormal HMM is subsequently calibrated. The posterior distributions from this model calibration are

summarised in Table 7.15. These results are relevant to this analysis as firstly the transition probabilities indicate a tendency for the HMM to remain in a wet state for a much longer period than the dry state, and secondly the posterior for the wet state location parameter includes the mean of the series of natural logarithms within a 90% credibility interval. These two features are consistent with the calibration of a two-state lognormal HMM to the District 66 data.

Table 7.15 Median and 90% credibility intervals for parameters of a two-state lognormal HMM calibrated to the simulated time series of length 1080

P_{WD}	P_{DW}	Wet state location	Dry state location	Wet state scale	Dry state scale
0.033	**0.820**	**4.636**	**3.625**	**0.460**	**1.082**
(0.02, 0.06)	(0.62, 0.95)	(4.61, 4.66)	(2.89, 4.13)	(0.43, 0.48)	(0.86, 1.34)

The difference between the calibrations of a two-state lognormal HMM and a two-state Gaussian HMM to this simulated series is clearly demonstrated in Figure 7.19, which displays the median state series (and 90% credibility interval) from the lognormal HMM for the same period of 120 values as previously used in Figure 7.18. This series shows a distinctly different form to that estimated with a Gaussian HMM. The state sequence used in simulation is not well identified, with the lognormal HMM estimating (with high probability) that most values were generated from a single wet state lognormal distribution.

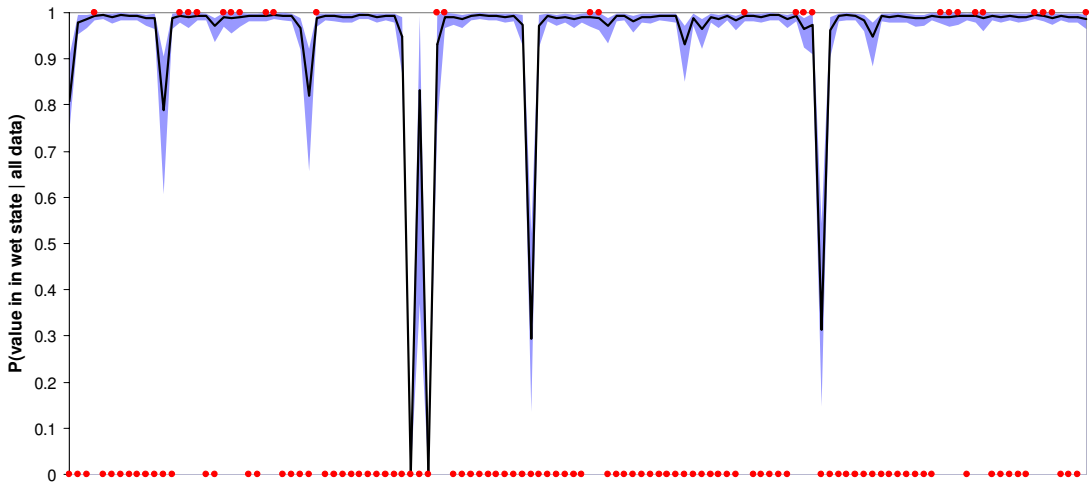

Figure 7.19 Median state series and 90% credibility interval for the same 120 values as Figure 7.18 from the calibration of a two-state lognormal HMM to the simulated series of length 1080

It is apparent that the posterior state series produced from the calibration of a two-state lognormal HMM to the simulated series has a similar structure to the series obtained from fitting this same model to the deseasonalised monthly rainfall for District 66. A single lognormal distribution appears to describe the majority of values, and only extreme low values are likely to have been simulated by the second lognormal distribution. In fact, only 31 of the

1080 values (2.9% of the series) have posterior state probabilities that suggest their simulation from a "dry state" distribution. The sum of transition probabilities for the state series has a posterior median of 1.009 and 90% credibility interval of (0.89, 1.08). This result provides strong evidence that the lognormal HMM degenerates to a mixture of two lognormals, with the inherent persistence that was simulated from a Gaussian HMM unable to be detected.

Although the simulation results appear similar to those obtained through the calibration of a lognormal HMM to the series of deseasonalised monthly anomalies for District 66, it is important to note that the spatially-averaged series is not modelled sufficiently by a two-state Gaussian HMM. Although the model identifies two underlying distributions that are well-separated, estimated as $N(179,88^2)$ and $N(68,32^2)$ from posterior medians of parameters, the sum of transition probabilities has a posterior median of 0.979, with 90% credibility interval of (0.89, 1.06). This indicates that the Gaussian HMM degenerates to a mixture of two Gaussians, and the assumption of underlying climate states is not supported.

State conditional distributions for the series of deseasonalised monthly anomalies in District 66 are inconsistent with random draws from either Gaussian, lognormal or gamma distributions. Although it is possible that a wet state conditional distribution has different parametric form to the dry state distribution, a more rational perspective would be that climate processes distribute rainfall data similarly in each discrete state. The important question of what form the state conditional distributions in monthly time series should assume is investigated in Chapter 8 using a novel approach.

7.5 Summary of chapter

In response to an inability to detect significant two-state persistence in annual rainfall data, HMMs have been calibrated in this chapter to time series of deseasonalised monthly totals. The monthly time scale is more appropriate given both the frequencies of dominant climate modes that underlie hydrological persistence across Australia, and also the increase in information over annual data that is obtained. However the analysis of monthly-scale persistence has not been previously reported in the literature. The calibration of HMMs to monthly rainfall from cities across Australia, after the removal of seasonal nonstationarity, identifies significant two-state persistence in these data that is associated with climatic fluctuations detected within time series of other variables.

This chapter analysed the monthly records associated with the seven point-rainfall and one spatially-averaged rainfall series that were introduced analysed in Chapter 6. The District 66 rainfall provided a comparison to the calibration results for the Sydney series. The first part of

these analyses involved the calibration of two-state lognormal HMMs to the eight deseasonalised monthly rainfall series. This parametric form of HMMs was chosen in preference to the two-state Gaussian HMMs used in the previous chapter as the latter showed a tendency to degenerate to a mixture of Gaussians, a result that weakens the supposition of persistent climate states.

From the calibrations of two-state lognormal HMMs, the monthly rainfall from Adelaide, Sydney and Darwin showed evidence for significant persistence using the test derived in Chapter 6. Furthermore these calibrations estimated transition probabilities that were approximately equal, a feature that was deemed consistent with the assumptions of the broader climate fluctuating between two stable states. Calibrations of HMMs to the monthly data for Sydney achieved much tighter bounds around the series of posterior state probabilities than achieved using 12-month aggregations of these data. This clearly illustrates the benefit from the increase in information achieved from using hydrological data of higher frequency than annual totals. Importantly these calibrations identified persistence at a sub-annual scale, such that it could not be detected in time series of annual totals.

Although two-state lognormal HMMs identified significant persistence in various monthly rainfall series, a number of variations to this model were also introduced in this chapter to offer improved descriptions of persistence. The first variation was to explore alternative parametric forms for conditional distributions, with two-state gamma HMMs shown to be superior in a Bayesian context to lognormal HMMs for certain monthly rainfall series. Subsequently the number of model states was also investigated, and three-state lognormal HMMs and three-state gamma HMMs were also shown to be suitable models for hydrological persistence in these monthly rainfall data. Three-state models were not previously investigated as possible models for persistence in annual rainfall by Thyer and Kuczera (2000) or Srikanthan *et al.* (2002d), however Bayesian model selection identifies these as a suitable modelling approach. In later chapters, certain hydrological data are shown to have strong three-state persistence, which demands this very approach.

Analyses in this chapter made it apparent that model selection results cannot be used in isolation from model calibration results. The need for correct interpretation of modelling results was illustrated with the spatially-averaged monthly rainfall from the meteorological district surrounding Sydney. Although these data are highly correlated to the point rainfall data from Sydney, calibrations of two-state and three-state parametric HMMs to the two series produced very different results. Two-state lognormal and gamma HMMs tended to describe the entire District 66 series with a single distribution, whereas the Sydney data demonstrated persistence of similar strength in two climate states.

Furthermore although model selection results showed that a three-state lognormal HMM was the best model for the District 66 data, transition probability estimates indicated that there was only weak evidence for three-state persistence in these data. Through simulation studies, it was shown that incorrect assumptions about the form of state conditional distributions can lead to inaccurate estimates of both transition probabilities and underlying model states. This leads to the requirement for the form of conditional distributions to be assumed prior to HMM calibration. This poses a potentially difficult problem, however the next chapter introduces a technique that remedies this dilemma.

Chapter 8 Non-parametric hidden Markov models

A non-parametric (NP) HMM is developed in this chapter as an alternative to parametric HMM that have been utilised throughout the preceding two chapters. This model has been termed *non-parametric*, as no assumptions concerning the underlying state distributions are made. The non-parametric approach offers a more robust and flexible interpretation for HMMs, effectively re-sampling observed values to estimate state conditional distributions. This model provides unbiased estimates of state transitions, and is therefore a valuable method for identifying persistence within hydrologic observations.

8.1 Model structure

8.1.1 Background to the NP HMM

The procedure used to calibrate the NP HMM to a known data series is summarised in the following manner. Let $\{y_t, t = 1,2,...,T\}$ represent the observed data in time order, such that y_t is the datum at time t, with $1 \le t \le T$. Now let $\{y\}$ be the data sorted into ascending order. A transform into the (0, 1) interval is defined by $y_\tau \to \frac{m}{(T+1)}$ where m is the rank of y_τ when the observations are sorted. Define $u_\tau = \frac{m}{(T+1)}$, then $\{u_1, u_2, ..., u_T\}$ is the time series transformed into the (0, 1) interval. The hidden model states are identified within this transformed time series. The sorting procedure that generates the transformed time series can be undertaken on both discrete and continuous variables, such that the NP HMM can be fitted, without modification, to time series of either type.

Using the same model states as the parametric two-state HMMs described previously, it is assumed that a value u_τ has arisen from either a wet state (W) or a dry state (D) and that the higher values of u_τ are more likely to be from the former. For a given value of u_t, $1 \le t \le T$, the probability of that value having arisen from a wet state is complementary to the probability of that value having arisen from a dry state

$$P(x_t = W \mid u_t) + P(x_t = D \mid u_t) = 1 \qquad \textbf{(8.1)}$$

Transitions between states are defined in the same way as for the parametric HMM. The states of the transformed series are then represented geometrically by the partition of a unit square

having u_t on the abscissa and $P(s_t^j \mid u_t, \theta)$ on the ordinate (where $j = W, D$). The parameter vector θ is included here as a conditional vector of unknown quantities, which includes the HMM transition probabilities and a relationship governing the partition between the two states. This partition divides the square and defines the probabilities of u_t having been generated from either state. Figure 8.1 shows the unit square of the NP model, with a partition curve separating the wet and dry states.

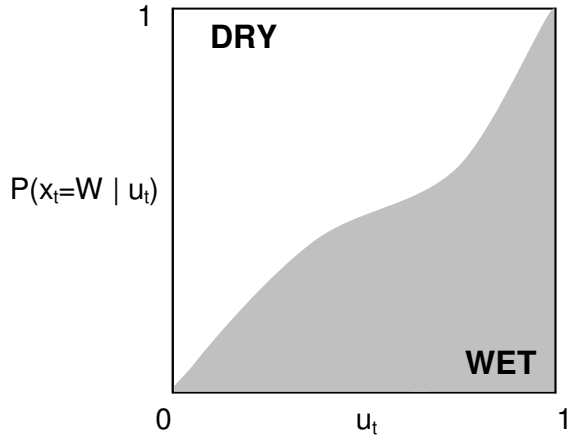

Figure 8.1 An example of the partition of a unit square into wet and dry model states

8.1.2 Partitioning the unit square

The partition of the square on which the NP HMM is based can take a variety of shapes dependent upon underlying state distributions. At one extreme, if the distributions are well separated, the partition will be a vertical line with areas either side corresponding to the proportion of the marginal series that is derived from each distribution. To illustrate this, data from a simulation of 10,000 draws are taken both from a wet state distribution $N(2000,200^2)$ and from a dry state distribution $N(1000,200^2)$, and placed into ascending order. The complimentary histograms in Figure 8.2a show the proportion of data from the dry and wet distributions for u_t bin widths of 0.01.

At the other extreme, the partition will be horizontal, with areas either side of the partition corresponding to proportions of the data from either distribution. As an illustration of this, 10,000 draws are taken from wet state and dry state distributions, both of which are $N(2000,200^2)$, and placed into ascending order. Figure 8.2b shows complimentary histograms with the proportion of data from both distributions for u_t bin widths of 0.01. With the ratio between the states approximately 1 within each histogram bin, this simulation indicates that there is equal probability for any value of u_t to have been generated from either model state.

To illustrate a partition that is between the horizontal and the vertical, data from a simulation of 10,000 draws are taken from two distributions $N(2000,200^2)$ and $N(1500,200^2)$, and placed into ascending order. Complimentary histograms from this simulation are shown in Figure 8.2c to change over the u_t values over the (0, 1) interval. These histograms show the proportion of data from wet and dry distributions using u_t bin widths of 0.01.

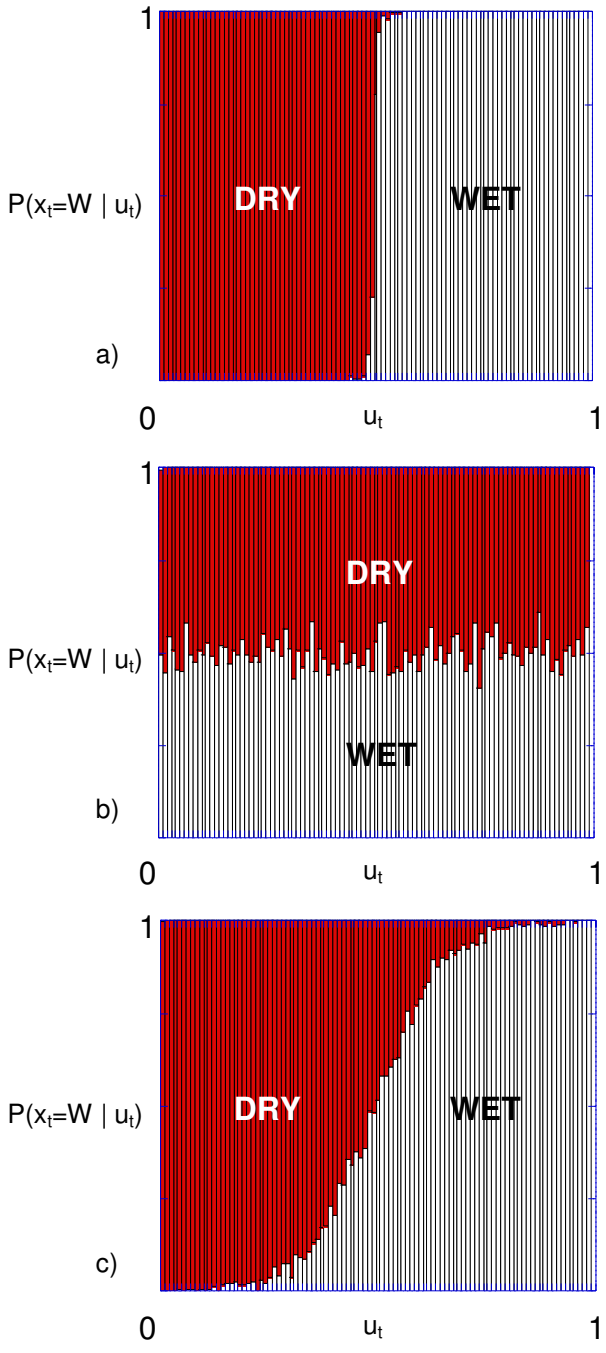

Figure 8.2 The estimated separation of unit squares from three simulated HMMs into wet and dry distributions, in equal proportion, that are a) well separated, b) identical, and c) overlapping

8.1.3 Approximating the partition

The division between the wet and dry histograms in Figure 8.2 can be characterised by a continuous curve. However, rather than specify a functional form of such a curve, a continuous division is effected by a discrete number of contiguous line segments. The approximation of a curve of partition by three segments, constrained at points (0,0) and (1,1), is determined by the coordinates of two points P_1 and P_2 as shown in Figure 8.3. A maximum likelihood procedure can determine the location of these two points, under the constraint that the coordinates of P_2 are greater than P_1. In this way, a wet state is identified as lying below these segments.

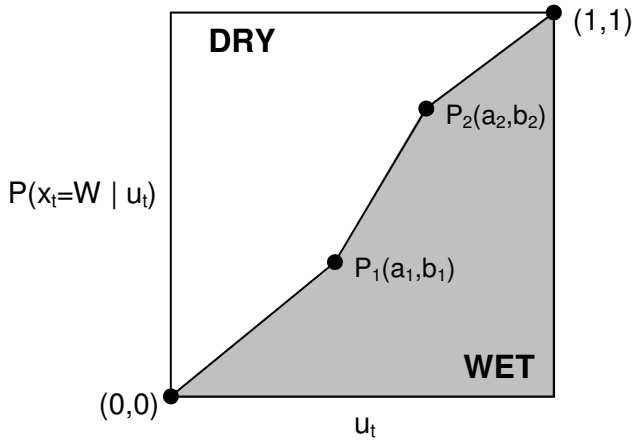

Figure 8.3 A two-point division of a unit square for a two-state NP HMM

It is possible to generalise the NP HMM to include more than two states. In the model illustrated here, two independent states are defined by one partition. Extending the model to allow for three model states requires two partitions and the locations of four points to be identified, as shown in Figure 8.4.

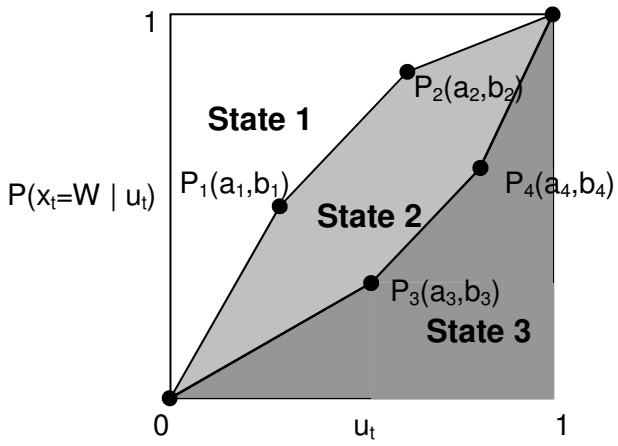

Figure 8.4 A two-point division of a unit square for a three-state NP HMM

In the three-state formulation, an orthogonal line through any point on the abscissa must pass firstly through a wet state (state 3 in Figure 8.4) before a neutral state (state 2) and then dry state

(state 1). It is a straightforward procedure to extend the model further to allow for more than three states. An alternative two-state model approach is to approximate the partition curve by a greater number of line segments. For example, ten contiguous lines are separated by nine points, which can be equally spaced along the abscissa. A maximum likelihood procedure determines the coordinates of these nine points, under the constraint that the distance of each point from the abscissa is not less than that of the preceding point. This possible "nine-point" model is shown in Figure 8.5, with the ten line segments again constrained at points (0,0) and (1,1).

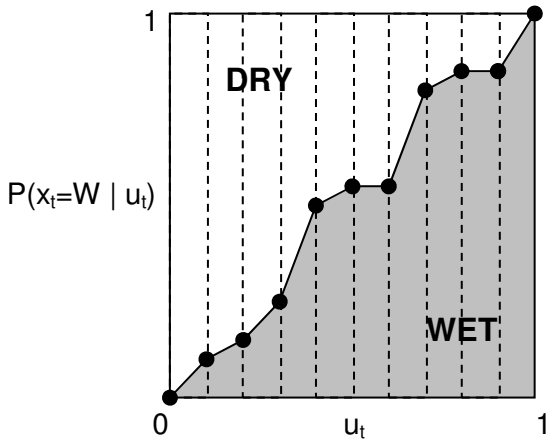

Figure 8.5 A nine-point division of a unit square for a two-state NP HMM

In both the "two-point" model and the "nine-point" model, the segment of a line through the point u_t between the abscissa and the partition curve is $P(x_t = W \mid u_t, \theta)$, and the length of this line above the partition is $1 - P(x_t = W \mid u_t, \theta)$, or $P(x_t = D \mid u_t, \theta)$. The parameter vector θ includes the locations of the points that define the partition curve. The two-point and nine-point models can be compared through the Bayesian model selection techniques described in Section 3.4.3. Applications of the NP HMM in this thesis are focused on two-point divisions only.

8.1.4 Likelihood function for the NP HMM

In the likelihood function of the conventional HMM described in Section 3.3.4, the probability density function (pdf) of observations is represented by $p(y_t \mid s_t^j, \theta)$ when showing a conditional relationship to unknown model parameters. This density is related to the density of the NP HMM through the uniform transformation. The non-parametric HMM pdf in state j $(j = W, D)$ at time t is represented as $p(u_t \mid s_t^j, \theta)$, which can be expressed using Bayes' theorem as

$$p(u_t \mid s_t^j, \theta) = \frac{P(s_t^j \mid u_t, \theta) \times p(u_t \mid \theta)}{P(s_t^j \mid \theta)} \tag{8.2}$$

Here, $P(s_t^j \mid u_t, \theta)$ is the probability of u_t being in state j at time t, which is also the distance from the abscissa to the partition at that point. Also $p(u_t \mid \theta)$ is the marginal pdf for u_t, which is equal to unity since the transformed series is mapped to a uniform distribution. Furthermore, $P(s_t^j \mid \theta)$ is the marginal probability of each state (the steady-state probabilities P_{wet} and P_{dry}) given the shape of the partition curve. These marginal probabilities are equal to the proportions of the uniform distribution described as wet or dry by the partition and are related to the transition probabilities. The marginal state probabilities are associated through $P_{wet} = 1 - P_{dry}$, such that only one transition probability in the NP HMM requires estimation, with the other being described through the location of the partition.

A three-state HMM is described by six transition probabilities; P_{WN}, P_{WD}, P_{NW}, P_{ND}, P_{DW} and P_{DN} using N as an abbreviation for the neutral climate state. If the sum of these six probabilities is simplified as ΣTP, the three steady-state probabilities are described as

$$P_{wet} = \frac{P_{NW} + P_{DW}}{\Sigma TP}$$

$$P_{neutral} = \frac{P_{WN} + P_{DN}}{\Sigma TP} \tag{8.3}$$

$$P_{dry} = \frac{P_{WD} + P_{ND}}{\Sigma TP}$$

These probabilities lead to the relationship

$$\frac{P_{NW} + P_{DW}}{P_{wet}} = \frac{P_{WN} + P_{DN}}{P_{neutral}} = \frac{P_{WD} + P_{ND}}{P_{dry}} \tag{8.4}$$

Rearranging this equality produces

$$P_{NW} = \frac{P_{wet}}{P_{neutral}}(P_{WN} + P_{DN}) - P_{DW} \tag{8.5}$$

and

$$P_{ND} = \frac{P_{dry}}{P_{neutral}}(P_{WN} + P_{DN}) - P_{WD} \tag{8.6}$$

Using these relationships, it is clear that in a three-state NP HMM, only four transition probabilities need to be estimated, with others being described through the locations of the two partitions.

By using Bayes' theorem, the partition is scaled to form a pdf that can be incorporated into the standard HMM likelihood function of Section 3.3.4. The SCE algorithm is used with the NP likelihood function to identify maximum likelihood estimates of model parameters, given suitable prior distributions. MCMC procedures are utilised as with the parametric HMM in order to estimate the posterior distribution of unknown parameters $p(\theta|U_T)$ where U_T represents the entire series of transformed observations.

8.1.5 Identification and estimation of state conditional distributions

Following the identification of the partition, state conditional distributions are estimated with a Monte Carlo sampling procedure that is summarised in the following four steps

- A uniform random number is generated, relating to a value u_τ and its corresponding position on the abscissa of the NP HMM unit square
- A corresponding value (y_τ) from the original un-transformed time series is interpolated from the transformed value u_τ.
- A second uniform random number (p_τ) is generated and yields a distance along the orthogonal line through u_τ on the square.
- If p_τ lies above the partition, then the value of y_τ is assigned to the dry state distribution, and vice versa for the wet state.

Estimates of the two underlying state distributions are obtained by repeating this sampling procedure multiple times (2,000 repetitions are used to obtain the results shown throughout this work, unless otherwise noted). These distributions are guided only by the location of the partition line, which is identified through a maximum likelihood procedure. In this way, assumptions about the underlying state distributions in the two climate states are prevented. This is the key advantage of the non-parametric HMM methodology, which avoids deficiencies of the modelling approach described in Chapter 7 that impeded the calibration to monthly rainfall data. Estimates of state conditional distributions are made using maximum likelihood estimates and posterior medians of the partition location parameters.

8.2 Calibration of NP HMMs to various simulated data series

The accuracy of the NP HMM to identify two-state persistence in a range of continuous and discrete data is demonstrated through the calibration of the model to time series generated from a range of stochastic models. Before interpreting results from the calibration of the NP HMM to

observed data, it is useful to have familiarity with the accuracy of the NP HMM to identify known model parameters from simulated time series. The simulated series used in these analyses share a common length of 1000 values, approximating the length of continuous monthly rainfall series analysed previously.

8.2.1 Time series simulated from two-state Gaussian HMMs

A time series of length 1000 was simulated from a two-state Gaussian HMM with conditional distributions $N(1500,200^2)$ and $N(1000,200^2)$, and transition probabilities $P_{WD} = 0.20$ and $P_{DW} = 0.20$. After calibrating the NP HMM to these data, wet and dry distributions are estimated, using posterior medians of partition parameters, to have means 1510.4 and 1026.6 and standard deviations 207.3 and 209.4 respectively. Furthermore these estimated distributions approximate a series of Gaussian-distributed variables as shown in Figure 8.6. Posterior distributions for the two transition probabilities, P_{WD} and P_{DW}, have median values 0.214 and 0.200 with 90% credibility intervals of (0.179, 0.252) and (0.167, 0.235) respectively. The NP HMM identifies the known transition probability values within 90% credibility intervals, and estimates accurately the parameters of the state conditional distributions.

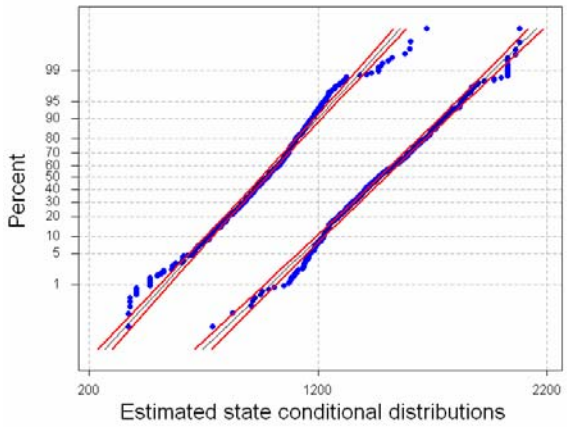

Figure 8.6 Gaussian probability plot showing estimates for conditional model states for a simulation of length 1000

For comparison, a Gaussian HMM was calibrated to the same simulated series and posterior estimates for the HMM parameters are shown in Table 8.1. As the assumptions of the parametric model are valid, it is expected that the intervals around transition probability estimates would be narrower than for the NP HMM. However the influence of sample error on the simulation of the HMM series leads to inaccuracies in parameter estimation, such that transition probability estimates are less accurate for the parametric model. These results indicate that the NP HMM is an accurate and efficient implementation of the HMM framework for modelling two-state persistence.

Table 8.1 Posterior medians and 90% credibility intervals for parameters of a two-state Gaussian HMM fitted to the series simulated from a two-state Gaussian HMM

P_{WD}	P_{DW}	μ_W	μ_D	σ_W	σ_D
0.215	**0.191**	**1483.3**	**966.2**	**208.1**	**188.3**
(0.178, 0.254)	(0.159, 0.225)	(1459, 1505)	(947, 987)	(194, 224)	(177, 201)

The influence of the length of the input series upon the ability of the NP HMM to identify parameters accurately is now investigated by simulating a second time series, using the same parameter values as earlier, yet over the shorter length of only 100 values. This length reflects annual rainfall time series. This calibration produces posteriors for P_{WD} and P_{DW} that have medians 0.192 and 0.272, with 90% credibility intervals of (0.037, 0.466) and (0.088, 0.698) respectively. The large uncertainty around these parameters demonstrates that shorter time series create difficulties for the NP HMM to identify accurately the underlying series of model states. This has obvious implications for the identification of climatic persistence within annual rainfall time series. The increase in parameter uncertainty is associated with inferior estimation of state conditional distributions, as indicated in Figure 8.7.

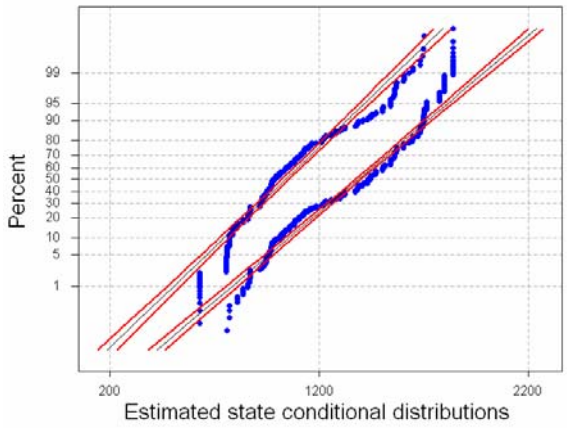

Figure 8.7 Gaussian probability plot showing estimates for conditional model states for a simulation of length 100

The following simulation investigates the performance of the NP HMM to identify known model states that are associated with skewed conditional distributions.

8.2.2 Time series simulated from a two-state lognormal HMM

For this simulation, a time series of length 1000 was simulated from a two-state HMM with transition probabilities $P_{WD} = 0.3$ and $P_{DW} = 0.1$, such that the stationary distribution of the Markov chain will have a majority of values in the dry state. Random samples for the wet and dry states are initially drawn from $N(7.5, 0.35^2)$ and $N(6.8, 0.25^2)$, with exponentials of these samples then taken to produce two conditional lognormal distributions. After calibrating a two-

point NP HMM to these data, posterior distributions for estimates of P_{WD} and P_{DW} have respective medians of 0.254 and 0.082, with 90% credibility limits of (0.188, 0.332) and (0.061, 0.109) that contain the values used in simulation. Although the assumptions of parametric models are infringed by the NP framework, transition probabilities are nonetheless estimated accurately by this model. The state conditional distributions, estimated from posterior medians of location parameters, approximate random draws from lognormal distributions, as demonstrated in Figure 8.8. Taking logarithms of these estimates produces wet and dry distributions that have means 7.457 and 6.805, and standard deviations 0.373 and 0.249 respectively. These results indicate that the NP HMM estimates accurately the underlying skewed distributions; a useful result for modelling persistence in monthly hydrologic data.

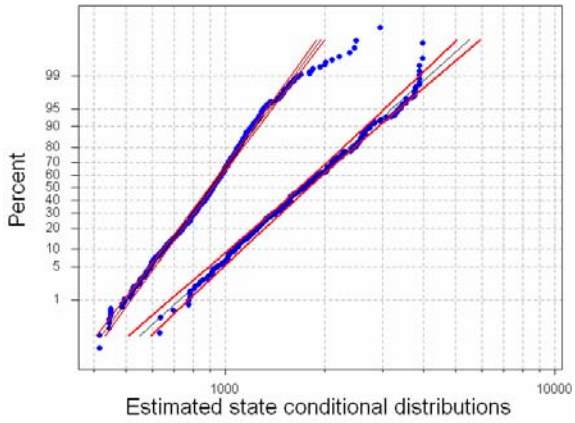

Figure 8.8 Lognormal probability plot showing estimates for conditional model states for a simulation of length 1000

8.2.3 Time series simulated from a two-state HMM having one Gaussian and one lognormal conditional distribution

For this third example, a time series of length 1000 is again constructed from a two-state HMM with transition probabilities $P_{WD} = P_{DW} = 0.20$. Random samples in a dry state were drawn from $N(1000, 200^2)$, with wet state samples being the exponentials of random draws from $N(7.5, 0.35^2)$. Therefore, this simulation allows the form of state conditional distributions to vary, and tests the ability of the NP model to estimate these changes and identify accurately HMM transition probabilities.

Calibrating the NP HMM to this series produces posterior medians of 0.215 and 0.199 for P_{WD} and P_{DW} respectively, with 90% credibility intervals of (0.172, 0.263) and (0.162, 0.238). This result indicates that a series with state conditional distributions of two different parametric forms fails to negatively impact upon the NP HMM estimation. The two state conditional

distributions are estimated with posterior medians of partition parameters and shown in Figure 8.9. These plots illustrate dry state samples approximating a series of random draws from a Gaussian distribution with mean 1017.3 and standard deviation 224.4. Wet state samples approximate a series of lognormally-distributed variates with their logarithms having mean 7.512 and standard deviation 0.362. Estimates of underlying conditional distributions are consistent with the simulated series, departing from expected distributions most distinctly in their tails. The dry state distribution has a tendency for a slightly heavier upper tail than would be expected for a series of Gaussian variates. The ability of the NP HMM to estimate discrete-valued conditional distributions accurately is analysed in the following simulation.

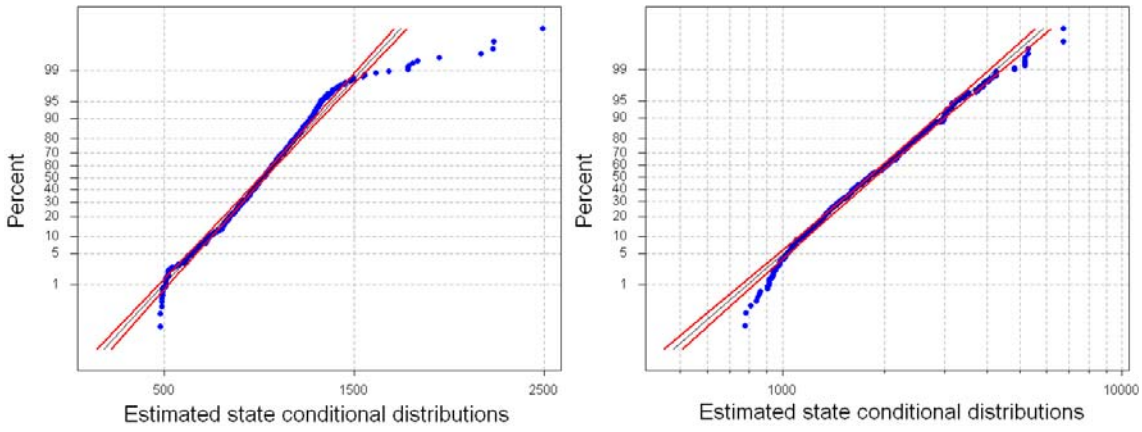

Figure 8.9 Estimated conditional distributions for dry state shown on Gaussian probability plot (left) and wet state shown on lognormal probability plot (on right)

8.2.4 Time series simulated from a two-state Poisson HMM

In this fourth example a discrete-valued series is simulated from a two-state HMM, with random samples in a wet state being drawn from Poisson distribution with a mean of 13 and dry state samples being drawn from a Poisson distribution with mean 8. Again, this simulation was undertaken with transition probabilities $P_{WD} = P_{DW} = 0.20$. The marginal distribution of this simulated series is shown in the histogram of Figure 8.10.

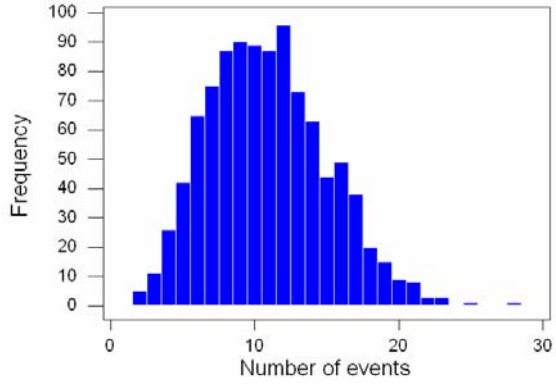

Figure 8.10 Marginal distribution of a series simulated from a two-state Poisson HMM

After calibrating the NP HMM to this series, 2000 random samples around posterior medians of partition locations were used to generate estimates of the two conditional distributions, displayed in Figure 8.11. These wet and dry distributions have means 13.05 and 8.16, and variances 13.20 and 8.46 respectively. The estimates of the variances are consistent with Poisson distributions for which the mean equals the variance. The posteriors for P_{WD} and P_{DW} have median values of 0.182 and 0.219, with 90% credibility limits of (0.126, 0.249) and (0.158, 0.298). These two credibility intervals contain the values used in simulation.

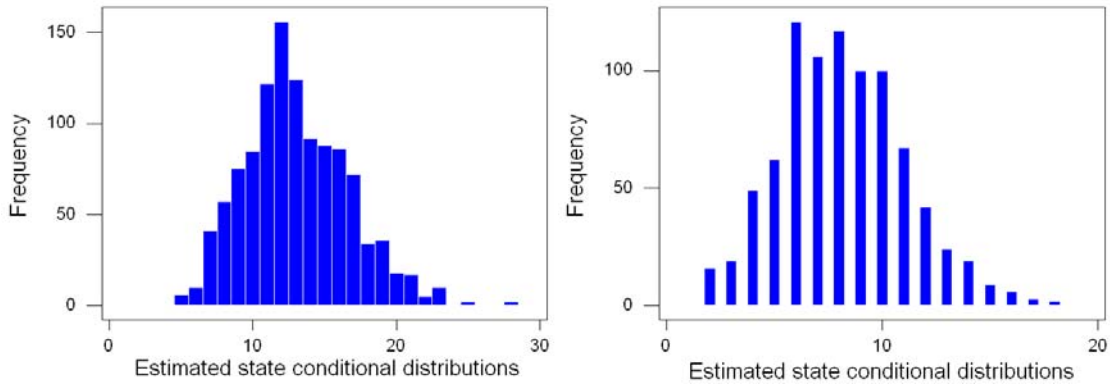

Figure 8.11 Histograms showing state conditional distributions for the wet state (on left) and dry state (on right)

When a two-state HMM with Poisson state conditional distributions is calibrated to this series, posterior distributions for P_{WD} and P_{DW} have median values of 0.179 and 0.229, with 90% credibility limits of (0.131, 0.240) and (0.172, 0.292) that are marginally narrower than those estimated by the NP HMM, which is expected for models having valid parametric assumptions. The wet and dry state means are estimated with posterior medians of 13.03 and 8.12 respectively, with 90% credibility intervals of (12.62, 13.50) and (7.68, 8.56).

8.2.5 Time series simulated from a three-state Gaussian HMM

The various simulations described in this section have shown the NP HMM to be suitable for identifying underlying two-state persistence in a range of data. To determine the efficacy of this model to identify three-state persistence, a time series of length 1500 is now simulated from a three-state Gaussian HMM with distributions $N(1500,300^2)$, $N(1000,200^2)$, $N(500,100^2)$, and transition probabilities $P_{WN} = P_{DN} = 0.3$, $P_{WD} = P_{DW} = 0.1$ and $P_{NW} = P_{ND} = 0.25$. Three-state parametric and non-parametric HMMs are calibrated to this series.

The posterior distributions of transition probabilities from the calibration of a three-state Gaussian HMM to this series are summarised in Table 8.2. These results indicate that the parametric model correctly estimates the six transition probabilities within 90% credibility intervals; hence it identifies the three-state persistence in this simulated series.

Table 8.2 Posterior medians and 90% credibility intervals for transition probabilities of a three-state Gaussian HMM calibrated to the series simulated from a three-state Gaussian HMM

P_{WN}	P_{WD}	P_{NW}	P_{ND}	P_{DW}	P_{DN}
0.287	**0.082**	**0.246**	**0.245**	**0.101**	**0.339**
(0.25, 0.33)	(0.06, 0.11)	(0.21, 0.28)	(0.21, 0.29)	(0.07, 0.13)	(0.28, 0.40)

The posterior distributions for estimates of the conditional state distribution parameters from this model calibration are summarised in Table 8.3. Once again, the three-state Gaussian HMM accurately estimates each parameter within 90% credibility intervals, which is expected as the modelling assumptions are justified.

Table 8.3 Posterior medians and 90% credibility intervals for other parameters for a three-state Gaussian HMM calibrated to the series simulated from a three-state Gaussian HMM

μ_W	σ_W	μ_N	σ_N	μ_D	σ_D
1521	**232**	**1010**	**156**	**502**	**106**
(1473, 1560)	(208, 263)	(994, 1028)	(142, 172)	(493, 512)	(99, 113)

The three-state NP HMM, which makes no assumptions about the form of state conditional distributions, is now calibrated to this simulated series. Posterior distributions for the six transition probabilities are summarised in Table 8.4, which shows that the correct value of three of these probabilities are identified within 90% credibility intervals. These posterior distributions have slightly wider credibility intervals than those summarised in Table 8.2, which is an accepted consequence of using the non-parametric approach.

Table 8.4 Posterior medians and 90% credibility intervals for transition probabilities of a three-state NP HMM calibrated to the series simulated from a three-state Gaussian HMM

P_{WN}	P_{WD}	P_{NW}	P_{ND}	P_{DW}	P_{DN}
0.210	**0.096**	**0.167**	**0.247**	**0.132**	**0.189**
(0.16, 0.26)	(0.06, 0.14)	(0.12, 0.22)	(0.20, 0.30)	(0.09, 0.17)	(0.13, 0.25)

After taking 3000 random samples around posterior medians for partition locations, estimates of the three conditional distributions from this calibration are shown in the probability plot of Figure 8.12. These 3000 samples are divided in a wet: neutral: dry ratio of 813: 1169: 1000, which contrasts to the expected value of 923: 1154: 923 obtained from the steady-state probabilities of each state. In Figure 8.12, the wet and neutral state distributions more closely approximate straight lines, with some high values in the dry state producing a much heavier tail than would be expected for a Gaussian distribution. The estimated wet and neutral distributions have sample means of 1488 and 1038 and sample standard deviations 293 and 235 that are slightly greater than values used in simulation. Their respective Anderson-Darling goodness-of-

fit statistics of 3.90 and 10.34 are much lower than the 62.31 of the dry state distribution, indicating a closer approximation to series of Gaussian variates, although these are still greater than would be expected for series of Gaussian variates. Interestingly of the 1000 points in the dry state estimate, the 919 that have a value below 800 have a sample mean of 503 and sample standard deviation of 105, with an AD statistic of 0.62 that is consistent with a series of normally-distributed variates. This illustrates that accurate estimation of state conditional distributions may be limited in certain NP HMM applications.

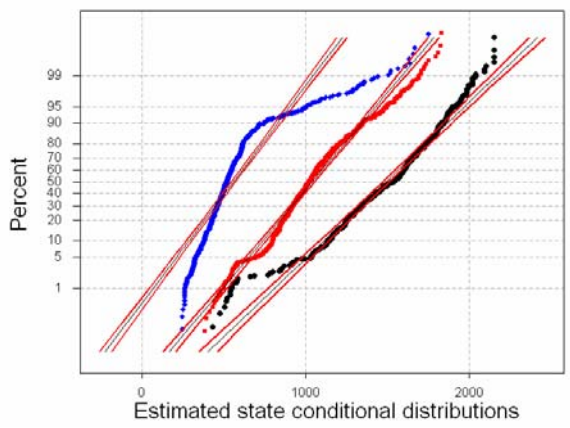

Figure 8.12 Gaussian probability plot showing estimates for conditional model states from the calibration of a three-state NP HMM to a simulated series of length 1500

The series of most likely states from the calibration of both the three-state NP HMM and three-state Gaussian HMM closely match the true state sequence (correlations $r = 0.90$ in each case). Furthermore the state series from the two models are also highly correlated $(r = 0.97)$, so without making assumptions about the form of conditional distributions, the NP HMM can identify the correct three-state persistence.

This point is further illustrated through the calibration of a two-state NP HMM to this series, and evaluating the Bayes Factor relative to the three-state NP HMM. This model comparison obtains $\ln BF_{3NP,2NP} = 20.4$, which demonstrates a superiority of the three-state model. This example has shown that the three-state NP HMM can correctly identify three-state persistence within a simulated time series. In the following section, both two-state and three-state NP HMMs are calibrated to monthly rainfall data from across Australia.

8.3 Calibration of NP HMMs to Sydney and District 66 rainfall

Having demonstrated that NP HMMs provide accurate estimates of HMM parameters and state conditional distributions within simulated data, these models are now calibrated to the Sydney and District 66 rainfall. The distribution of the sum of transition probabilities is used as an

unbiased test of the hypothesis that annual and monthly rainfall data are influenced by persistent climate states.

8.3.1 Calibration of two-state NP HMMs to annual rainfall data

The two-state NP HMM is calibrated to the time series of annual rainfall totals from both Sydney and District 66. The results shown in Section 6.4 suggested that a two-state Gaussian HMM degenerated to a mixture of two Gaussian distributions when calibrated to both series, with 90% credibility intervals around estimates of $P_{WD} + P_{DW}$ including a value of 1 in each case. Figure 8.13 shows state conditional distributions from the calibration of the NP HMM to the Sydney annual rainfall series, obtained from 2000 random samples around posterior medians of the partition locations. These distributions are approximate straight lines on a Gaussian probability plot, indicting that the assumption of a two-state Gaussian HMM is valid.

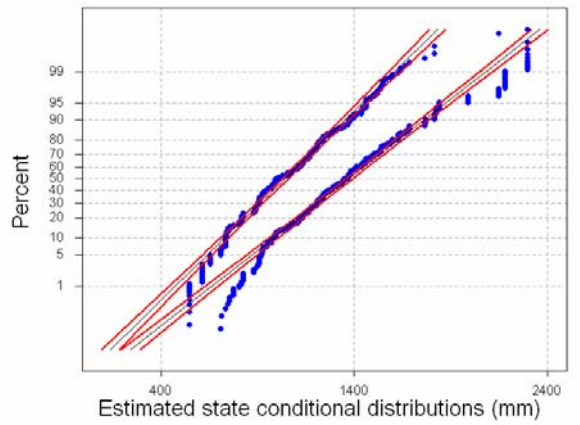

Figure 8.13 Gaussian probability plot showing estimates for conditional model states from the calibration of a two-state NP HMM to the annual rainfall for Sydney

Table 8.5 shows a comparison of transition probability estimates obtained from the calibration of both the two-state Gaussian HMM and two-state NP HMM models. The NP HMM estimates for the point data are similar to those for the spatial data, although wide credibility intervals mean that the sum of transition probabilities includes a value of 1. Consequently, even through the removal of modelling assumptions for state conditional distributions approximating a series of Gaussian variates, the annual rainfall series fails to demonstrate significant two-state persistence.

Table 8.5 Posterior medians and 90% credibility intervals for transition probabilities from calibrating two-state Gaussian HMMs and NP HMMs to annual rainfall series

		P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Sydney	Gaussian HMM	**0.530** (0.139, 0.922)	**0.200** (0.030, 0.880)	**0.833** (0.334, 1.415)
	NP HMM	**0.383** (0.089, 0.752)	**0.491** (0.112, 0.935)	**0.944** (0.276, 1.443)
District 66	Gaussian HMM	**0.516** (0.194, 0.908)	**0.150** (0.038, 0.521)	**0.705** (0.320, 1.176)
	NP HMM	**0.361** (0.090, 0.760)	**0.323** (0.087, 0.905)	**0.738** (0.235, 1.462)

The effect of high uncertainty around estimates for the transition probabilities in the Sydney data is illustrated through the posterior state series shown in Figure 8.14. High median estimates for P_{WD} and P_{DW} in Table 8.5 manifest themselves in a tendency for the state series to have alternating high and low values, indicative of a lack of persistence in either model state. Moreover the large interval around these state probabilities indicates uncertainty about the most likely model state at each time step. The posterior state series for District 66, although not shown, has a general form that is similar to Figure 8.14. The median state series for District 66 is significantly correlated $(r = 0.85)$ to the median state series for Sydney.

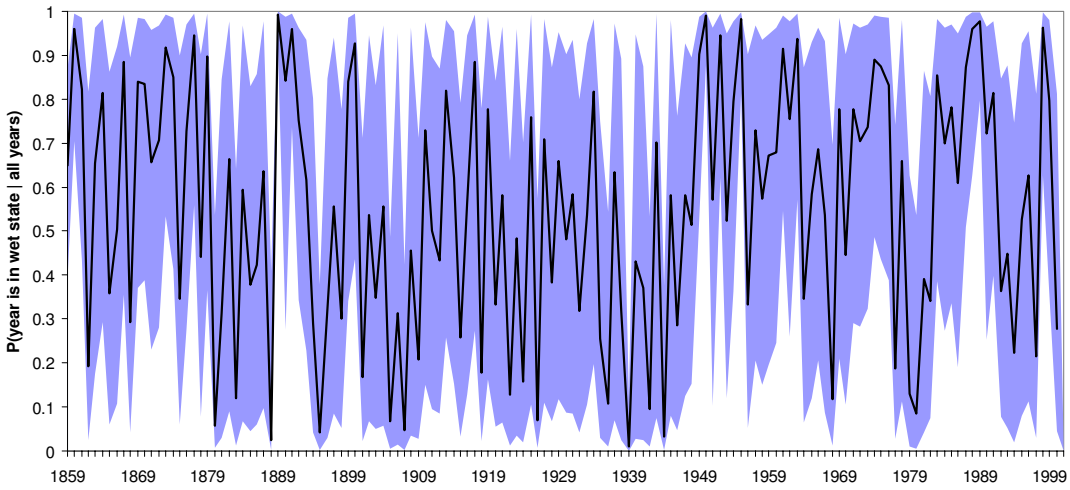

Figure 8.14 Median state series and 90% credibility interval from th calibration of a two-state NP HMM to the annual rainfall for Sydney

Whereas Figure 6.6 showed the posterior median probabilities from the calibration of a two-state Gaussian HMM to the annual rainfall series for Sydney remaining in a dry state for a majority of years, this construction is not apparent in the state series derived from the NP HMM. If this multi-decadal dry period was a dominant feature of persistence in the climate that influences Sydney, it is likely that this would also be evident in the state series of the NP HMM,

which estimates state conditional distributions to approximate random Gaussian draws. Consequently Figure 8.14 provides further evidence against significant climatic persistence at an annual time scale. As runs analyses suggested, the monthly time scale may be more appropriate for identifying persistence in time series such as Australian rainfall.

8.3.2 Calibration of two-state NP HMMs to monthly rainfall data

The two-state NP HMM is now calibrated to the deseasonalised monthly rainfall series for Sydney. The lognormal HMM is suitable for identifying two-state persistence in these data, although when calibrated to the time series of district-averaged observations, this model identified a wet state distribution with parameters that matched the marginal. Transition probabilities showed the lognormal HMM to remain in this single wet state, with an estimate of $P_{WD} + P_{DW}$ showing the model degenerates to a mixture of two lognormal distributions. The NP HMM is now used to determine whether relaxing assumptions about the form of conditional distributions improves the identification of persistence within these series. Posterior samples are estimated from 6,000 samples taken from 10 Markov chains in the Adaptive Metropolis algorithm. Convergence diagnostics using the variance ratio method are summarised in Table 8.6, indicating that this combination of independent Markov chains and samples is sufficient to obtain accurate estimates from the posterior distributions.

Table 8.6 Convergence diagnostics from using the variance ratio method to analyse 6,000 samples from 10 Adaptive Metropolis chains to obtain estimates of posterior distributions for a two-state NP HMM calibrated to deseasonalised monthly Sydney rainfall

	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$	$n = 2000$	$n = 2500$	$n = 3000$
P_{DW}	2.434	1.087	1.035	1.034	1.042	1.020	1.009
a_1	1.362	1.109	1.031	1.059	1.060	1.038	1.014
a_2	1.467	1.282	1.078	1.071	1.025	1.034	1.019
b_1	1.926	1.173	1.041	1.049	1.047	1.023	1.009
b_2	1.486	1.205	1.035	1.034	1.022	1.013	1.009
Average	**1.735**	**1.171**	**1.044**	**1.049**	**1.039**	**1.026**	**1.012**

Table 8.7 compares posterior estimates of transition probabilities from the calibrations of both the NP HMM and the lognormal HMM to the Sydney monthly rainfall series. These results show estimates of transition probabilities to be similar for both the NP HMM and lognormal HMM, with neither model showing credibility intervals of $P_{WD} + P_{DW}$ including a value of 1. This is a useful result, as the NP HMM would be expected to identify the two-state persistence previously isolated with the lognormal HMM.

Table 8.7 Comparison of posterior distributions for transition probabilities, showing medians and 90% credibility intervals, from the calibration of a two-state lognormal HMM and two-state NP HMM to the deseasonalised monthly rainfall of Sydney

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Lognormal HMM	**0.343** (0.294, 0.490)	**0.379** (0.215, 0.506)	**0.726** (0.588, 0.826)
NP HMM	**0.301** (0.208, 0.406)	**0.420** (0.305, 0.539)	**0.728** (0.592, 0.841)

The posterior state series derived from calibrating the NP HMM to the Sydney monthly rainfall is shown in Figure 8.15, using the same 10-year period as in Figure 7.6 to show the lognormal HMM state series. These figures show that the state series from both models are closely related, and over the 1716 months of this monthly rainfall series, the series of median state probabilities from both models are strongly correlated $(r = 0.995)$. State estimation with the NP HMM is slightly better than that achieved with the parametric model however, with the 90% credibility bound around the posterior state series having an average size of 0.306 as opposed to an average of 0.393 for the lognormal HMM.

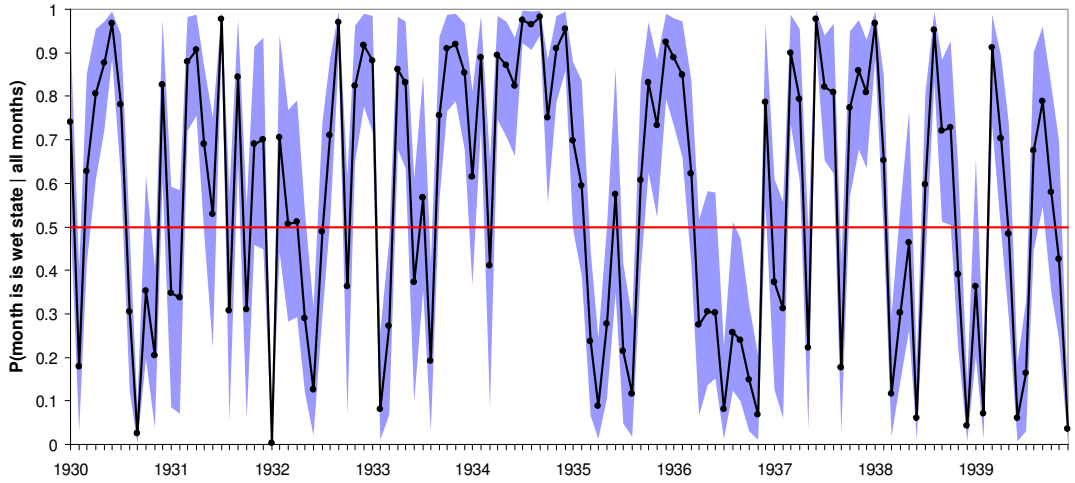

Figure 8.15 Median state series and 90% credibility interval for a 10-year period from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for Sydney

The strong relationship between estimates from the NP and lognormal HMMs for transition probabilities and the median state series suggests that the assumption of lognormal conditional distributions in this series is appropriate. This is further demonstrated through observing estimates for these distributions, as shown in Figure 8.16. After the calibration of the NP HMM to obtain posterior medians for the partition locations, 2000 random draws were used to estimate wet and dry state distributions. These distributions (with 1186 values in the wet state and 814 in the dry) are approximate straight lines in the lognormal probability plot of Figure 8.16. With

both distributions having low values for the Anderson-Darling (AD) goodness-of-fit statistic (1.47 for wet and 0.75 for dry), it is clear that the assumption of lognormal distributions is consistent for this series.

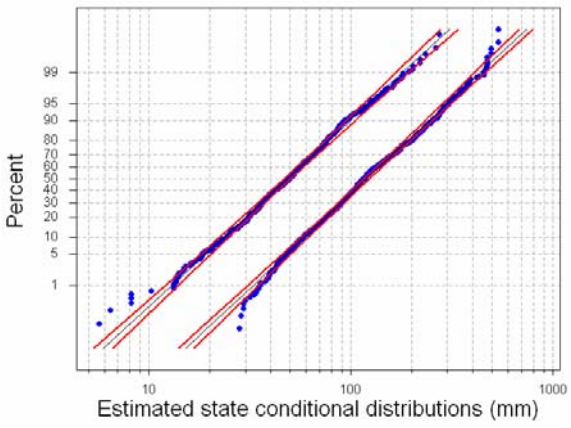

Figure 8.16 Lognormal probability plot showing estimates for conditional model states from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for Sydney

After taking logarithms of the two conditional distributions, the wet state distribution in Figure 8.16 is approximately distributed as $N(4.78,0.56^2)$ with the dry state distributed as $N(3.88,0.57^2)$. These estimates are similar to the posterior medians from the calibration of the lognormal HMM to the Sydney data, suggesting these models identify comparable two-state persistence. This is a clear benefit of the NP HMM, by which it can identify suitable parametric families from which state conditional distributions can be simulated. The use of the NP HMM, which makes no assumption about climate state distributions, provides a clear advantage over parametric models in the identification of these distributions in hydrologic data. The possible loss in efficiency in using a non-parametric model is compensated for by using the results of its calibration to fit a suitable parametric HMM.

The NP HMM is now fitted to the deseasonalised monthly rainfall series for District 66. With the lognormal and gamma HMMs both shown to be inappropriate descriptors for these data, the NP HMM can determine whether the assumption of parametric forms for the state conditional distributions masks significant two-state or three-state persistence. Table 8.8 shows estimates for the two transition probabilities from calibrating both a lognormal HMM and a NP HMM to the District 66 data. These results show that the NP HMM identifies a very different persistence structure within these data.

Table 8.8 Comparison of posterior distributions for transition probabilities, showing medians and 90% credibility intervals, from the calibration of a two-state lognormal HMM and two-state NP HMM to the deseasonalised monthly rainfall for District 66

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Lognormal HMM	**0.040** (0.008, 0.193)	**0.642** (0.347, 0.959)	**0.706** (0.450, 1.014)
NP HMM	**0.248** (0.108, 0.387)	**0.460** (0.317, 0.627)	**0.721** (0.521, 0.873)

The posterior medians for P_{WD} and P_{DW} from the calibration of the NP HMM show that on average, 65% of months are in a wet state with 35% in the dry. This is a vastly different result than that achieved from the calibration of a two-state lognormal HMM, which estimated 94% of months to be in a wet state. Importantly, the sum of transition probabilities has a 90% interval that does not include 1, thereby rejecting the notion that the HMM degenerates to a mixture distribution at a 10% significance level. Furthermore with the median value of $P_{WD} + P_{DW}$ being very close to the value estimated from the Sydney monthly series; a similar pattern of persistence is identified in these two series. This result is consistent with a hidden climate indicator influencing hydrologic observations across the Sydney region, and this is expected as the district-averaged series is derived from point rainfall data that includes the Sydney gauge.

The posterior state series for the District 66 data identified by the NP HMM is shown in Figure 8.17 over the same 10-year period that was used for the Sydney data. It is clear that this state series identifies a different pattern of persistence to that obtained from the calibration of a lognormal HMM in Figure 7.17. While the lognormal HMM state series showed months being predominantly within a wet state, the NP HMM state series has various changes in state that parallel the changes identified within the Sydney data. Over the period 1913-2001, the median state series identified by the NP HMM for the Sydney and District 66 data are strongly correlated $(r = 0.938)$, providing further evidence for consistent two-state persistence. The two-state persistence of the district data is less accurate than the point rainfall series however, with credibility intervals being slightly wider (average interval 0.393).

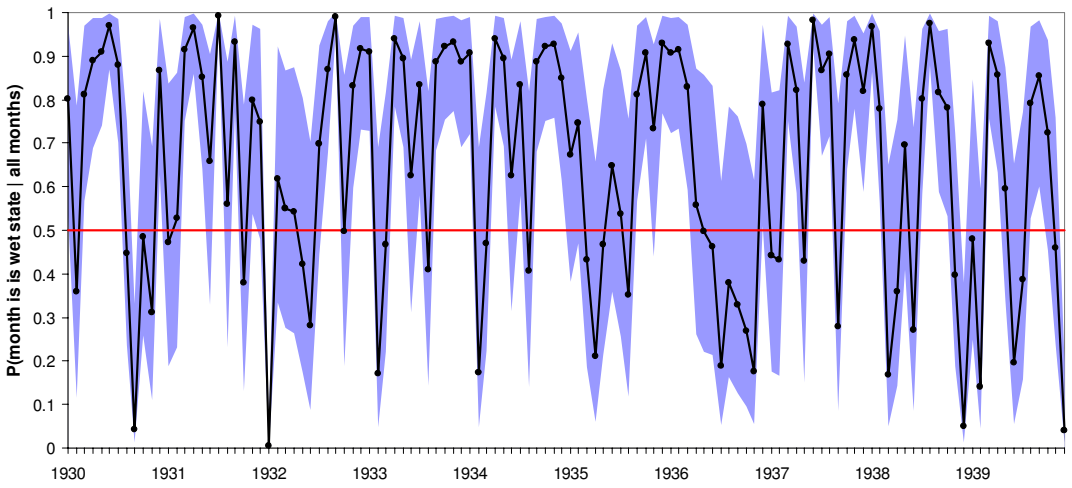

Figure 8.17 Median state series and 90% credibility interval for a 10-year period from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for District 66

With the NP HMM identifying hydrological persistence in the District 66 data, it is important to investigate reasons for the inefficiency in assuming lognormal state conditional distributions. Figure 8.18 shows the estimated state conditional distributions from the District 66 data that are obtained using 2000 random samples around posterior medians for partition parameters. From these random samples, 1376 were estimated to be in a wet state and 624 in a dry, producing a ratio of 0.69:0.31 that is close to the stationary distribution of the Markov chain. Although these two states seem to plot as straight lines, the dry state is weaker in its lower tail than would be expected for a series consistent with random draws from a lognormal distribution. This lighter tail produces a slightly higher Anderson-Darling goodness-of-fit statistic in the dry state than for the wet state (3.77 as opposed to 3.64), both of which are slightly higher than expected for a series of lognormal variates of this length.

It is possible that the two conditional distributions are better modelled by gamma distributions, and this can be determined by calculating the AD statistic using parameters for gamma cdfs estimated from the conditional samples. This shows that although the dry state samples have a lower AD statistic (2.81) for a gamma distribution, this value remains higher than would be expected for a series of gamma variates. Conversely, the wet state samples are closer to a series of lognormal variates, with an AD statistic (22.77) from a gamma cdf that is significantly higher. Although the lognormal is a better overall choice than the gamma as the single parametric form to describe the two conditional distributions in the District 66 data, the goodness-of-fit statistics show that this assumption may still be inaccurate, and this may explain the misleading results that are obtained through the calibration of a two-state lognormal HMM.

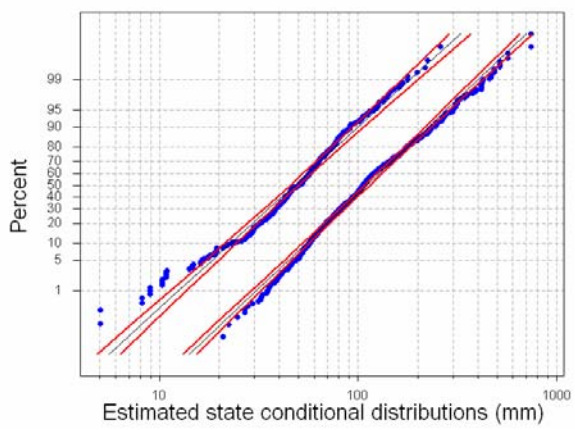

Figure 8.18 Lognormal probability plot showing estimates for conditional model states from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for District 66

The relationships between the persistent climate states identified in the Sydney monthly rainfall data and measures of broad-scale climate fluctuations previously described are now analysed. Using the 5-mrm values of the NINO3 index to define each month as being in El Niño, La Niña and ENSO neutral phases, Table 8.9 shows the numbers of months in which these regimes coincide with HMM states. A posterior median state probability of 0.5 is used as a threshold to categorise each month during the period of the Sydney record as being in a wet or dry state.

These results indicate a tendency for a lot of months in which the climate is most likely within a wet state to coincide with La Niña periods, which generally lead to increased rainfall in the area surrounding Sydney. The inclination for the HMM calibration to identify El Niño months as being in dry states is only modest however, with a majority of El Niño months having wet state probability greater than 0.5. This result may indicate that ENSO is not the sole global circulation phenomenon that influences rainfall variability in Sydney. This is further demonstrated through the rank correlation between the median state series from this calibration and the monthly NINO3 series. Although this correlation is significant $(p < 0.001)$, the NINO3 series only explains approximately 1.5% of the variability in the state series $(r = -0.121)$. Table 8.10 shows similar results from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for District 66. As with the Sydney data, the District 66 rainfall shows a bias towards wet states coinciding with La Niña periods. Also there is little evidence for dry states to occur during El Niños, reinforcing the observations in Table 8.9.

Table 8.9 Numbers of months in which most probable HMM states from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for Sydney coincide with ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	280	429	334
Dry state	224	293	156

Table 8.10 Numbers of months in which most probable HMM states from the calibration of a two-state NP HMM to the deseasonalised monthly rainfall for District 66 coincide with ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	225	378	206
Dry state	95	115	61

Two-state NP HMMs are now calibrated to the various deseasonalised monthly rainfall series from across Australia, with results summarised in Table 8.11. These results show evidence for significant two-state persistence in the monthly rainfall series of Adelaide, Alice Springs, Brisbane and Melbourne, together with the Sydney and District 66 data. These are important observations, as neither two-state gamma nor two-state lognormal HMMs showed persistence in the Brisbane or Melbourne data. With the monthly rainfall for Perth, the NP HMM was unable to reject the notion of two-state persistence degenerating to a mixture of two distributions.

Table 8.11 Medians of posteriors for HMM transition probabilities and their sum, with 90% credibility intervals from the calibration of two-state NP HMMs to deseasonalised monthly rainfall

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Adelaide	**0.232** (0.081, 0.457)	**0.323** (0.169, 0.548)	**0.578** (0.346, 0.819)
Alice Springs	**0.274** (0.122, 0.475)	**0.068** (0.027, 0.211)	**0.348** (0.159, 0.644)
Brisbane	**0.214** (0.147, 0.292)	**0.252** (0.158, 0.412)	**0.472** (0.348, 0.643)
Darwin	**0.388** (0.112, 0.796)	**0.558** (0.220, 0.865)	**0.954** (0.550, 1.334)
Melbourne	**0.328** (0.108, 0.539)	**0.373** (0.173, 0.581)	**0.727** (0.422, 0.911)
Perth	**0.402** (0.122, 0.741)	**0.551** (0.215, 0.869)	**0.976** (0.621, 1.222)

Table 8.11 shows that the posterior distribution of $P_{WD} + P_{DW}$ from the calibration of a two-state NP HMM to the Darwin data fails to provide significant evidence for two-state persistence.

Table 7.4 and Table 7.5 however demonstrated two-state persistence in the Darwin rainfall through assuming lognormal and gamma conditional distributions respectively. With the two-state NP HMM identifying monthly-scale persistence effectively, the presence of three-state persistence in these monthly rainfall observations is now investigated in the following section through the calibration of three-state NP HMMs. It is possible that the Darwin data is better described as three-state persistence.

8.3.3 Calibration of three-state NP HMMs to monthly rainfall data

Three-state NP HMMs are now calibrated to the deseasonalised monthly rainfall from Sydney and District 66. In Section 7.3, the possibility of a three-state HMM with lognormal conditional distributions degenerating to merely a mixture distribution when calibrated to either series could not be rejected at a 10% significance level.

In order to investigate whether the estimation of three-state persistence in these two monthly rainfall series is biased by the assumption of the form of conditional distributions, three-state NP HMMs are now calibrated to both series. The posterior distributions for the sums of self-transition probabilities from these calibrations each have medians values well in excess of unity (1.322 for Sydney and 1.346 for District 66), and more importantly the 90% credibility intervals around these estimates remain above a value of 1. As stated in Section 7.3, this condition is consistent with credible three-state persistence; hence the possibility of the NP HMM calibration degenerating to a mixture distribution is rejected at a 10% level. By removing the assumption of lognormal conditional distributions, evidence of three-state persistence in these data is more clearly identified.

After calibrating the three-state NP HMMs to these deseasonalised monthly rainfall series, 3000 random samples are taken around posterior medians of the partition locations to estimate the state conditional distributions. The estimated distributions for the Sydney rainfall series are shown on a lognormal probability plot in Figure 8.19. The three distributions plot as approximate straight lines, which suggest that these are consistent with random draws from lognormals, supported by low Anderson-Darling goodness-of-fit statistics (1.922 for the wet state, 0.683 for the neutral state and 1.282 for the dry state).

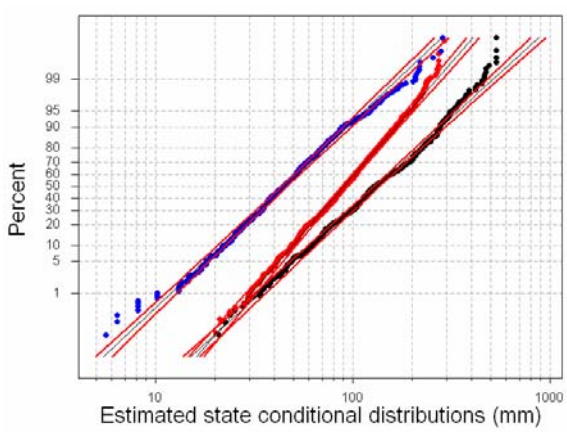

Figure 8.19 Estimates of state conditional distributions from the calibration of a three-state NP HMM to the deseasonalised monthly rainfall for Sydney

The estimated conditional distributions from the calibration of a three-state NP HMM to the District 66 series, again using 3000 random samples, are shown in the lognormal probability plot of Figure 8.20. These distributions approximate straight lines, with the neutral distribution consistent with random draws from a lognormal distribution (AD statistic 0.535). The slightly higher goodness-of-fit statistics for the wet state (2.301) and dry state (3.208) may lead to the assumptions of the three-state lognormal HMM being inadequate.

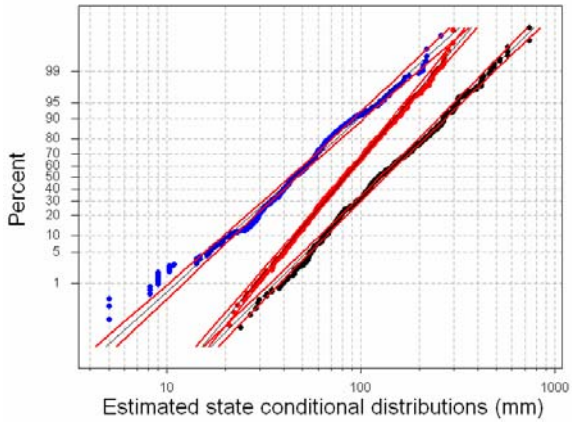

Figure 8.20 Estimates of state conditional distributions from the calibration of a three-state NP HMM to the deseasonalised monthly rainfall for District 66

In order to investigate the relationship between hidden climate states in the Sydney monthly rainfall and ENSO phases, Table 8.12 shows the numbers of months in which predominantly wet, neutral and dry states coincide with El Niño, La Niña and ENSO neutral regimes. These results demonstrate that a majority of months during La Niña periods are most likely to be identified within wet climate states. However Table 8.12 shows little tendency for months within El Niño periods to be identified as dry states, suggesting that the ENSO signature is not revealed clearly in the Sydney monthly rainfall through NP HMM calibration.

Table 8.12 Numbers of months in which most probable HMM states from the calibration of a three-state NP HMM to the deseasonalised monthly rainfall for Sydney coincide with ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	216	378	268
Neutral state	68	110	66
Dry state	220	288	156

Table 8.13 shows the results from the calibration of a three-state NP HMM to the deseasonalised monthly rainfall for District 66. An interesting aspect of these results is that a majority of all months in this series (55% of months) are identified as being most likely generated from a wet climate state. This underlies the fact that each of the three ENSO phases are most likely to coincide with wet hidden states as opposed to either neutral or dry climate states. Such results demonstrate that the significant three-state persistence identified in both the Sydney and District 66 data is not influenced solely by ENSO variability.

Table 8.13 Numbers of months in which most probable HMM states from the calibration of a three-state NP HMM to the deseasonalised monthly rainfall for District 66 coincide with ENSO phases

	El Niño	ENSO Neutral	La Niña
Wet state	168	260	162
Neutral state	68	127	50
Dry state	84	106	55

Model selection results show that two-state NP HMMs are superior to three-state NP HMMs for both the Sydney and District 66 data than with Bayes Factors of $\ln BF_{2NP,3NP} = 6.0$ and $\ln BF_{2NP,3NP} = 6.7$ respectively. Table 8.14 summarises the calibrations of three-state NP HMMs to the various deseasonalised monthly time series. These results indicate significant three-state persistence in each time series, with the exception of Melbourne and Perth. These latter series show credibility intervals around $P_{WW} + P_{NN} + P_{DD}$ that include unity. By observing Bayes Factors that compare the calibrations of two-state NP HMMs to the calibrations of three-state NP HMMs, it is clear that the simpler model is favoured for each series apart from Darwin. The failure of the two-state NP HMM to identify persistence in the Darwin time series suggests that its weak persistence is indeed better modelled by a three-state NP HMM.

Table 8.14 Medians of posterior distributions for the sums of self-transition probabilities, with 90% credibility intervals, from the calibrations of three-state NP HMMs to deseasonalised monthly rainfall series, with Bayes Factors comparing these calibrations to two-state NP HMMs

	$P_{WW} + P_{NN} + P_{DD}$	$\ln BF_{2NP,3NP}$
Adelaide	**1.616** (1.198, 2.017)	**4.6**
Alice Springs	**1.662** (1.200, 2.265)	**5.4**
Brisbane	**1.785** (1.312, 2.133)	**6.3**
Darwin	**1.391** (1.271, 1.530)	**-1.5**
Melbourne	**1.380** (0.973, 1.804)	**6.5**
Perth	**1.167** (0.752, 1.617)	**5.8**

Notwithstanding the results shown in Table 8.14, the three-state NP HMM is a suitable approach for the identification and modelling of hydroclimatic persistence. The benefits of both two-state and three-state NP HMMs are further examined in the following section, in which the persistence of monthly hydrologic data from across Australia is examined.

8.4 Calibration of NP HMMs to Australian hydrologic data

In this section, the non-parametric HMM approach is used to investigate the spatial extent of persistence in the hydrology of Australia. Two-state and three-state NP HMMs are calibrated to various spatially-averaged rainfall and monthly streamflow data to obtain an understanding of the role of hydrological persistence, and its relationship to broad-scale climate fluctuations identified through a range of indices.

8.4.1 Spatially-averaged monthly rainfall for New South Wales

The NP HMM is an unbiased method to elucidate hydrological persistence within rainfall observations from the Sydney region. The El Niño Southern Oscillation (ENSO) phenomenon strongly modulates hydrologic responses in this region; however it is pertinent to investigate whether similar persistence is identified throughout regions in which the influence of ENSO is weaker. Two-state NP HMMs are now calibrated to spatially-averaged rainfall across various meteorological districts of New South Wales. In order to determine whether persistence identified by the NP HMM is consistent with variability within rainfall observations, monthly data for districts along a transect from Sydney towards the northwestern corner of the state are used. This transect is shown in Figure 8.21 to pass through seven meteorological districts. The

two-state NP HMM is calibrated to the deseasonalised monthly rainfall for each of these seven districts, with monthly persistence in each district compared.

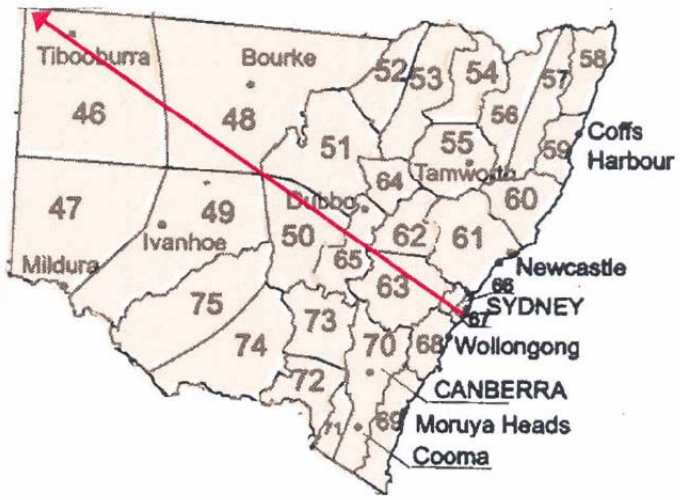

Figure 8.21 Meteorological districts in New South Wales, with arrow showing the transect from Sydney to northwestern corner of the state

Along this transect in a direction away from the coast, districts show a gradual reduction in mean monthly rainfall, as described in Table 8.15. This same relationship between rainfall in each district is also maintained when observing median monthly rainfalls. Following the calibration of two-state NP HMMs to the deseasonalised monthly rainfall in each of these districts, transition probability estimates are also summarised in Table 8.15.

Table 8.15 Mean and median monthly rainfall for various meteorological districts in New South Wales, together with the posterior median and 90% credibility intervals for the sum of transition probabilities from calibrating two-state NP HMMs to deseasonalised monthly rainfall

District	Mean monthly rainfall	Median monthly rainfall	$P_{WD} + P_{DW}$
66	96.0	69.1	**0.708** (0.525, 0.858)
67	69.5	50.0	**0.683** (0.500, 0.820)
63	74.6	63.2	**0.524** (0.326, 0.702)
65	50.1	42.7	**0.444** (0.281, 0.643)
50	39.8	31.0	**0.417** (0.279, 0.569)
48	29.6	20.6	**0.349** (0.231, 0.503)
46	19.2	10.4	**0.247** (0.142, 0.421)

The posterior medians for $P_{WD} + P_{DW}$ show a pattern that is consistent with mean monthly rainfalls, with regions of lower rainfall showing stronger hydrological persistence. Furthermore, in each of these districts, HMM calibrations reject the notion of persistence degenerating to mixtures of distributions.

The calibrations of two-state NP HMMs to these monthly data are utilised further to investigate the relationships between persistence and rainfall observations along this transect. Linear correlations between monthly rainfall in each of these districts and the monthly series from District 66 provide an indication of the spatial distribution of rainfall. These correlations are presented in Table 8.16, and show weaker correlations associated with districts located further from District 66. Alongside these values in Table 8.16 are linear correlations between the series of median state probabilities from NP HMM calibrations to each district and District 66. These results show that the NP HMM preserves underlying relationships in the rainfall regime of New South Wales, with monthly persistence closely linked to spatial rainfall variability.

Table 8.16 Linear correlations between the monthly rainfall in each district and the monthly rainfall in District 66, and between the median state series for each district and the median state series in District 66 from calibrating two-state NP HMMs to the deseasonalised monthly rainfall

District	Correlation of monthly rainfall to District 66	Correlation of median state series to District 66
67	0.847	0.858
63	0.670	0.680
65	0.403	0.468
50	0.385	0.449
48	0.307	0.355
46	0.179	0.222

8.4.2 Spatially-averaged monthly rainfall across Australia

Following the calibrations of two-state NP HMMs to deseasonalised monthly rainfall across New South Wales, this modelling approach can now be employed to ascertain monthly persistence across the whole country. A useful approach to determining Australia-wide characteristics of persistence is to use the spatially-averaged rainfall data for each of the 107 meteorological districts. The district-averaged rainfall data provides a spatially-consistent set of monthly rainfall, and each series (apart from two) has an identical length of 1080 months.

There are three main benefits from calibrating NP HMMs to the monthly rainfall for each district. Firstly, the magnitude of the sum of transition probabilities from the calibration of two-state NP HMMs to the monthly rainfall in each district provides a measure to highlight areas of strongest (and weakest) persistence, with Section 8.4.1 showing regions of lower rainfall to demonstrate stronger persistence. Secondly, credibility bounds around posterior estimates for

$P_{WD} + P_{DW}$ highlight districts in which the calibrations of two-state NP HMMs degenerate to merely mixture distributions. This quantity can then be used to evaluate the number (and location) of districts that fail to show evidence of statistically significant persistence. The third benefit of the NP HMM is that posterior state series are produced for each district. These series can be analysed to determine whether HMMs preserve spatial relationships between hydrological regimes across Australia. It is expected that for two-state (or three-state) hydrological persistence to reflect broad-scale climatic fluctuations, the hydrology of neighbouring meteorological districts would be similarly influenced. As a consequence, it is expected that the posterior state series from the calibrations of NP HMMs to monthly rainfall of neighbouring districts would be closely related. This was shown with the Sydney data and the spatially-averaged rainfall for the meteorological district in which Sydney is located, District 66.

Statistical tests described in Chapter 4 suggested significant two-state persistence in the spatial rainfall across Australia. Two-state NP HMMs are now calibrated to these spatial rainfall data, and Bayesian credibility limits for posterior estimates of $P_{WD} + P_{DW}$ show evidence for two-state persistence in the monthly rainfall data for 103 of the 107 districts. Of the remaining 4 districts that have 95th percentiles of this sum being greater than 1 (Districts 8, 10 and 10A in Western Australia and 99 in Tasmania), District 8 has the highest value at 1.084, which only marginally exceeds unity. These results demonstrate that two-state hydrological persistence is prevalent across Australia at a monthly scale.

The posterior medians for $P_{WD} + P_{DW}$ are now analysed to determine the varying extents of persistence across Australia. The 107 districts are initially ranked in increasing order of the magnitude of posterior medians for $P_{WD} + P_{DW}$, and then divided into four groups: $P_{WD} + P_{DW} < 0.4$, $0.4 < P_{WD} + P_{DW} < 0.5$, $0.5 < P_{WD} + P_{DW} < 0.6$ and $P_{WD} + P_{DW} > 0.6$. The 18 districts sharing the darkest shade in Figure 8.22 represent the “most persistent” group, with increasingly lighter shades representing comparatively weaker persistence.

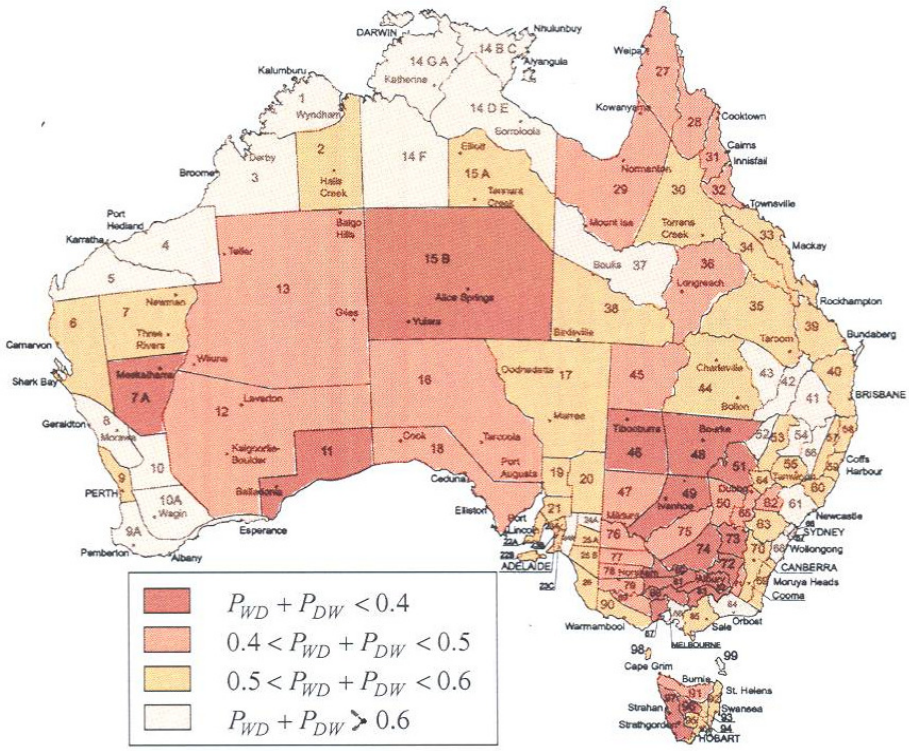

Figure 8.22 Spatial distribution of the magnitude of persistence from the calibration of two-state NP HMMs to the deseasonalised monthly rainfall for each district

A comparison of Figure 8.22 with Figure 5.1 shows that 14 of the 28 districts that failed to show persistence with each of the four runs tests described in Chapter 4 are also included in the least persistent group of districts (having the lightest shade) in Figure 8.22. Furthermore there is a slight tendency for the strongly persistent districts to correspond to the lower rainfall regions of Australia, with none of the districts in the most persistent group located along the generally wetter eastern coast of the country.

The consistency between persistence identified with the NP HMM approach and the results of spells analyses is now investigated. The series of 107 $P_{WD} + P_{DW}$ values from the calibrations of two-state NP HMMs are correlated to the series of 107 LORT probabilities, run skews, mean and maximum run lengths, and lag-1 autorun statistics in Table 8.17. Each correlation shown here is significant; indicating that NP HMMs identify similar persistence characteristics to those demonstrated with spells analyses. This is an important result, as it justifies the use of the non-parametric HMM approach for representing hydrological persistence. Furthermore the series of $P_{WD} + P_{DW}$ estimates fail to demonstrate significant correlation to the series of monthly Hurst exponents ($r = 0.182$, $p = 0.077$).

Table 8.17 Rank correlations (with p-values in brackets) between various runs statistics and posterior medians for the sums of transition probabilities from the calibration of two-state NP HMMs to deseasonalised monthly rainfall in each district

	LORT	Run skew	Mean runs	Lag-1 autorun	Max. runs
$P_{WD} + P_{DW}$	**0.542** (<0.001)	**0.614** (<0.001)	**0.437** (<0.001)	**0.470** (<0.001)	**0.410** (<0.001)

In order to demonstrate that the NP HMM identifies climatic features in the hydrology of Australia, spatial relationships between the monthly rainfall of each district should be conserved in the output of HMM calibrations. The spatial signature of persistence across Australia is analysed through a comparison of the median state series from each district to the median state series of District 66. For broad-scale climate influences to be revealed in HMM calibrations, districts surrounding 66 should demonstrate similar sequences of wet and dry months. Figure 8.23 shows rank correlations between the median state series for each district and the series from District 66. Alongside these correlations are the rank correlations between the untransformed monthly observations in each district with those from District 66, which reflect variability within the spatial rainfall data of Australia. Figure 8.23 shows a very close relationship across the 107 districts between correlations from observed monthly totals and correlations from posterior state probabilities. This indicates that two-state NP HMMs retain the inherent variability in the Australian climate. Furthermore regions that have state probabilities that are most strongly correlated to the state series of District 66 are districts within close proximity. This further supports the notion that the two-state NP HMM identifies underlying persistence within spatially-averaged rainfall data.

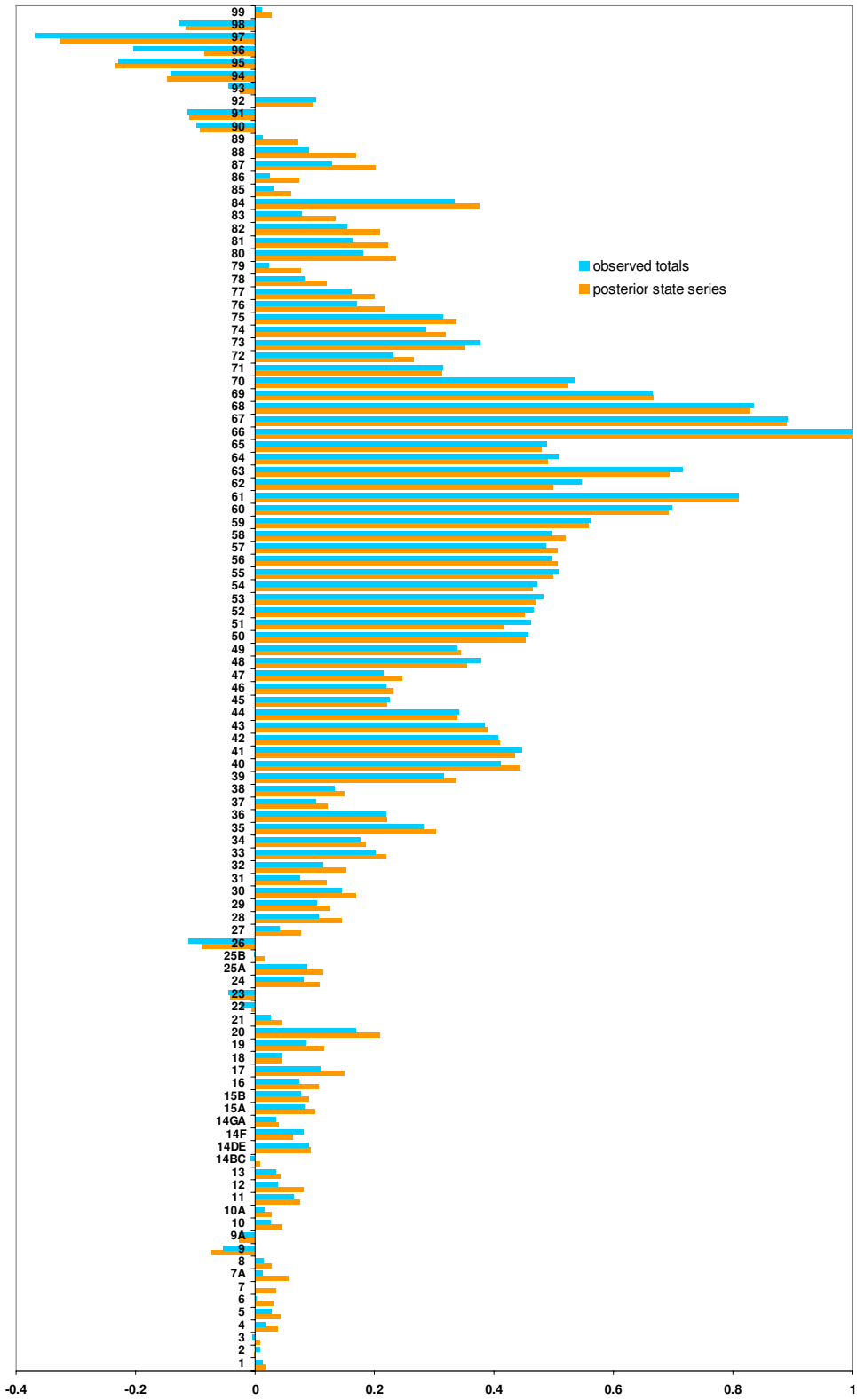

Figure 8.23 Rank correlations between monthly rainfall for each district and monthly rainfall of District 66 (blue), and rank correlations between the median state series for each district and the median state series for District 66 (orange) from the calibrations of two-state NP HMMs to the deseasonalised monthly rainfall for each district

Non-parametric HMMs demonstrate clear two-state persistence in the spatial rainfall data of Australia. The three-state NP HMM that was introduced in the previous section is now calibrated to the deseasonalised monthly rainfall for each district to analyse three-state persistence. Credibility bounds around the sums of self-transition probabilities from these calibrations indicate that 102 of 107 districts have significant three-state monthly persistence. Of the five remaining districts (Districts 6, 8, 9A, 10 and 10A all in Western Australia), three also fail to show significant two-state persistence.

The ranked order across the 107 districts of the posterior medians for sums of self-transition probabilities from calibrations of three-state NP HMMs is significantly correlated to the ranked order from calibrating two-state NP HMMs $(r = 0.82)$. This provides further evidence that the NP HMM identifies true modes of climatic persistence in these spatial rainfall series. Although the NP HMM approach provides an overwhelming justification for three-state persistence in monthly rainfall across Australia, Bayesian model selection suggests that these hydroclimatic features are better modelled as two-state persistence. In fact, 96 of the 107 districts have positive values of $\ln BF_{2NP,3NP}$, illustrating that the two-state NP HMM is superior to the three-state NP HMM (in a Bayesian context) for the various deseasonalised monthly rainfall series. The eleven districts better modelled by three-state NP HMMs are shaded in Figure 8.24.

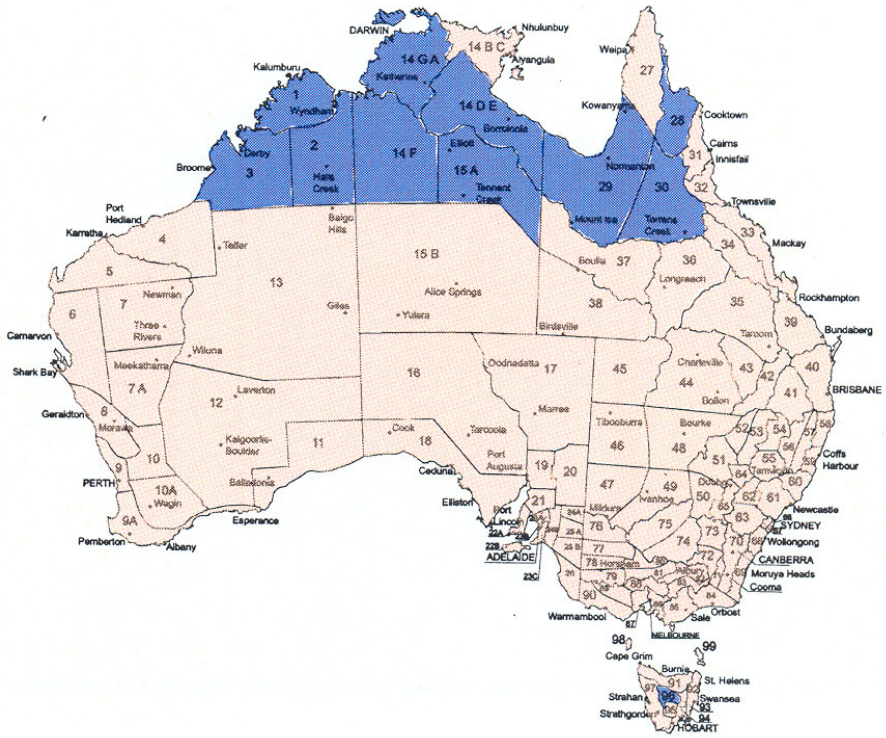

Figure 8.24 Districts in which three-state NP HMMs are superior models for monthly rainfall than two-state NP HMMs using Bayes Factors

Interestingly the meteorological districts in which monthly rainfall is better modelled through a three-state NP HMM are located in the tropical north of the country, except for District 96 in Tasmania, which also demonstrates the strongest three-state persistence of all districts. Once again, the spatial distribution of persistent districts shows apparent uniformity, further supporting previous findings that the NP HMM approach identifies underlying patterns of hydroclimatic persistence.

8.4.3 Monthly streamflows

The previous section showed that the unbiased NP HMM identifies widespread monthly persistence in the spatial rainfall data across most of Australia. Following these results, it is likely that streamflow data from across Australia will also demonstrate persistence through calibration of these models. Spells analyses described in Section 5.1 showed evidence for persistence in seven flow series that was stronger than the persistence shown in the various monthly rainfall series. Table 8.18 summarises the calibrations of two-state NP HMMs to the deseasonalised monthly rainfall series from each of the 107 districts.

Table 8.18 Comparison of posterior distributions for transition probabilities from the calibration of two-state NP HMMs to time series of deseasonalised monthly streamflows, with posterior medians and 90% credibility intervals shown

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Murray	**0.085** (0.070, 0.102)	**0.091** (0.077, 0.108)	**0.177** (0.148, 0.209)
Darling	**0.116** (0.099, 0.137)	**0.109** (0.093, 0.126)	**0.225** (0.193, 0.260)
Cooper	**0.244** (0.181, 0.308)	**0.115** (0.082, 0.151)	**0.359** (0.271, 0.449)
Diamantina	**0.268** (0.207, 0.330)	**0.086** (0.066, 0.109)	**0.354** (0.276, 0.436)
Burdekin	**0.153** (0.098, 0.212)	**0.102** (0.070, 0.141)	**0.254** (0.175, 0.344)
Dumaresq	**0.199** (0.142, 0.274)	**0.096** (0.058, 0.159)	**0.298** (0.210, 0.409)
Condamine	**0.182** (0.141, 0.225)	**0.236** (0.115, 0.297)	**0.416** (0.272, 0.506)

Table 8.18 shows strong two-state persistence in the monthly flows from these seven rivers, with each series characterised by low values for the posterior medians of $P_{WD} + P_{DW}$. The least persistent series is the Condamine River; however the median estimate of $P_{WD} + P_{DW}$ from this series is still lower than that observed in the monthly rainfall records of 80 meteorological districts, illustrating a tendency towards stronger persistence in streamflow data. The 90% credibility intervals for estimates of $P_{WD} + P_{DW}$ are all well below unity, indicating that

persistence in each river is strongly significant at a 10% level. The monthly flows from the Murray and the Darling show the strongest two-state persistence, which reflects spells analyses.

The calibrations of three-state NP HMMs showed significant three-state persistence in the monthly rainfall of many districts. To investigate this characteristic in monthly streamflows, three-state NP HMMs are now calibrated to the seven flow series, with results summarised in Table 8.19.

Table 8.19 Posterior distributions for the sums of self-transition probabilities from the calibration of three-state NP HMMs to time series of deseasonalised monthly streamflows, with medians and 90% credibility intervals shown, and Bayes Factors comparing the calibrations of two-state NP HMMs to the calibrations of three-state NP HMMs

	$P_{WW} + P_{NN} + P_{DD}$	$\ln BF_{2NP,3NP}$
Murray	**2.520** (2.463, 2.570)	**-175.4**
Darling	**2.447** (2.377, 2.511)	**-142.1**
Cooper	**1.767** (1.549, 2.119)	**3.7**
Diamantina	**1.883** (1.566, 2.197)	**4.0**
Burdekin	**2.499** (2.338, 2.631)	**-11.4**
Dumaresq	**2.225** (2.009, 2.429)	**-6.9**
Condamine	**1.812** (1.657, 1.972)	**-19.6**

The Bayesian credibility intervals shown in Table 8.19 are well above unity for each flow series, indicating significant three-state persistence in each case, which is consistent with results from spatial rainfall data. Whereas a majority of these latter series were better modelled by the simpler two-state NP HMM, Bayes Factors indicate that three-state NP HMMs are superior models for five of the seven monthly flow series. Monthly data from the Murray and the Darling, which showed the strongest two-state persistence, have the clearest preference for the three-state model. The two flow series better modelled by two-state NP HMMs are the two arid zone rivers; the Cooper and Diamantina, which also showed the weakest persistence from the spells analyses summarised in Chapter 4. Monthly persistence is a major feature within the different flow series analysed here, and is clearly more prevalent than within monthly rainfall observations. This outcome is consistent with the influence of catchment storage in rainfall-runoff transformations.

8.5 Investigating persistence in alternative hydrological variables

The investigations described in this chapter have demonstrated that hydrological persistence is most clearly observed at a monthly scale, primarily due to the frequencies of dominant climate processes, and also the increase in information available over annual records. The approach used up to this point has relied upon time series of total monthly rainfall or streamflow within which intra-annual seasonality is removed. In this section, other sources of hydrologic data that are associated with total monthly rainfall in Sydney are analysed, to provide evidence towards the most suitable scale of persistence. The benefit of the NP HMM to provide an unbiased estimate of persistence in a variety of hydrologic data is revealed through these analyses.

8.5.1 Analysing persistence within the number of monthly rain-days

In order to obtain a clearer depiction of monthly persistence in the hydrology of Sydney, alternative sources of hydrologic information are now considered. The number of days in each month during which rain is recorded provides a discrete-valued alternative to the continuous-valued total rainfall data, and is the focus of this section. These data sets describe a different aspect of the hydrological cycle; the number of rain-days interpreted as a representation of the number of rain events passing the observation point in Sydney and total monthly rainfall interpreted as the total amount of atmospheric moisture. Persistence in the broader climate is likely to influence each characteristic.

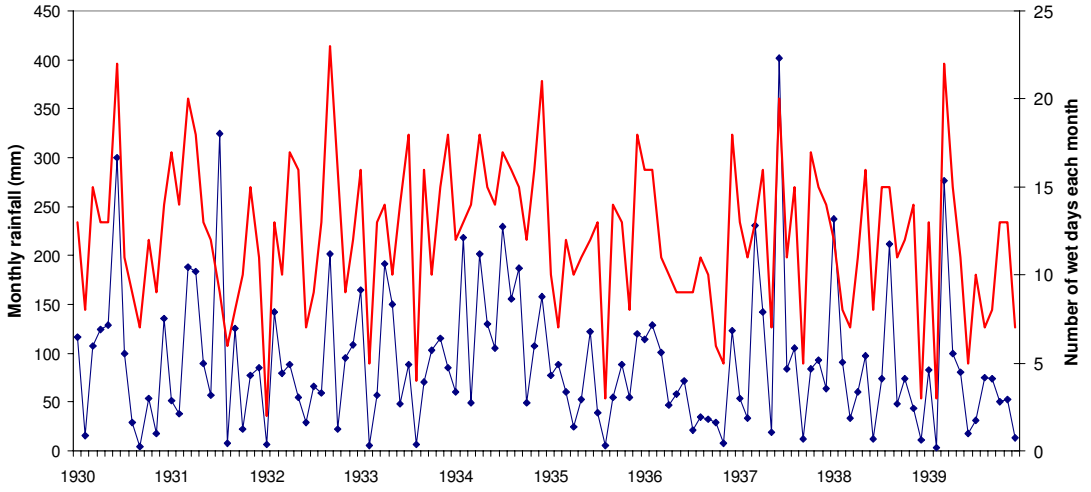

Figure 8.25 Time series of monthly rainfall observations for Sydney (blue line with points) and the number of days on which rainfall was recorded each month (red line) over a ten-year period

Figure 8.25 shows the time series of total monthly rainfall alongside the number of monthly rain-days over a ten-year period, demonstrating that the latter series identifies features similar to those shown with monthly rainfall totals. For the 1716 values over the 1859-2001 record, these two time series have significant linear correlation $(r = 0.61)$, which increases to $r = 0.68$ for

non-parametric rank correlation. Although these time series represent diverse aspects of the hydrologic cycle, their significant relationship indicates that common sources of climate variability affect each series.

The seasonal nonstationarity in the total monthly rainfall data was demonstrated in Figure 7.2, with a minimum monthly average rainfall in September, and maximum in March. This nonstationarity is compared with the intra-annual variability in the series of monthly rain-days in Figure 8.26 through the ratios of monthly averages to annual averages of each data series. The latter series shows similar nonstationarity, with a maximum monthly average again in March and minimum in August.

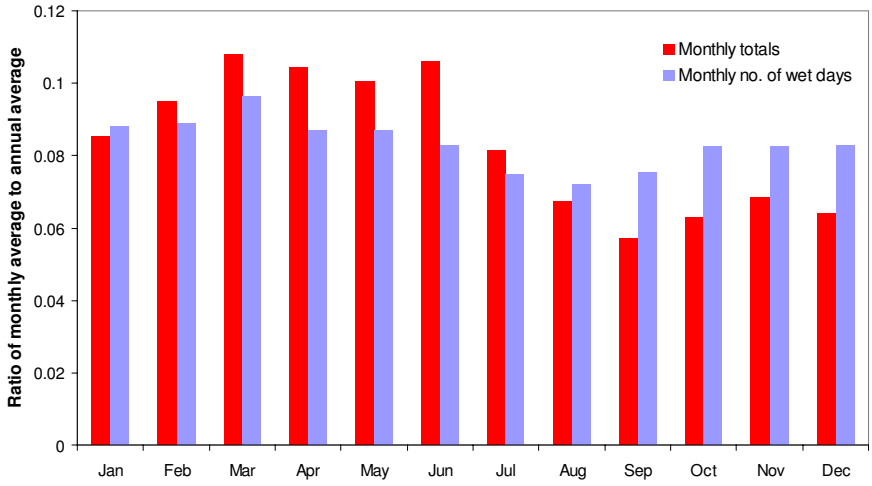

Figure 8.26 Seasonal variations in the ratio of monthly averages to annual averages for total monthly rainfall and number of rain-days in each calendar month for Sydney

As previously described, a benefit of the NP HMM is its ability to be calibrated to both continuous and discrete-valued time series. Consequently this model provides a suitable approach to investigating both two-state and three-state persistence in the series of monthly rain-days. However in order to maintain consistency with the analysis of persistence within the series of monthly rainfall totals, it is important to remove the seasonal nonstationarity in the discrete series prior to the calibration of NP HMMs. The discrete-valued time series of monthly rain-days is subsequently transformed to a series of deseasonalised monthly variates having zero monthly mean and unit standard deviation through subtracting the average number of rain-days in each month and dividing by the monthly standard deviation.

The calibration of two-state NP HMMs to the time series of deseasonalised number of rain-days each month identifies stronger persistence than that shown in the time series of deseasonalised monthly rainfall totals. Transition probabilities from the calibrations of NP HMMs to each series are summarised in Table 8.20, with posterior medians demonstrating two-state persistence within the former series that is approximately twice as long as in the latter.

Table 8.20 Comparison of posterior distributions for transition probabilities, showing medians and 90% credibility intervals, from the calibration of two-state NP HMMs to the deseasonalised monthly rainfall of Sydney and the deseasonalised number of monthly rain-days

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
Monthly rainfall totals	**0.301** (0.208, 0.406)	**0.420** (0.305, 0.539)	**0.728** (0.592, 0.841)
Number of rain-days per month	**0.231** (0.144, 0.343)	**0.121** (0.052, 0.234)	**0.361** (0.205, 0.547)

Figure 8.27 shows a further comparison of the calibration of NP HMMs to the two monthly series, with a time series plot of the two median state series over a ten-year period. The state series associated with the number of rain-days shows stronger persistence through reduced variability between consecutive months. Over the entire 143-year record, the two median state series are significantly correlated $(r = 0.62)$, providing further evidence for underlying climate persistence that influences not only monthly rainfall totals but also the number of rain events.

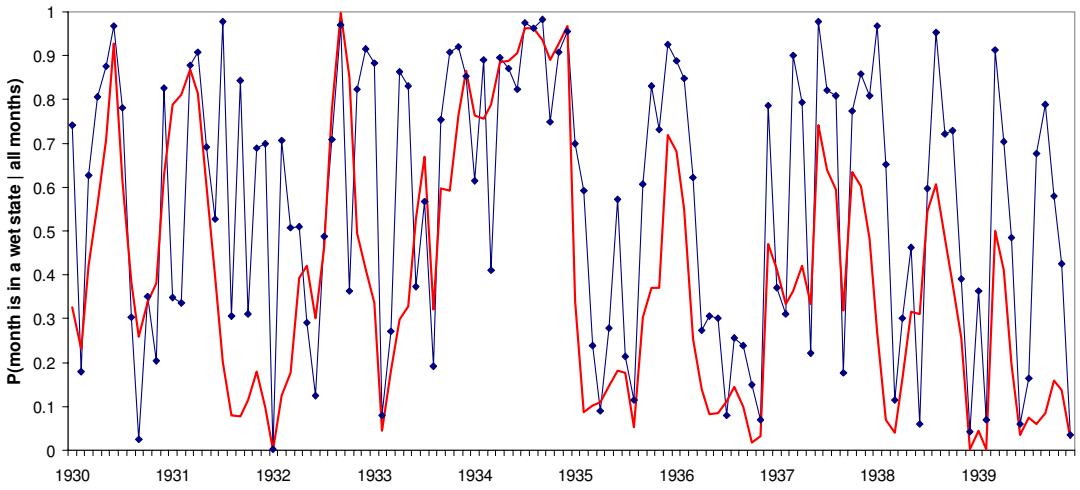

Figure 8.27 Median state series from the calibrations of two-state NP HMMs to the time series of deseasonalised monthly rainfall totals for Sydney (blue line with points) and deseasonalised variates of the number of monthly rain-days (red line)

The calibration of NP HMMs suggests that the time series of deseasonalised variates of the number of monthly rain-days more clearly identifies the influence of broad-scale climate fluctuations than the time series of monthly rainfall totals. Furthermore rank correlations between median state series and monthly NINO3 values show that the influence of ENSO is reflected almost twice as strongly $(r = -0.21)$ as the series of total monthly rainfall $(r = -0.13)$.

It is important to investigate whether the time series of monthly rain-days also demonstrates significant three-state persistence. The calibration of a three-state NP HMM produces a sum of self-transition probabilities that has a posterior median of 1.931 and a 90% credibility interval

(1.548, 2.313) that is well in excess of 1. Bayesian model selection identified the simpler two-state NP HMM as a superior model for total monthly rainfall through $\ln BF_{2NP.3NP} = 6.0$. Bayes Factors show a similar result for the number of monthly rain-days, with strong evidence for the simpler model presented through $\ln BF_{2NP,3NP} = 6.5$. This analysis of monthly rain-days reinforces the significance of monthly-scale persistence in the hydrology of Sydney. For other series these discrete analogues of total monthly rainfall may provide a more useful description of hydrologic variability.

8.5.2 Analysing monthly persistence in rainfall extremes

The previous section indicated that the time series of monthly rain-days for Sydney preserves the two-state hydrological persistence observed in the time series of total monthly rainfall. It was noted that data such as these were representative of the number of rain events, as opposed to the total amount of atmospheric moisture. An alternative source of hydrological data to now investigate for similar persistence is extreme rainfall events, which are analysed through rainfall intensities over various durations. These data provide important information in the design and management of water resources infrastructure, such that the accurate modelling of hydroclimatic interactions that underlie their variability is a vitally important concern. The unbiased NP HMM is now calibrated to these rainfall extremes in order to investigate persistence.

Using pluviograph records from the same Sydney gauge, it is possible to extract time series of maximum rainfall intensities for each calendar month over durations of 6 minutes, 1 hour and 24 hour. Continuous data are available for a period of 960 months (1921-2000), with three months during this period missing. Monthly mean values were then used for these months to infill the record. The statistics of these observed data are shown in Table 8.21.

Table 8.21 Sample statistics for maximum rainfall intensities for Sydney over selected durations

	Mean intensity (mm/hr)	Median intensity (mm/hr)	Standard deviation	Skew
6-minute maxima	37.92	30.2	28.85	2.013
1-hour maxima	13.63	10.295	11.82	2.663
24-hour maxima	1.77	1.28	1.64	2.232

These three time series of extreme rainfall are closely related, indicated by the linear correlations shown in Table 8.22 alongside correlations between each series and the total monthly rainfall over the period (1921-2000). Each of these correlations is significant, with $p < 0.001$ in each case. The two strongest correlations are firstly between 6-minute maxima and 1-hour maxima, and secondly between 24-hour maximum intensities and total monthly rainfall. The latter suggests that months having extreme day-long rain events are likely to produce large monthly totals, which is an expected result.

Table 8.22 Linear correlations between time series of maximum rainfall intensities for Sydney over selected durations and total monthly rainfall for the period (1921-2000)

	1-hour maxima	24-hour maxima	Monthly totals
6-minute maxima	0.810	0.569	0.560
1-hour maxima	–	0.728	0.641
24-hour maxima	–	–	0.849

The seasonal nonstationarity in the time series of total monthly rainfall and also the number of monthly rain-days for Sydney, as shown in Figure 8.26, is also apparent in the series of extreme rainfall intensities. Figure 8.28 shows the ratios of average monthly intensities to annual averages for the three time series alongside the ratios for the total monthly rainfall taken for the period (1921-2000). It is clear that the close relationship between the two shorter duration intensities and also between the 24-hour intensities and monthly totals that were highlighted previously are maintained within these ratios. These four time series each show maximum monthly averages in March.

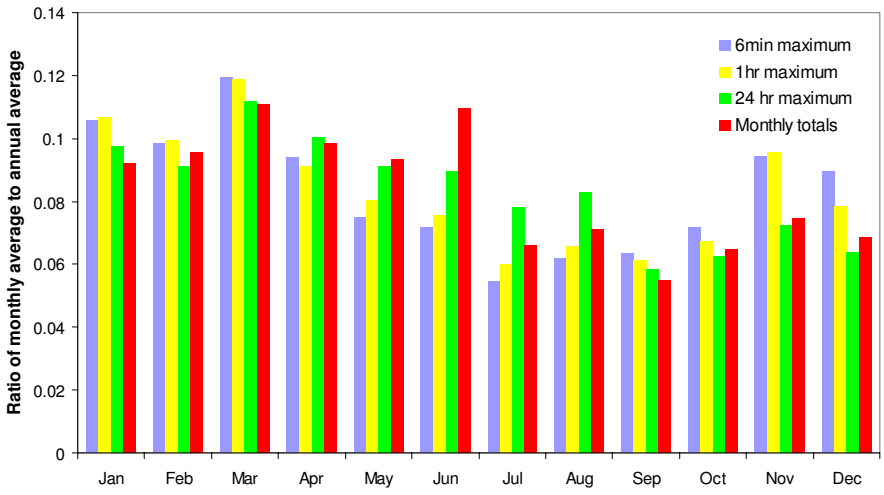

Figure 8.28 Seasonal variations in the ratio of monthly averages to annual averages for maximum rainfall intensities for Sydney recorded over various durations, and total monthly rainfall

To maintain consistency between the calibrations of NP HMMs to the extreme rainfall data and the monthly totals previously described, the seasonal nonstationarity in each series must first be removed through subtracting the average intensity in each month and dividing by the monthly standard deviation. Once again, this produces three series of zero mean and unit variance. The first stage of investigating hydrological persistence in these rainfall extremes is to calibrate two-state NP HMMs to each deseasonalised monthly series. The results of these calibrations are demonstrated through the following three figures, in which the median state series for each data set is compared to the median state series obtained from the calibration of a two-state NP HMM to the time series of total monthly rainfall for Sydney that is deseasonalised over the period

(1921-2000). This latter state series differs slightly from the series of median state probabilities obtained from calibrating NP HMMs to the time series of monthly totals deseasonalised over the longer period (1859-2001) as used previously in this thesis, although over the period (1921-2000) these state series are strongly correlated ($r = 0.99$). The median state probabilities from these calibrations are displayed in the following figures over identical 10-year periods.

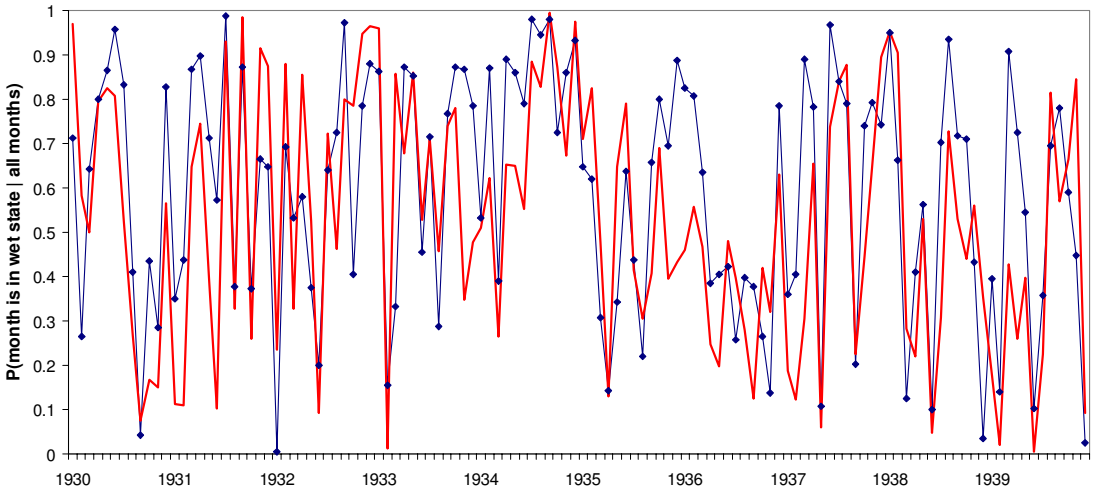

Figure 8.29 Median state series from the calibrations of two-state NP HMMs to the time series of deseasonalised monthly rainfall totals for Sydney (blue line with points) and deseasonalised maximum rainfall intensities for each month recorded over 6-minute durations (red line)

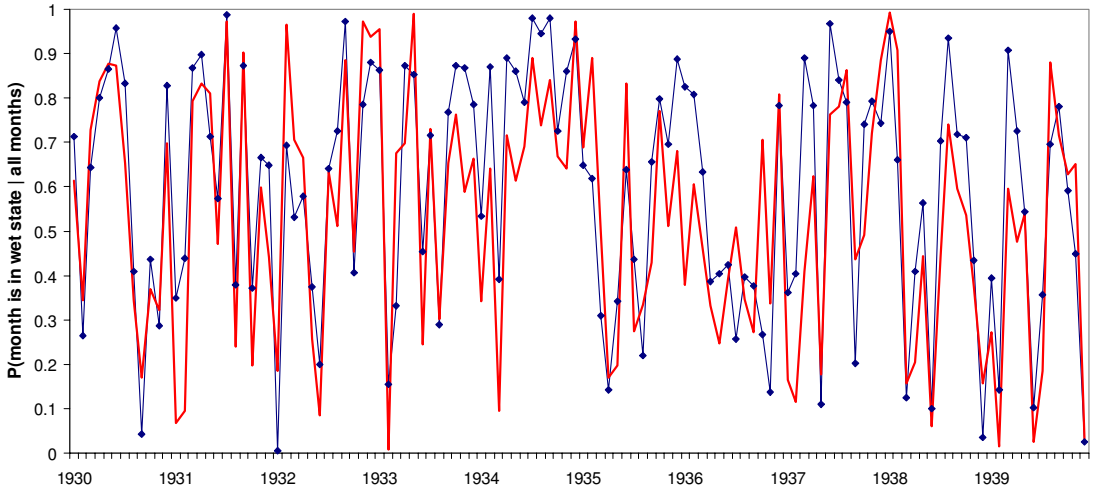

Figure 8.30 Median state series from the calibrations of two-state NP HMMs to the time series of deseasonalised monthly rainfall totals for Sydney (blue line with points) and deseasonalised maximum rainfall intensities for each month recorded over 1-hour durations (red line)

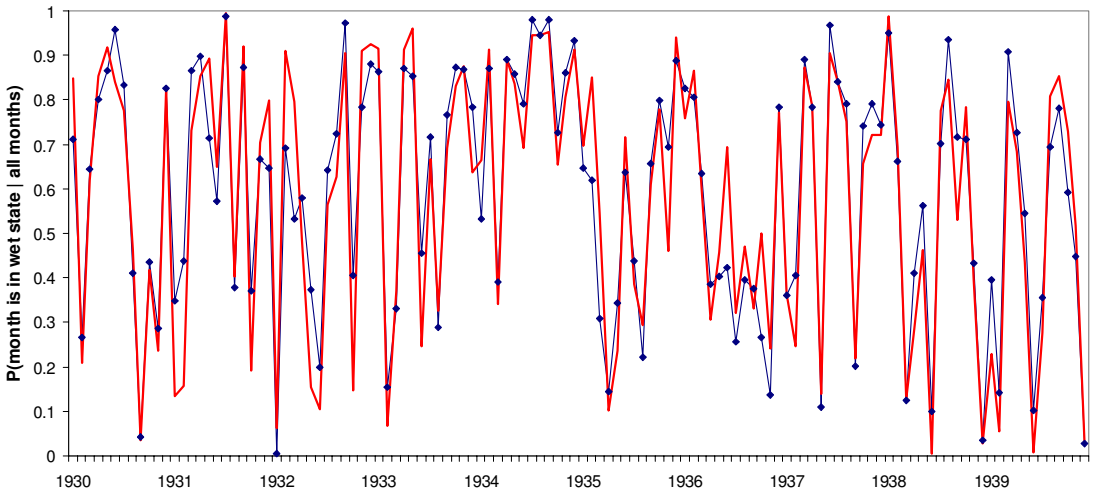

Figure 8.31 Median state series from the calibrations of two-state NP HMMs to the time series of deseasonalised monthly rainfall totals for Sydney (blue line with points) and deseasonalised maximum rainfall intensities for each month recorded over 24-hour durations (red line)

The preceding three figures demonstrate that the two-state persistence observed in time series of monthly rainfall totals is also clearly observed in the time series of maximum monthly intensities. In particular the median state series from the 24-hour maxima appears very similar to the median state series from deseasonalised monthly totals. The relationships between the median state series from these four data sets are also described through linear correlations shown in Table 8.23. These correlations indicate that through the calibrations of HMMs, relationships between these four observed series are retained.

Table 8.23 Linear correlations between median state series obtained from calibrating two-state NP HMMs to the time series of deseasonalised maximum monthly rainfall intensities for Sydney over selected durations and deseasonalised monthly rainfall totals for the period (1921-2000)

	1-hour maxima	24-hour maxima	Monthly totals
6-minute maxima	0.859	0.683	0.668
1-hour maxima	–	0.818	0.760
24-hour maxima	–	–	0.887

The magnitude of two-state persistence in the time series of maximum rainfall intensities is analysed through the transition probability values shown in Table 8.24. These results indicate that these three data series reveal two-state persistence that is similar to the persistence in the series of monthly totals, with transition probabilities having similar posterior medians. Although the two-state persistence in the time series of extremes is not significant at a 10% level (with 95th percentile of sums of transition probabilities slightly exceeding unity), these are significant at level slightly exceeding 10%. By comparing the transition probability estimates for the time series of monthly rainfall totals deseasonalised over the period (1921-2000) with estimates

obtained from calibrating two-state NP HMMs to the related time series that is deseasonalised over the period (1859-2001), it is clear that the shorter series has wider posteriors. This demonstrates the influence of data series length on the ability of the NP HMM to identify accurately underlying persistence. As a result, it is apparent that hydrological persistence observed in the time series of monthly totals is also evident in time series of monthly extreme rainfall intensities. Results such as these have not been previously reported in the literature.

Table 8.24 Posterior distributions for transition probabilities, showing medians and 90% credibility intervals, from the calibration of two-state NP HMMs to the time series of deseasonalised maximum monthly rainfall intensities for Sydney over selected durations and deseasonalised monthly rainfall totals for the period (1921-2000)

	P_{WD}	P_{DW}	$P_{WD} + P_{DW}$
6-minute maxima	**0.343** (0.073, 0.601)	**0.377** (0.137, 0.757)	**0.793** (0.257, 1.044)
1-hour maxima	**0.363** (0.103, 0.634)	**0.460** (0.180, 0.802)	**0.855** (0.516, 1.069)
24-hour maxima	**0.342** (0.132, 0.601)	**0.497** (0.257, 0.745)	**0.861** (0.596, 1.048)
Monthly totals	**0.300** (0.142, 0.519)	**0.463** (0.255, 0.677)	**0.783** (0.553, 0.967)

The next part of this investigation into persistence within the time series of maximum rainfall intensities is to calibrate three-state NP HMMs, and to compare these results to those obtained from calibrating two-state NP HMMs. Table 8.25 shows correlations between the series of most likely climate states from the calibrations of three-state NP HMMs to the four monthly time series. Although each of these correlations is statistically significant $(p < 0.001)$, all six correlations are weaker than those shown in Table 8.23. While such a result does not imply that the three-state models are inferior for these data, it does suggest that the relationships between observed data are not observed as strongly in the output of these models.

Table 8.25 Linear correlations between series of most likely states from calibrating three-state NP HMMs to the time series of deseasonalised maximum monthly rainfall intensities for Sydney over selected durations and deseasonalised monthly rainfall totals for the period (1921-2000)

	1-hour maxima	24-hour maxima	Monthly totals
6-minute maxima	0.762	0.567	0.562
1-hour maxima	–	0.665	0.634
24-hour maxima	–	–	0.799

In order to investigate whether persistence in the time series of monthly maximum rainfall intensities is better modelled with three climate states, transition probabilities from the calibrations of three-state NP HMMs are now analysed. Table 8.26 presents the sums of self-

transition probabilities together with Bayes Factors comparing the calibrations of three-state NP HMMs to two-state NP HMMs. The 90% credibility intervals for posteriors of $P_{WW} + P_{NN} + P_{DD}$ include a value of unity for each series, suggesting that there is no evidence at a 10% level for significant three-state persistence in these data series. This is similar to the results from calibrating two-state NP HMMs, which failed to support evidence for two-state persistence at 10% levels. The results for the monthly totals here show that with fewer data points the three-state NP HMM is unable to detect the significant three-state persistence that was evident in the longer time series of monthly totals (1859-2001). The Bayes Factors shown in Table 8.26 demonstrate that the persistence in each data set is better modelled with a two-state NP HMM, and Bayes Factors of similar magnitude provide further indication that hydrological persistence is evident in a similar fashion through these four time series.

Table 8.26 Posterior medians and 90% credibility intervals for the sums of self-transition probabilities from the calibrations of three-state NP HMMs, and Bayes Factors comparing the calibrations of two-state NP HMMs to the calibrations of three-state NP HMMs

	$P_{WW} + P_{NN} + P_{DD}$	$\ln BF_{2NP,3NP}$
6-minute maxima	**1.289** (0.881, 1.748)	6.0
1-hour maxima	**1.249** (0.804, 1.720)	5.9
24-hour maxima	**1.140** (0.801, 1.570)	5.7
Monthly totals	**1.146** (0.814, 1.649)	5.6

This section has provided further evidence for significant underlying persistence in the hydrological regime of Sydney. The investigations of persistence in hydrological variables such as the number of monthly rain-days and maximum monthly rainfall intensities are novel developments that have not been previously reported in the literature. Although persistence in the maximum rainfall intensity data was weaker than observed with monthly totals, it was apparent that by having to analyse fewer data of this type, the accuracy in identifying persistence was compromised. It is indeed possible that if pluviograph data was available for the same period as available for the record of monthly rainfall totals, persistence of similar magnitude may be observed.

With significant two-state persistence being observed within time series of monthly rain-days, it is clear that persistent climate states modulate the number of rain events that pass the observation point each month. This suggests that these persistent climate states may influence either the speed of upper-level winds that guide the movement of cloud formations or the rate at which evaporative processes produce atmospheric moisture. Furthermore, with significant two-

state persistence identified within time series of extreme short-duration rainfall events, it is clear that within each of these rain events, the level of moisture in each rain event is modulated. A wet climate state is therefore associated with both an increased number of rain cells passing the observation point, and also an increased level of moisture within each rain cell, which leads to an amplification of the intensity of short-duration rain bursts. It follows from these suppositions that the persistence observed in time series of monthly rainfall totals in an artefact of persistence in both the number of monthly rain-days and the intensity of extreme rain events each month.

8.6 Summary of chapter

This chapter has presented a thorough analysis of a novel non-parametric HMM for modelling persistence in hydrological records. Without relying upon assumptions that concern the parametric form of underlying state conditional distributions, this model provides unbiased estimation of persistence, such as transition probabilities and statistics of underlying conditional distributions. This model retains the parsimonious nature of conventional parametric HMMs, and it was shown to be easily extended to more than two model states.

One of the strengths of the non-parametric modelling approach is its ability to identify persistence in hydrologic records, and to provide unbiased estimates of the comparative strength of persistence across a range of data series. The advantage of this approach over conventional parametric HMMs is illustrated with its calibration to the monthly rainfall data from both Sydney and its associated meteorological district. The difficulty shown by parametric HMMs to identify significant two-state persistence in this latter series is in contrast with results shown in this chapter that demonstrate similar patterns of persistence within these closely related series.

Using the NP HMM, an Australia-wide analysis using spatially-averaged rainfall indicated widespread hydrological persistence within monthly data. Slight tendencies were shown towards stronger persistence in the low rainfall areas of the country, with significant correlations observed between the strength of two-state persistence and the results of various runs statistics. Importantly, the spatial signature of persistence identified with NP HMMs preserved the relationships between different rainfall series. Furthermore, persistence at an annual scale is not prevalent in the rainfall data of Australia, likely due to the insufficient data available from aggregating at this level. Monthly streamflow data demonstrated much stronger persistence than monthly rainfall data, reflecting the results of spells analyses described in Chapter 5.

The final investigations in this chapter focused upon persistence within alternative hydrological variables. The advantages provided by the NP HMM were demonstrated in the analysis of the discrete number of days each month upon which rainfall was recorded in Sydney, Australia.

This analysis provided evidence for underlying persistence at a monthly scale that was stronger than persistence observed in the time series of total monthly rainfall at this gauge. Furthermore, the NP HMM demonstrated that significant two-state persistence was also apparent in time series of rainfall extremes, an interesting result that was obtainable through calibrations with this model. These latter two sets of results indicate that persistent climate states modulate both the number of rain events that pass the observation point in question and the quantity of moisture contained within such rain events. The combination of these effects is then observed through persistent within time series of monthly rainfall totals.

The NP HMM is a valuable development of conventional parametric HMMs, allowing unbiased estimates of the strength of persistence within hydrologic data. This model can be calibrated to both continuous-valued and discrete-value data, and in Chapter 11, is shown to provide accurate simulations of monthly hydrologic data.