Appendix A - *Estimation of Hurst Coefficient*

A.1   Introduction

Montanari et al. [1997] provide three heuristic methods for the estimation of the Hurst coefficient, h. The purpose of these methods is only to detect long memory and provide a rough estimation of the h value. They will be applied to the monthly Sydney rainfall data to determine if there is any evidence of long-term memory. Monthly data is used because daily data has zero values which make the calculation of long-term memory useless [*Montanari et al., 1997*]. Annual data was not used because the reliable estimation of long-memory can only be performed when the sample size is large enough that the asymptotic properties hold [*Montanari et al., 2000*]. All the monthly values were deseaonalized to remove the effects of any annual periodic component. Only the results for the three methods will be presented. For full details of each method refer to *Montanari et al.* [1997,2000].

A.1.1 Rescaled range statistic

The Rescaled Range (R/S) statistic is the best known method for estimating h. The R/S statistic is calculated for 10 equally spaced values of the time step t and for 50 logarithmically spaced values of the lag, k. Figure A.1 shows a plot of logarithm of k against the logarithm of the R/S statistic. The line of best fit through these points should scatter along a line with slope h. The line best fit for this plot was estimated using only the central region, as points in extreme left and right are subject to biases [*Montanari et al.,* 1997]. Using this method an estimate of $h = 0.60$ was obtained.
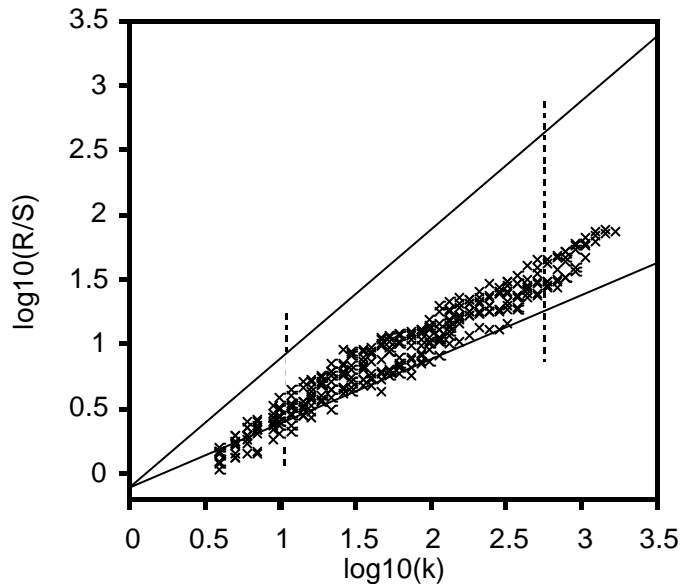

Figure A.1 – R/S Statistic for Sydney's deseasonalized monthly rainfall values. The two vertical dashed lines delineate the region in which the slope of the best straight line fit was estimated. The two solid lines represent the slopes corresponding to $h = 0.5$ and $h = 1.0$.

A.1.2 Aggregated variance method

Figure A.2 shows the results of applying the aggregated variance (AV) method. The logarithm of the sample variance when the data is divided into N/m blocks of size m was calculated and plotted against the logarithm of m. The line of best fit though these points should have a slope of $2h - 2$. 50 logarithmically spaced values of m were used. Again, only the central region of the plot was used. Using this method a value of $h = 0.52$ was obtained.
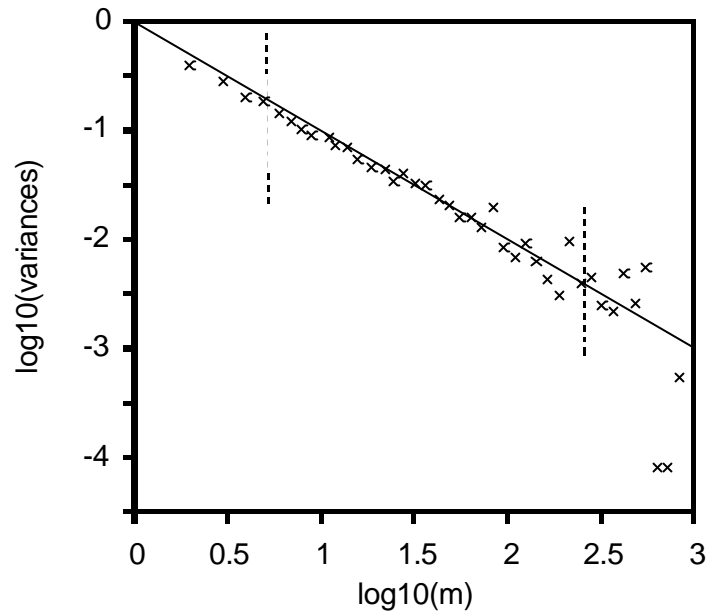
Figure A.2 – Aggregated variance plot for Sydney's deseaonalized monthly rainfall values. The two vertical dashed lines delineate the region in which the slope of the best straight line fit was estimated. The solid lines represent the slopes corresponding to $h = 0.5$ **.**

A.1.3   Differenced variance method

The two heuristic methods used above can give $h > 0.5$ for time series that do not have long term memory, but have shifts in the mean or a slowly decaying trend. When such nonstationarity is present the differenced variance (DV) method can be used to detect the long memory [*Montanari et al.,* 1997]. Figure A.3 shows the results of applying this method to the Sydney deseaonalized monthly rainfall data. The plot shows values of the difference in the variance for 50 logarithmically spaced values of m (note 50 points are not shown on the plot because those corresponding to a negative difference in the variance must be ignored). Again, only the central region of the plot was used to estimate the line of best fit. The slope of the line of best fit through these points should have a slope of $2h - 2$. Hence, an estimate of $h = 0.66$ was obtained.
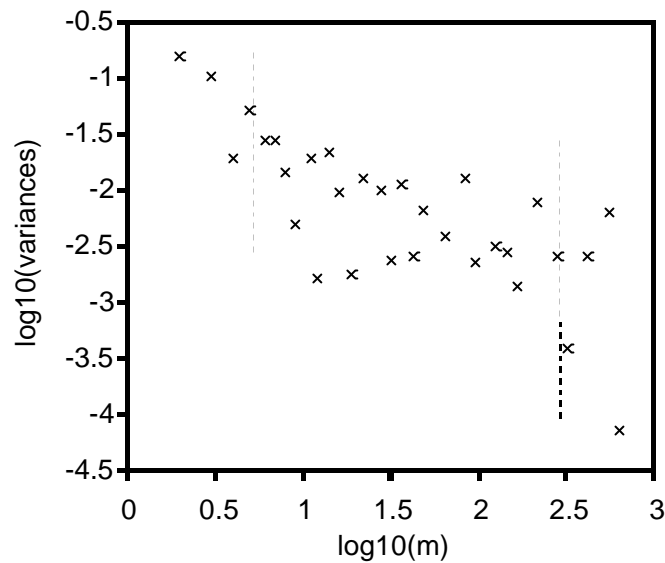
Figure A.3 – Differenced variance plot for Sydney's deseaonalized monthly rainfall values. The two vertical dashed lines delineate the region in which the slope of the best straight line fit was estimated.

A.2   Summary

A value for the Hurst coefficient h equal to 0.5 means the absence of long memory. The higher the h value is (up to a maximum value of 1.0) the higher is the intensity of long memory. These three heuristic methods give conflicting results for the estimation of h and hence the detection of long memory. Table A.1 shows that the R/S and the DV method both indicate the presence of long memory, although it is not very strong, whereas the AV method indicates the absence of long memory.

Table A.1 – Estimates obtained for the *Hurst* coefficient, h for three different heuristic methods

Heuristic Method	Estimated h value
Rescaled Range Statistic	0.60
Aggregated Variance Method	0.52
Differenced Variance Method	0.66

Montanari et al. [2000] showed that these three methods can give unreliable estimates of h when there is a strong periodicity in the time series. The autocorrelation function for the Sydney deseasonalized monthly rainfall values was calculated and no evidence of a strong periodic component was found.

It is difficult to draw any clear conclusions from these results, except to say that there is not strong evidence to suggest that there is long memory present in the Sydney monthly time series.

Appendix B - Filling-In and Correction of Hydrological Data

B.1 Filling-in of Missing Data

In all the monthly rainfall records from the selected sites in the Warrgamba catchment there were some months with missing data. This missing data was filled-in using the following methodology. Firstly, the daily rainfall record for a particular station was checked, because often the missing monthly value was due to only a missing day or two. If this was the case the daily record was filled-in using data from neighbouring stations, and the new monthly values were calculated. If the entire month was missing from the daily record then the fill-in value was calculated using the rainfall values from neighbouring stations and the correlations between these stations. If two or more neighbouring stations were available then the rainfall values from each were averaged to produce the fill-in value. In this case often this filled-in value was rounded to the nearest 5 or 10 mm. A higher degree of accuracy was not considered justifiable given the inherent uncertainties when in-filling using data from neighbouring stations. In addition, in this thesis these monthly values were aggregated to annual values. Hence any errors produced by this in-filling process would be reduced by the aggregation procedure. It was also considered that a 5 mm or 10 mm inaccuracy in an annual rainfall total which averages around 1000 mm is a relatively small error.

Table B.1 provides a summary of the neighbouring stations used for each of the selected sites and the monthly correlations. For Mt. Victoria, Blackheath PO, Moss Vale PO and Taralga PO, the monthly correlations with the majority of the neighbouring stations were quite high $\left(r^2 > 0.8 \right)$. For Yarra they were somewhat lower, but still considered reasonably good, with average $r^2 \approx 0.7$. Table B.2 provides a list of the months with missing data and the stations used to provide the fill-in values.

B.2   Correction of Inconsistencies in Monthly Data

Once the missing data was infilled, each of the monthly time series was examined to ensure the data was consistent with neighbouring stations. Scatter plots of the monthly values between the selected sites and the neighbouring stations were examined. If a monthly value appeared as an outlier in these scatter plots it was highlighted as potentially inconsistent data. Fortunately the data supplied by Sydney Water was quality coded, code 26 referred to good quality data, while code 80 referred to average quality data. If an outlier had quality code 80 this provided more evidence that the data maybe inconsistent. The daily records for the suspect months were also examined for further evidence. If strong evidence of an inconsistency was found the monthly value was corrected. Examples of strong evidence include when a two or three day rain event occurs at every neighbouring station but not at the selected station (e.g. Yarra, March 1976). For rainfall records taken at a post office the rainfall over a weekend would often be aggregated and entered as Monday's rainfall value. When the weekend coincided with a large rain event and a change of month this would produce an inconsistency in the two monthly values (e.g. Moss Vale PO, April/May 1995). Other instances include cases where the decimal place was clearly incorrectly placed e.g. 51 mm was entered instead of 5.1 mm (e.g. Moss Vale PO, January 1990). Similar to the technique used for the missing data the corrected monthly values were calculated using the rainfall information from neighbouring stations. Table B.2 provides a list of the monthly values that were corrected, the neigbouring stations used for the correction and brief description of the reasons why the correction was deemed necessary.

Table B.1 – Summary of the neighbouring sites and their monthly correlations for each of the selected sites from the Warragamba catchment.

Selected Site	Neighbouring Sites	Monthly Correlation (r^2)
Blackheath PO (063009)	Mt. Victoria (063056)	0.93
	Little Hartley (Sheepcombe) (063048)	0.80
	Leura (063045)	0.90
Mt. Victoria (063056)	Moss Vale PO (068045)	0.63
	Blackheath PO (063009)	0.93
Moss Vale PO (068045)	Goulburn (070037)	0.50
	Mittagong Pool (068044)	0.80
	Bowral PO (068005)	0.90
	Exeter (068025)	0.87
	Bundanoon PO (068008)	0.87
Yarra (Wollogorang) (070088)	Goulburn (Pomeroy) (070071)	0.75
	Forest Lodge (070033)	0.71
	Woodhouse Lee (Leeston) (070131)	0.71
	Chatsbury (Maryland) (070020)	0.64
	Goulburn (070037)	0.80
Taralga PO (070080)	Woodhouse Lee (Leeston) (070131)	0.83
	Chatsbury (Maryland) (070020)	0.80
	Golspie (Aryston) (063032)	0.77

Table B.2 – Monthly rainfall data filled-in and corrected for the selected sites from the Warragamba catchment region.

Gauge No	Location	Start Date	End Date	Missing Data		Corrected Data				
				Month	**Neighbouring Station(s)**	**Month**	**Reason**	**Neighbouring Station(s)**	**Old Value**	**New Value**
063009	Blackheath PO	1/1898	06/1994	1/1901-2/1901	063056	04/1940	up to 15th no rainfall recorded, while three others did	063056/063048/063045	28.9	100.0
						11/1940	Code 80 over weekend, divided	-	76.2	123.7
						12/1940	equally over two days	-	255	199.6
						12/1947	High rainfall in all other stations, little rainfall at 063009	063056/063048/063045	36	450.0
						04/1955	Moved daily rainfall from 30/4/1955 to 1/5/1955	-	179.6	248.2
						05/1955			146.6	78.0
				2	**Total**	**6**				
063056	Mt Victoria	1/1872	12/1989	7/1874-11/1874	068045					
				12/1987-12/1988	063009					
				17	**Total**					
068045	Moss Vale PO	08/1870	06/1994	12/1870	Daily Record	11/1905	Code 80 over end of month, divided	068008/068044/068005	7.9	97.9
				1/1871	70037	12/1905	use proportions from:	"	157	67.0
				3/1878	Daily Record	02/1931	Rainfall on 18th v. high compared to 4 other gauges in area, monthly total reduced	068044/068005/068025/068008	172.3	70.0
				4/1878-8/1878	070037/063056	04/1955	Code 80 over end of month, divided	068008/068044/068005/068025	55.3	185.3
						05/1955	use proportions from:		247.4	117.0
				02/1991	Daily Record	01/1990	Rainfall on 14th v. high compared to 068044/068008, reduced from 51mm to 5.1mm	068044/068008	184.1	138.1
				03/1991	68044					
				10	**Total**	**6**				
070080	Taralga PO	01/1882	07/1994	01/1884	Daily Record	12/1989	No rainfall recorded in month 3 other surrounding gauges recorded 50 mm min.	070020/070131/063032	0.0	65.0
				02/1884	07131					
				12/1892	Daily Record	06/1994	Only 9.01mm recorded, while 2 other srrounding gauges had over 50mm	070131/063032	9.01	50.0
				1/1893 - 12/1894	070020					
				07/1994	070080					
				28	**Total**	**2**				

Table B.2(cont) – Monthly rainfall data filled-in and corrected for the selected sites from the Warragamba catchment region.

Gauge No	Location	Start Date	End Date	Missing Data		Corrected Data				
				Month	Neighbouring Station(s)	Month	Reason	Neighbouring Station(s)	Old Value	New Value
070088	Yarra (Wollogorang)	9/1896	06/1994	10/1923	Daily Record	12/1947	Missed Events 4th - 31st	070071/ 070033/ 070131/ 070020/ 070037	25.96	200.0
				11/1923-12/1923	070071/ 070020/ 070037	01/1904	Values too low, compared to Goulburn	070071/ 070131/ 070020	50.62	180.0
				08/1947	Daily Record	05/1980	Daily Record Zero for period 4/1980 to 6/1980, other stations give values	"	0	100.0
				09/1947	070071/ 070033/ 070131/ 070020/ 070037	10/1981	Daily Record Zero for month, except 31st, other stations show rain	"	1.57	60.0
				06/1951	Daily Record	06/1980	Daily Record Zero for month, except 1st, other stations show rain	"	0.49	45.0
				7/1951-12/1951	070071/ 070033/ 070131/ 070020/ 070037	11/1979	Monthly value too low, compared to other stations	"	0	50.0
						04/1965	Rain recorded on 8th to 14th deemed too low	070071/ 070131/ 070020/ 070037	46.5	100.0
				12/1969	Daily Record	03/1976	No rain recorded on 1st to 4th , surrounding stations had rain	070071/ 070131/ 070020	14.61	60.0
				1/1970-12/1970	070071/ 070131/ 070020	06/1984	Monthly value too low, compared to other stations	070071/ 070131/ 070020	5.39	40.0
				11/1978-12/1978	"					
				10/1985-11/1985	"					
				07/1987	Daily Record					
				08/1987	070071/ 070131/ 070020					
				12/1987	Daily Record					
				12/1988	"					
				09/1989	"					
				10/1989-12/1989	070071/ 070131/ 070020					
				37	Total	6				

Appendix C - Markov Chain Monte Carlo Methods: An Overview

C.1 Introduction

Markov chain Monte Carlo (MCMC) methods provide a means to randomly sample from virtually any multivariate distribution. *Chib and Greenberg* [1995] give an excellent tutorial exposition of the theory behind MCMC methods. The explanation given here will take the reader through MCMC theory with particular care taken to explain the concepts behind the theory. The idea is to provide an intuitive explanation of how MCMC methods work. Detailed descriptions of the three commonly used members of the MCMC family, the Metropolis-Hastings (M-H) algorithm, the Metropolis algorithm and the Gibbs sampler will be given - the second two being special cases of the first. Because *Chib and Greenberg* [1995] provide possibly the best explanation for understanding MCMC methods the material in this appendix borrows heavily from their paper. As the emphasis is placed on explaining the concepts which underlay MCMC theory, on occasions the derivations given will not be theoretically rigourous. The presented theory focuses on continuous distributions. However, it equally applies to discrete distributions. The usual convention of denoting vectors as nonitalic bold type used in the rest of this thesis is not used this appendix.

C.2   Markov Chain Monte Carlo Simulation

The aim of MCMC simulation is to draw samples from a target distribution. All that is required is that the target density, denoted by $\partial(x)$, be calculable for a given vector x up to a known multiple, such that:

$$\pi(x) = f(x)/K \qquad\qquad (C.1)$$

where $f(x)$ is the unnormalized density and K is the (possibly unknown) normalizing constant. The approach used is to generate a Markov chain iterative sequence of samples $x^1, x^2, \ldots, x^n$. Now, due to the construction of the MCMC algorithm the distribution of these samples will eventually converge to a stationary distribution which corresponds to the target distribution.

The proof that this Markov chain sequence converges to the target distribution has two parts. The first part is to demonstrate that the sequence of samples is a Markov chain that converges to a stationary distribution. This holds if the Markov chain is irreducible, aperiodic and not transient. The latter two conditions hold for a random walk on any proper probability distribution (except for certain trivial cases) [*Gelman et al.,* 1995]. The term random walk is used to describe a sequence of samples where the next value is equal to the current value plus some noise. Irreducibility holds if the Markov chain has a positive probability of visiting every part of the target distribution from any other part of the target distribution. The second part of the proof is that the stationary distribution is the target distribution. This part is a little harder to prove.

In Markov chain theory the usual approach is to start with a transition kernel $P(A|x)$ which is the conditional probability function that represents the probability of moving from a point x to a point in the region defined by A. The notation "a region A" is used because the probability of moving to an exact point is zero. Also, it is permitted that the chain make a transition from x to x, that is, the probability of moving from x to x is not necessarily zero. Now, the major concern of Markov chain theory is to determine under what conditions a stationary distribution π^* exists and under what conditions the iterations based on this transition kernel $P(\cdot|\cdot)$ converge to the stationary distribution. The stationary distribution satisfies:

$$\pi^*(dy) = \int P(dy|x)\pi(x)dx \qquad\qquad (C.2)$$

where $\pi(\cdot)$ represents the probability density of the stationary distribution $\pi^*(\cdot)$. A subtle, but important point regarding the notation is explained here. The term dy is used because the probability of sampling a point y is zero, whereas the probability of sampling a region with "width" dy represented by $\pi(y)dy$ has some measure, hence $\pi^*(dy) = \pi(y)dy$. A conceptual interpretation of Equation (C.2) is that if the distribution is stationary then the probability of moving to a region dy from any point x is equal to the probability of sampling that region; that is, it is independent of the starting point x. The necessary conditions for the convergence of the iterations of the Markov chain to this stationary distribution were discussed earlier.

MCMC methods turn this theory around. The stationary density is already known (up to a constant multiple) − it is $\pi(\cdot)$, the target density from which samples are desired − but the transition kernel is unknown. Therefore to generate samples from $\pi(\cdot)$ MCMC methods find and utilize a transition kernel $P(dy|x)$ whose nth iterate converges to $\pi(\cdot)$ for large n. The process is started at an arbitrary x and iterated a large number of times. After this large number the distribution of samples generated by the transition kernel is approximately the target distribution.

Of course the problem then is to find a suitable transition kernel. MCMC theory simplifies this choice by supposing that for some function $p(y|x)$ the transition kernel can be expressed as:

$$P(dy|x) = p(y|x)dy + r(x)\delta_x(dy) \qquad\qquad (C.3)$$

where $p(x|x) = 0$, $\delta_x(dy) = 1$ if $x \in dy$ and 0 otherwise and

$$r(x) = 1 - \int p(y|x)dy \qquad\qquad (C.4)$$

which represents the probability that the chain remains at x. The $r(x)\delta_x(dy)$ part of Equation (C.3) takes care of the requirement that the probability of the chain not moving is nonzero, even if $p(x|x) = 0$.

Now, the reversibility condition is introduced, which is stated as:

$$\pi(x)p(y|x) = \pi(y)p(x|y).$$ (C.5)

This intuitively says that the unconditional probability of moving from x to y where x is sampled from $\pi(\cdot)$ [the left hand side of (C.5)] is equal to the unconditional probability of moving from y to x where y is sampled from $\boldsymbol{p}(\cdot)$ [the right hand side of (C.5)]. *Chib and Greenberg* [1995] show that if the reversibility condition holds then $\boldsymbol{p}(\cdot)$ is the stationary density of $P(\cdot|x)$. This proof is reiterated here, by first considering the right hand side of Equation (C.2):

$$
\begin{aligned}
\int P(A|x)\boldsymbol{p}(x)dx &= \int\left[\int_A p(y|x)\,dy\right]\boldsymbol{p}(x)dx + \int r(x)\boldsymbol{d}_x(A)\boldsymbol{p}(x)dx \\
&= \int_A\left[\int p(y|x)\boldsymbol{p}(x)dx\right]dy + \int_A r(x)\boldsymbol{p}(x)\,dx \\
&= \int_A\left[\int p(x|y)\boldsymbol{p}(y)dx\right]dy + \int_A r(x)\boldsymbol{p}(x)\,dx \\
&= \int_A[1-r(y)]\boldsymbol{p}(y)\,dy + \int_A r(x)\boldsymbol{p}(x)dx \\
&= \int_A \boldsymbol{p}(y)\,dy \\
&= \boldsymbol{p}^*(A)
\end{aligned}
$$ (C.6)

The first line of the above proof uses the definition given in (C.3). In the second line the variables of the integration are exchanged. In the third line the reversibility condition given in (C.5) is applied. In the fourth line the definition in (C.4) is applied. Equation (C.6) shows that if the reversibility condition holds then $\pi^*(\cdot)$ is the stationary distribution for $P(\cdot|\cdot)$. Hence the reversibility condition provides a sufficient condition that must the satisfied by $p(y|x)$. Now, it will be explained how the Metropolis-Hastings algorithm constructs a transition kernel that satisfies this condition.

C.3   The Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm is the most general implementation of the MCMC methods. It requires a candidate generating density from which the samples can be selected. As these candidates are to form a Markov chain, the density must be dependent on the current state of the process. Thus it is denoted as $q(y|x)$. This is interpreted as meaning that when the process is at a point x, the density samples a value y from $q(\cdot|x)$.

It is now required that this density $q(y|x)$ satisfies the reversibility condition given in (C.5), which it is not likely to do all by itself. Suppose that for some (x, y):

$$\pi(x)q(y|x) > \pi(y)q(x|y) \tag{C.7}$$

which says that the process is more likely to move from x to y then to move from y to x. This does not satisfy the reversibility condition. To correct this situation in the M-H algorithm a *probability of move* $\alpha(y|x) < 1$ is introduced to reduce the likelihood of the process moving from x to y. If this move is not made the process returns x as a value from the target distribution. Thus the transitions from x to y $(y \neq x)$ are made according to:

$$p_{\text{MH}}(\text{y}|\text{x}) \equiv q(y|x)\alpha(y|x), \qquad x \neq y \tag{C.8}$$

where $\alpha(y|x)$ is yet to be determined. To define the corresponding probability $\alpha(x|y)$ the inequality given in (C.7) is considered again. It is desirable that the likelihood of the process moving from y to x is made as large as possible. Thus $\alpha(x|y)$ is defined to be as large as possible, and since it is a probability, its upper limit is 1. To determine the probability of move $\alpha(y|x)$ we apply the reversibility condition to $p_{MH}(y|x)$, because then:

$$\begin{aligned} \pi(x)q(y|x)\alpha(y|x) \; &= \pi(y)q(x|y)\alpha(x|y) \\ &= \pi(y)q(x|y) \end{aligned} \tag{C.9}$$

and therefore it follows that $\alpha(y|x) = \pi(y)q(x|y)/\pi(x)q(y|x)$. If the inequality in (C.7) is reversed then $\alpha(y|x)$ is set to 1 and $\alpha(x|y)$ is derived similarly as above. The interpretation of the probabilities $\alpha(y|x)$ and $\alpha(x|y)$ is that they are introduced to "balance" both sides of the inequality given in (C.7) to ensure that $p_{MH}(y|x)$ satisfies the reversibility condition. Hence for $p_{MH}(y|x)$ to be reversible the formal definition of the *probability of move* is:

$$\alpha(y|x) = \begin{cases} min\left[\dfrac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right], & if \ \pi(x)q(y|x) > 0 \\ 1 & otherwise \end{cases} \qquad (C.10)$$

To provide a complete definition of the M-H process transition kernel, the possibility that the process remains at x must be considered. Using (C.4) and (C.8) from above we get the result:

$$r_{MH}(x) = 1 - \int q(y|x)\alpha(y|x)\ dy \qquad (C.11)$$

Hence the complete definition of the M-H transition kernel denoted by $P_{MH}(dy|x)$ is given by:

$$P_{MH}(dy|x) = q(y|x)\alpha(y|x)dy + \left[1 - \int q(y|x)\alpha(y|x)\ dy\right] \qquad (C.12)$$

As shown above by its construction $P_{MH}(dy|x)$ is reversible and hence it follows that the M-H kernal has $\pi(x)$ as its stationary density.

This completes the bulk of the theory that proves that the Markov chain sequence of samples generated by the M-H algorithm converges to the target distribution. The Metropolis-Hastings algorithm is potentially a very powerful algorithm because, theoretically at least, it can be used to sample from virtually any distribution. However there are several fundamental issues regarding the implementation of this algorithm which must be highlighted.

Firstly, the Metropolis-Hastings algorithm is specified by its candidate generating density, $q(y|x)$. The selection of an appropriate density is crucial to the efficiency of the algorithm. This issue will be discussed further in the following sections. Secondly, if

a candidate value is rejected then the current value is taken as the next item in the sequence. Thirdly, the calculation of $\alpha(y|x)$ *does not* require knowledge of the normalizing constant of $\pi(\cdot)$ because it appears in both the numerator and denominator of (C.10). This result means the application of the M-H algorithm for simulating Bayesian posteriors is particularly appealing because calculation of the posterior normalizing constant $p(y)$ [refer to Equation (5.2) in Section 5.2] is not needed. Finally, the implementation of the general M-H algorithm can be summarized in the following steps:

Step 1. Initialise x with arbitrary starting value x^0.

Step 2. Repeat for $i = 1, 2, ..., n$

Generate y from $q\left(\cdot \mid x^i\right)$ and u from $U(0,1)$

If $u \leq \alpha\left(y \mid x^i\right)$

set $x^{i+1} = y$

else

set $x^{i+1} = x^i$

Step 3. Return the values $\left\{x^1, x^2, ..., x^n\right\}$.

With any MCMC method the draws are only regarded as a sample from the target density $\boldsymbol{p}(x)$ only after the chain has passed the transient stage and the effect of starting value has become so small that it can be ignored. Now, we know the conditions that are required for a Markov chain to converge, these have been previously discussed. However, these conditions do not determine the rate of convergence. This leaves us with the empirical question of how large is the sample size, say b, which should be discarded and how long should the sampling run be. This issue of MCMC convergence is a crucial one. It has been and still is the subject of a large amount of research work. Section 5.2.4.c provides further discussion on this topic including an outline of the methods used to determine if convergence has been achieved.

As mentioned, the selection of an appropriate candidate generating density has a large influence on the performance of the M-H algorithm. Typically, this density is chosen

from a family of distributions that requires the specification of such tuning parameters as the location and the scale. *Chib and Greenberg* [1995] provide a good discussion of this subject with several examples. In this appendix the discussion will be centred on the families of candidate generating densities which lead to two special cases of the M-H algorithm, the Metropolis algorithm and the Gibbs sampler, as these are the MCMC methods which were applied in this thesis.

C.4   The Metropolis Algorithm

The Metropolis algorithm is a special case of the general Metropolis-Hastings algorithm where a symmetrical candidate generating density is chosen. Given that a symmetrical density satisfies $q(y|x) = q(x|y)$ the probability of a move reduces to $\alpha(y|x) = \min[\pi(y)/\pi(x),1]$. Therefore if $\pi(y) \geq \pi(x)$ the chain will always make the move from x to y; otherwise it moves with probability $\pi(y)/\pi(x)$. An intuitive interpretation of this result is that if a jump goes "uphill" then it is always accepted, whereas if a jump goes "downhill" it is accepted with nonzero probability.

In the implementation of the Metropolis algorithm used in this thesis the multivariate Gaussian density was chosen as suitable symmetrical candidate generating density. The algorithm was set up such that it was from the family of candidate generating densities that are characterized by the form $q(x|y) = q_1(y-x)$. Thus the candidate y is drawn according to the process $y = x + z$, where z is the incremental random variable that follows the distribution q_1 (which is multivariate Gaussian in this case). Because the candidate is equal to the current value plus some noise, this case is called a *random walk* chain. The advantage of this setup is that only the scale parameter of the candidate generating density is required to be tuned, the location parameter is given by the current value of the process *x*.

This still leaves the important question of choosing the scale, or spread, of the candidate generating density. This has important implications for the efficiency of the algorithm. The spread of the candidate generating density affects the behaviour of the chain in at least two ways: One is the "acceptance" rate (the percentage of times a move is made to a new point) and the other is the ability of the chain to explore all regions of the target distribution. To understand this, consider the situation where the chain has converged

and is located at the mode of the target distribution. If the spread is too large then the samples from the candidate generating density will be a long way from the current value, and hence will have a low probability of being accepted (because the target probability density for the samples a long way from the mode will have a very low value compared to the density at the mode). Conversely, if the spread is too small the chain will take a long time to explore the entire target density and it is likely that low probability regions will be undersampled. Again, *Chib and Greenberg* [1995] provide a good discussion on this topic, including some references which provide guidance for appropriate acceptance rates. For the application of the Metropolis algorithm in this thesis, refer to Section 8.3.1 for a description of how the spread of the candidate generating density was determined.

C.5   The Gibbs Sampler

If the M-H algorithm is set up such that it is applied in turn to subblocks of the vector x rather than simultaneously to all elements of the vector then some powerful algorithms, such as the Gibbs sampler, can be formed. These "block-at-a-time" algorithms have many advantages, the main one is that often it simplifies the search for a suitable candidate generating density.

The "block-at-a-time" concept is illustrated by considering the situation where the vector x is divided into two blocks $x = (x_1, x_2)$. Suppose there exists a conditional transition kernel $P_1(dy_1 | x_1, x_2)$ which for a fixed value of x_2 has $\pi_{1|2}^*(\cdot | x_2)$ as its stationary distribution [with corresponding density $\pi_{1|2}(\cdot | x_2)$]. Therefore applying Equation (C.2) to this subblock case we get:

$$\pi_{1|2}^*(dy_1 | x_2) = \int P_1(dy_1 | x_1, x_2)\pi_{1|2}(x_1 | x_2)\, dx_1 \qquad \text{(C.13)}$$

Also suppose the existence of a conditional transition kernel $P_2(dy_2 | x_2, x_1)$ which has $\pi_{2|1}^*(\cdot | x_1)$ as its stationary distribution, analogous to (C.13). For example P_1 could be the transition kernel generated by a M-H chain applied to block x_1 with x_2 fixed for all iterations. Now initially it may seem there are some serious drawbacks to this subblock approach. For example, one may think that each of these kernels would have to run to

convergence for every fixed value of the conditioning variable. Thankfully, this is not the case because the *product of kernels principle* saves the day. This principle says that the product of these two transition kernels has $\pi(x_1,x_2)$ as its stationary density. The practical significance of this result is enormous. Instead of having to run each kernel to convergence for each value of the conditioning variable it allows us to take draws in succession from each of the kernels. This has the added advantage because, as mentioned above, it is often easier to find several conditional kernels that converge to their respective conditional densities rather than to find one kernel that converges to the joint density.

To prove the product of kernels principle, it is first necessary to specify the order that the elements of x will be sampled. Suppose the transition kernel $P_1(\cdot\,|\,\cdot,x_2)$ samples y_1 given x_1 and x_2, and the transition kernel $P_2(\cdot\,|\,\cdot,y_1)$ samples y_2 given x_2 and y_1. This means that the vector $x=(x_1,x_2)$ is sampled in two steps. In the first step y_1 is sampled from $P_1(\cdot\,|\,\cdot,x_2)$ to replace x_1 and in the second step y_2 is sampled from $P_2(\cdot\,|\,\cdot,y_1)$ to replace x_2. The kernel formed by multiplying these two conditional kernels together has $\eth^*(x_1,x_2)$ as its stationary distribution, as shown below:

$$
\begin{aligned}
&\iint P_1\left(dy_1|\,x_1,x_2\right)P_2\left(dy_2|\,x_2,y_1\right)\eth\left(x_1,x_2\right)dx_1\ dx_2 \\
&= \int P_2\left(dy_2|\,x_2,y_1\right)\Big[\int P_1\left(dy_1|x_1,x_2\right)\eth_{1|2}\left(x_1|\,x_2\right)dx_1\ \Big]\eth_2(x_2)\ dx_2 \\
&= \int P_2\left(dy_2|\,x_2,y_1\right)\eth^*_{1|2}\left(dy_1|\,x_2\right)\eth_2(x_2)dx_2 \\
&= \int P_2\left(dy_2|\,x_2,y_1\right)\frac{\eth_{2|1}\left(x_2|\,y_1\right)\eth^*_1\left(dy_1\right)}{\eth_2(x_2)}\eth_2(x_2)\ dx_2 \\
&= \eth^*_1\left(dy_1\right)\int P_2\left(dy_2|\,x_2,y_1\right)\eth_{2|1}\left(x_2|\,y_1\right)dx_2 \\
&= \eth^*_1\left(dy_1\right)\eth^*_{2|1}\left(dy_2|\,y_1\right) \\
&= \eth^*\left(dy_1,dy_2\right)
\end{aligned}
\tag{C.14}
$$

The first line follows from our assumption about the order that the elements of the vector x will be sampled. The second line is simply a rearrangement. The third line follows from (C.13). In the fourth line Bayes theorem is applied. The fifth line is another rearrangement. The sixth line follows from applying (C.13) to P_2, while the last is a result of the law of total probability.

The proof of the *product of kernels principle* can also be generalized to cases where the vector x is divided into more than two blocks. Consider the case where three blocks are used, $x = (x_1, x_2, x_3)$. Extending the two block example given above, x is now sampled in three steps. For the first step, suppose the transition kernel $P_1(\cdot \mid \cdot, x_2, x_3)$ samples y_1 given x_1, x_2 and x_3 (i.e. y_1 replaces x_1). In the second step the transition kernel $P_2(\cdot \mid \cdot, y_1, x_3)$ samples y_2 given x_2, x_3 and y_1, (i.e. y_2 replaces x_2). Finally for the third step the transition kernel $P_3(\cdot \mid \cdot, y_1, y_2)$ samples y_3 given x_3, y_1 and y_2, (i.e. y_3 replaces x_3). The kernel formed by multiplying these three conditional kernels together has $\eth^*(x_1, x_2, x_3)$ as its stationary distribution, as shown below:

$$
\begin{aligned}
&\iiint P_1(dy_1 \mid x_1, x_2, x_3) \, P_2(dy_2 \mid x_2, x_3, y_1) P_3(dy_3 \mid x_3, y_1, y_2) \eth(x_1, x_2, x_3) dx_1 \, dx_2 \, dx_3 \\
&= \int P_3(dy_3 \mid x_3, y_1, y_2) \left[\iint P_1(dy_1 \mid x_1, x_2, x_3) \, P_2(dy_2 \mid x_2, y_1, x_3) \eth(x_1, x_2 \mid x_3) dx_1 \, dx_2 \right] \eth_3(x_3) dx_3 \\
&= \int P_3(dy_3 \mid x_3, y_1, y_2) \eth^*_{1,2|3}(dy_1, dy_2 \mid x_3) \eth_3(x_3) dx_3 \\
&= \int P_3(dy_3 \mid x_3, y_1, y_2) \frac{\eth_{3|1,2}(x_3 \mid y_1, y_2) \eth^*_{1,2}(dy_1, dy_2)}{\eth_3(x_3)} \eth_3(x_3) dx_3 \\
&= \eth^*_{1,2}(dy_1, dy_2) \int P_3(dy_3 \mid x_3, y_1, y_2) \eth_{3|1,2}(x_3 \mid y_1, y_2) \, dx_3 \\
&= \eth^*_{1,2}(dy_1, dy_2) \eth^*_{1,2|3}(dy_3 \mid y_1, y_2) \\
&= \eth^*(dy_1, dy_2, dy_3)
\end{aligned}
\tag{C.15}
$$

This proof has a very similar format to the proof given in (C.14), except that the result from (C.14) is used in the third line. Furthermore this result follows on as x is divided into additional subblocks. The transition kernel formed by multiplying all the subblock transition kernels together has $\pi^*(dy)$ as its stationary distribution.

Using this result an important special case of the M-H algorithm, the Gibbs sampler can now be derived. This algorithm is obtained by letting the transition kernels (for the two block case), $P_1(dy_1 | x_1, x_2) = \pi^*_{1|2}(dy_1 | x_2)$ and $P_2(dy_2 | x_2, y_1) = \pi^*_{2|1}(dy_2 | y_1)$; that is, the samples are generated from their "full conditional distributions". Note this method requires that it is possible to generate independent samples from each of the full conditional densities.

The M-H probability of move $\alpha(y | x)$ for this Gibbs sampling setup will now be calculated. Considering the two block case given above. Because the transition kernel is set to $P_1(dy_1 | x_1, x_2) = \eth^*_{1|2}(dy_1 | x_2)$ then the candidate generating density for the first step $q_{1|2}(y_1 | x_1, x_2) = \eth_{1|2}(y_1 | x_2)$. Therefore, using Equation (C.10), the corresponding probability of move for this first step $\acute{a}_{1|2}(y_1 | x_1, x_2)$ can be determined:

$$
\begin{aligned}
\acute{a}_{1|2}(y_1 | x_1, x_2) &= \frac{\eth_{1|2}(y_1 | x_2) q_{1|2}(x_1 | y_1, x_2)}{\eth_{1|2}(x_1 | x_2) q_{1|2}(y_1 | x_1, x_2)} \\
&= \frac{\eth_{1|2}(y_1 | x_2) \eth_{1|2}(x_1 | x_2)}{\eth_{1|2}(x_1 | x_2) \eth_{1|2}(y_1 | x_2)} \\
&= 1
\end{aligned}
\tag{C.16}
$$

It also follows from this that the probability of move for the second step $\acute{a}_{2|1}(y_2 | x_2, y_1) = 1$. As the probability of move for both the first and second steps is one then for the Gibbs sampler the overall probability of move, $\alpha(y | x) = 1$. This means that the candidate samples are always accepted. This result also applies to the multi-block case. Furthermore it is important to note that because the samples are always accepted the Gibbs sampler is the most efficient implementation of the M-H algorithm.

The Gibbs sampler is very useful for drawing samples from Bayesian posteriors. Consider the case where a parameter vector θ with posterior density $p(\theta | Y_N)$ is split

into d subblocks. As given above, in the Gibbs sampler the candidate generating density for each j th subblock is set to the corresponding subblock target density. In a general sense, this says $q_j\left(y_j\mid x_j,\ldots x_d,y_1,\ldots,y_{j-1}\right)=\eth_j\left(y_j\mid x_{j+1},\ldots x_d,y_1,\ldots,y_{j-1}\right)$. Therefore, each subblock of the parameter vector θ_j can be sampled using (x^i refers to the i th sample of parameter subblock x):

$$\grave{e}^i_j \;\leftarrow\; \mathrm{p}\!\left(\grave{e}_j \;\middle|\; \grave{e}^i_1,\ldots,\grave{e}^i_{j-1},\grave{e}^{i-1}_{j+1},\ldots,\grave{e}^{i-1}_d, Y_N\right) \tag{C.17}$$

This is very useful for situations where drawing samples of the full parameter vector directly from the posterior is not possible. Often it is found that it is a simple procedure to draw samples from the posteriors of a subblock of the parameter vector conditioned on the remaining fixed parameter values (the HSM model is good example). Therefore, the Gibbs sampler is an extremely important special case of the M-H algorithm.

C.6 Summary

Phew! This explanation of Markov chain Monte Carlo methods requires a large amount of statistical theory, which to the practicing stochastic hydrologist may seem a bit daunting. It is important to remember that the statistical theory is often full of details which sometimes can detract from the main conceptual ideas. Hence, the important ideas will be recapitulated to conclude this appendix:

- The most general implementation of Markov chain Monte Carlo methods is the Metropolis-Hastings algorithm which provides a technique to sample from virtually any multivariate distribution. All that is required is that the "target" distribution be calculable up to a known multiple for a given vector x.

- To do this it utilizies a candidate generating density to generate samples. As these samples are made to be dependent on the current value of process, they form a Markov chain. Markov chains possess a transition kernel, which represents the probability of moving from say x to y.

- Given certain mild conditions, this Markov chain will converge to a stationary distribution. To ensure that this stationary distribution is the target distribution the

samples from the candidate generating density are only accepted with a certain probability, denoted as the *probability of move*.

- This *probability of move* modifies the transition kernel to ensure that the "reversibility" condition is satisfied.

- The reversibility condition says that given x is generated from the target distribution then the unconditional probability of moving from x to y is equal to the unconditional probability of moving from y to x, if y is also generated from the target distribution.

- Hence as the probability of move modifies the transition kernel such that the reversibility condition is satisfied, then the samples generated by the Markov chain induced by the Metropolis-Hastings algorithm will converge to the target distribution.

- The major stumbling blocks for the implementation of the M-H algorithm are the specification of a suitable candidate generating density, which has a great influence on the efficiency of the algorithm, and determining when the Markov chain has achieved convergence, which is not always a straightforward task.

- The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm where a symmetrical candidate generating density is chosen. This simplifies the calculation of the *probability of move*.

- "Block-at-a-time" algorithms refer to the implementation of the Metropolis-Hastings algorithm where the elements of x are sampled one subblock at a time, rather than the entire vector at once. This often simplifies the choice of suitable candidate generating densities. The *product of kernels principle* means that each individual kernel used to produce samples for each subblock does not have to run to convergence, rather samples from each of the kernels can be drawn in succession and the entire algorithm can be then assessed for convergence.

- For a special case of the M-H algorithm called the Gibbs sampler the actual conditional target distributions of each subblock are chosen as the candidate generating densities. "Conditional" target distributions refer to the target distribution

of one subblock of the vector x, conditioned on the remaining elements of x. This results in the probability of move always equaling one, i.e. the samples are always accepted. This is the most efficient M-H implementation.

Appendix D - Sampling Distributions for the Gibbs Sampler

D.1 Introduction

The procedures for sampling from each of the conditional posteriors required for the application of the Gibbs sampler to the single site and multi-site HSM model will be given. *Gelman et al.* [1995] provides an excellent reference text of the development of the Bayesian posteriors, and unless otherwise referenced, the majority of results derived here are from their book. As well as providing the details of the derivation of the sampling procedures a conceptual interpretation of the resulting expressions for the posteriors is also given. The aim is to give an appreciation of the respective roles played by the prior and the data in influencing the Bayesian posteriors. Also included is the derivation of the likelihood function for the HSM model.

D.2 Sampling the Hidden State Time Series

The general method for sampling the hidden state time series S_N for multi-state Markov mixture models as presented by *Chib* [1996] is terse. A fuller treatment for the two-state

case of the HSM model is presented here. For the following derivation it is convenient to adopt the notation, as used by *Chib* [1996], where:

$$S_N = \{s_1,....,s_n\} \qquad S_t = \{s_1,....,s_t\} \qquad S^{t+1} = \{s_{t+1},....,s_n\}$$

with a similar convention adopted for the observed data Y_N, Y_t, and Y^{t+1}.

The entire state time series is simulated using the distribution $p(S_N|Y_N,è)$ which is the joint posterior mass function of all the states given Y_N and $è$. The derivation aims to develop a simple expression for this joint distribution exploiting the Markovian property $p(s_t|S_{t-1},Y_{t-1}) = p(s_t|s_{t-1})$. This will lead to a recursive simulation procedure where at each step, starting with the terminal state, s_n, only a single state has to be drawn.

Step 1: By initially rewriting the joint distribution of the states, $p(S_N|Y_N,è)$ and applying the conditional probability theorem repeatedly to the right hand term a recursive expression results:

$$p(S_N|Y_N,è) = p(s_1,\{s_2,....,s_n\}|Y_N,è)$$
$$= p(s_1|\{s_2,....,s_n\},Y_N,è)\, p(\{s_2,....,s_n\}|Y_N,è)$$
$$= p(s_1|\{s_2,....,s_n\},Y_N,è)\, p(s_2|\{s_3,....,s_n\},Y_N,è)\, p(\{s_3,....,s_n\}|Y_N,è)$$

The summary of this recursion is:

$$p(S_N|Y_N,è) = p(s_1|S^2,Y_N,è)... p(s_t|S^{t+1},Y_N,è)... p(s_n|Y_N,è) \qquad (D.1)$$

The typical term, excluding the terminal point is therefore: $p(s_t|S^{t+1},Y_N,è)$

Step 2: Expand and split the Y_N term as $\{Y_t,Y^{t+1}\}$ from the typical term in (D.1), and apply Bayes theorem to the result. Further expand and split the S^{t+1} term as $\{s_{t+1},S^{t+2}\}$ and apply the conditional probability theorem:

$$p(s_t|Y_N,S^{t+1},è) = p(s_t|Y_t,Y^{t+1},S^{t+1},è)$$
$$= p(Y^{t+1},S^{t+1}|s_t,Y_t,è)\, p(s_t|Y_t,è)\big/ p(Y^{t+1},S^{t+1},Y_t,è)$$

$$\propto p\left(s_{t+1}, S^{t+2}, Y^{t+1} \mid s_t, Y_t, è\right) p(s_t \mid Y_t, è)$$

$$\propto p\left(S^{t+2}, Y^{t+1} \mid s_{t+1}, s_t, Y_t, è\right) p(s_{t+1} \mid s_t, Y_t, è) p(s_t \mid Y_t, è)$$

The term $p\left(Y^{t+1}, S^{t+1}, Y_t, è\right)$ is independent of s_t and hence becomes part of the normalizing constant. Due to the Markovian property of the states the term $p\left(S^{t+2}, Y^{t+1} \mid s_{t+1}, s_t, Y_t, è\right) = p\left(S^{t+2}, Y^{t+1} \mid s_{t+1}, Y_t, è\right)$ which is also independent of s_t and becomes part of the normalizing constant. Furthermore, the Markovian property of the states means that s_{t+1} is purely dependent on knowledge of s_t. Thus $p(s_{t+1} \mid s_t, Y_t, è)$ becomes $p(s_{t+1} \mid s_t, è)$. Therefore a simplified expression for the typical term of the joint posterior density is the product of two terms:

$$p\left(s_t \mid Y_N, S^{t+1}, è\right) \propto p(s_{t+1} \mid s_t, è) \, p(s_t \mid Y_t, è) \tag{D.2}$$

The first term is the transition probability of going from s_t to s_{t+1}, and the other term is the mass function of s_t given Y_t. The normalizing constant of this mass function is the sum of the numbers obtained using (D.2) as $s_t \in \{WET, DRY\}$.

The final stage of the calculation is to determine the mass function $p(s_t \mid Y_t, è)$ given in (D.2). The method developed is applied recursively for all s_t from $t = 1$ to n. Assume that the function $p(s_{t-1} \mid Y_{t-1}, è)$ is available. Then repeat the following steps:

Prediction Step: Determine $p(s_t \mid Y_{t-1}, è)$ using the total probability theorem:

$$p\left(s_t \mid Y_{t-1}, è\right) = \sum_{k \in \{WET, DRY\}} p\left(s_t \mid s_{t-1} = k, è\right) p(s_{t-1} = k \mid Y_{t-1}, è) \tag{D.3}$$

where the Markovian property of the states means that $p(s_t \mid s_{t-1}, Y_{t-1}, è) = p(s_t \mid s_{t-1}, è)$

Update Step: Determine $p(s_t \mid Y_t, è)$ first by splitting the Y_t term so that it becomes $p(s_t \mid y_t, Y_{t-1}, è)$ and then applying Bayes theorem:

$$
\begin{aligned}
p(s_t|Y_t,è) & = p(s_t|\ y_t,Y_{t-1},è) \\
& = p(y_t|\ s_t,Y_{t-1},è)\ p(s_t|\ Y_{t-1},è) \\
& \propto p(y_t|\ s_t,è)\ p(s_t|\ Y_{t-1},è)
\end{aligned}
\tag{D.4}
$$

The left hand term $p(y_t|s_t,Y_{t-1},è)$ becomes $p(y_t|s_t,è)$ because y_t only depends on s_t and $è$. This result is the probability density of the rainfall at time t, y_t, given the climate state, s_t. As the rainfall distribution is assumed Gaussian this is easily evaluated. The right hand term is calculated in the *prediction* step. The normalizing constant for mass function given in (D.4) is the sum of all the terms for $s_t \in \{WET, DRY\}$. At $t = 1$ these steps can be initialized by ignoring the *prediction* step and using the stationary Markovian state probabilities derived from the state transition probability matrix **P** for $p(s_1|Y_0,è)$.

Using the expressions derived previously, the simulation of the state time series is a relatively simple procedure. First the *prediction* and *update* steps are run recursively to compute the mass functions $p(s_t|Y_t,è)$, for all $t = 1$ to n. The sampling of the state time series starts by initially simulating s_n using $p(s_n|Y_N,è)$. The remaining states, from s_{n-1} through to s_1 can be simulated using the mass function $p(s_t|Y_N,S^{t+1},è)$ calculated using the expression given in (D.2). An example illustrates this procedure. Suppose that s_{t+1} was sampled as a wet state. Then the probability of sampling s_t as state k, is calculated as follows:

$$
\begin{aligned}
p(s_t = k|\ Y_N,S^{t+1},è) & \propto p(s_{t+1} = Wet|s_t = k,è)\,p(s_t = k/Y_t,è) \\
& \propto p(k \to Wet)* p(s_t = k/Y_t,è)
\end{aligned}
\tag{D.5}
$$

where $p(k \to Wet)$ is the probability of jumping from state k to a wet state.

D.3   Derivation of the Likelihood Function for the HSM Model

In an earlier work, *Chib* [1995] outlines a methodology for deriving the likelihood function for the HSM model which utilises some of the results given above. This is done without including the hidden state time series as a model parameter. Hence, the vector of unknown model parameters is as given in Equation (5.5), where:

$$\grave{e}^{*\prime} = (\ \mu_W, \sigma_W, \mu_D, \sigma_D, \mathbf{P})$$ (D.6)

The procedure for calculating the likelihood function $p(Y_N|\theta^*)$ is derived by first rewriting $p(Y_N|\theta^*)$ and repeatedly applying the conditional probability theorem, to develop a recursive expression, similar to Step 1 in Section D.2 above:

$$
\begin{aligned}
p\ (Y_N|\grave{e}^*) &= p(y_n, Y_{N-1}|\grave{e}^*) \\
&= p(y_n|Y_{N-1}, \grave{e}^*)p(Y_{N-1}|\grave{e}^*) \\
&= p(y_n|Y_{N-1}, \grave{e}^*)p(y_{n-1}|Y_{N-2}, \grave{e}^*)p(Y_{N-2}|\grave{e}^*)
\end{aligned}
$$

The summary of this recursion is:

$$p\ (Y_N|\grave{e}^*) = p(y_n|Y_{N-1}, \grave{e}^*)\dots p(y_t|Y_{t-1}, \grave{e}^*)\dots p(y_1|\grave{e}^*)$$ (D.7)

The typical term, excluding the terminal point is therefore: $p(y_t|Y_{t-1}, \grave{e}^*)$

Now, given the HSM modelling structure if the total probability theorem is applied to this typical term then:

$$
\begin{aligned}
p(y_t|Y_{t-1}, \grave{e}^*) &= \sum_{k \in \{WET, DRY\}} p\ (y_t|Y_{t-1}, s_t = k, \grave{e}^*)\ p(s_t = k\ |Y_{t-1}, \grave{e}^*) \\
&= \sum_{k \in \{WET, DRY\}} p\ (y_t|\ s_t = k, \grave{e}^*)\ p(s_t = k\ |Y_{t-1}, \grave{e}^*)
\end{aligned}
$$ (D.8)

The second line follows because the rainfall y_t is purely dependent on knowledge of s_t. The result is the probability density of the rainfall y_t given the state s_t. As mentioned, because the rainfall distribution is assumed Gaussian this is easily evaluated. The right hand term of Equation (D.8) is the time varying probability mass function that is given in the prediction step in Section D.2 above. For the terminal point $p(y_1|\grave{e}^*)$ the likelihood function can be evaluated using the stationary Markovian state probabilities derived from the state transition probability matrix $\mathbf{P}$ for $p(s_1|Y_0, \grave{e}^*)$.

Given the likelihood value for this typical term the full likelihood function can therefore be calculated using:

$$p\left(Y_N \mid \grave{e}^*\right) = \prod_{t=1}^{n} p\left(y_t \mid Y_{t-1}, \grave{e}^*\right) \qquad (D.9)$$

Hence a relatively simple procedure for calculating the likelihood function value for a given parameter vector θ^* of the HSM model is illustrated. This can be applied to either the single or multi-site HSM modelling framework. The only difference is the probability density function used to calculate $p\left(y_t \mid s_t = k, \grave{e}^*\right)$. For the single site case it is the univariate Gaussian and in the multisite case it is the multivariate Gaussian probability density function.

D.4   Sampling the State Transition Probabilities

Given knowledge of the hidden state time series the transition probabilities become independent of the data, Y_N. This leads to a straightforward procedure for deriving their conditional posterior $p\left(\mathbf{P} \mid S_N\right)$. *Chib* [1996] provides a method for sampling from the state transition probability matrix of a multi-state hidden Markov model. Presented here will be a method for the two-state case of the HSM model.

The conditional posteriors for both the transition probabilities p_{WD} and p_{DW} given in $\mathbf{P}$ are equivalent. Thus this derivation will be given for a single p_{ij} $(i \neq j)$, which can represent either p_{WD} or p_{DW}.

The first step in determining the conditional posterior $p\left(p_{ij} \mid S_N\right)$ is to determine the likelihood function $p\left(S_N \mid p_{ij}\right)$. Now given the number of times a particular state i appears in the hidden state time series n_i the number of times a transition is made from that state to the opposing state j, n_{ij} is referred to as the number of 'successes' (a success here refers to a successful state transition) in a sequence of n_i iid Bernouilli trials. Therefore the likelihood for the probability of a state transition p_{ij} follows a Binomial distribution such that:

$$S_N|p_{ij} \sim Bin(n_i, p_{ij})$$
$$p(S_N|p_{ij}) \propto (p_{ij})^{n_{ij}} (1-p_{ij})^{n_i - n_{ij}} \tag{D.10}$$

This likelihood represents the probability of n_{ij} state transitions in n_i trials with a given state transition probability p_{ij}. If we assume a Beta distribution for the prior of p_{ij} then:

$$p_{ij} \sim Beta(\acute{a}, \hat{a})$$
$$p(p_{ij}) \propto (p_{ij})^{\acute{a}-1} (1-p_{ij})^{\hat{a}-1} \tag{D.11}$$

where α and β are prior parameters. If this prior is updated by the likelihood given in (D.10) then the resulting posterior is:

$$
\begin{aligned}
p(p_{ij}|S_N) &\propto p(S_N|p_{ij}) p(p_{ij}) \\
&\propto (p_{ij})^{n_{ij}} (1-p_{ij})^{n_i - n_{ij}} (p_{ij})^{\alpha-1} (1-p_{ij})^{\beta-1} \\
&\propto (p_{ij})^{n_{ij}+\alpha-1} (1-p_{ij})^{n_i - n_{ij}+\beta-1}
\end{aligned} \tag{D.12}
$$

which has the form of a Beta distribution, and therefore:

$$p_{ij}|S_N \sim Beta(\acute{a} + n_{ij}, \hat{a} + n_i - n_{ij}) \tag{D.13}$$

This provides a good illustration of a *conjugate prior* distribution. A conjugate prior has the property that when it is updated using the likelihood it forms a posterior that is from the same parametric family. In the example above, the Beta distribution is the conjugate prior for the Binomial likelihood. Conjugate priors are mathematically convenient because it means that the posterior is from a known parametric family and is therefore easy to work with. To calculate the posterior the parameters of the prior are simply updated using information contained in the data, as shown above. The concept of a conjugate prior will be applied repeatedly in the following sections of this appendix.

Now, if the prior parameters α, β are not fixed then they would be known as *hyperparameters*. In this case the Beta distribution is uniform if $\alpha = \beta = 1$. As uninformative priors are desirable in a Bayesian framework, these parameter values will

be used for the prior. Often the choice of suitable prior and appropriate hyperparameter values is not so simple.

To sample values for both p_{WD} and p_{DW} from the posterior given in (D.12) a Gamma distribution can be used [*Chib,* 1996]:

$$p_{WD} = \frac{x_{WD}}{x_{WD} + x_{WW}} \quad p_{DW} = \frac{x_{DW}}{x_{DW} + x_{DD}} \quad x_{ij} \sim Gamma(1 + n_{ij}, 1)$$

D.5 Sampling the State Rainfall Parameters

In this section the methodology for sampling the state rainfall parameters in both the single site and multi-site context will be given. As explained in Appendix E, the results presented in this thesis for the single site HSM model used the prior specification for the state rainfall parameters given in Equations (5.10) and (5.11), referred to as the P3 prior. However, a different prior specification (referred to as the P1 prior) was originally used for the state rainfall parameters for the results published in *Thyer and Kuczera* [2000a]. The conditional posteriors for the state rainfall parameters slightly change depending on whether the P1 or P3 prior is used. The P3 prior assumes μ and σ are jointly unknown, hence their conditional posterior is $p(\mu, \sigma | Y_N)$. This can be further broken down using the relationship $p(\mu, \sigma | Y_N) = p(\mu | \sigma, Y_N) p(\sigma | Y_N)$. The conditional posterior for σ is therefore only conditioned on the data Y_N. In contrast using the P1 prior formulation the conditional posterior for σ is conditioned on known μ and the data, $p(\sigma | \mu, Y_N)$. For μ the conditional posterior is still $p(\mu | \sigma, Y_N)$. Both these formulations are allowable in the Gibbs sampling framework. However, as stated in Appendix E, changing the hyperparameter values for the P1 prior depending on the sampled hidden state time series is not allowable. This is the reason why the P1 prior is not used in this thesis.

For completeness the conditional posteriors for the P1 prior (unknown μ conditioned on known σ and unknown σ conditioned on known μ) will be given. This will provide an excellent introduction to working with Gaussian distributions for the evaluation of Bayesian posteriors. As the motivation for using the P3 prior stems from the multi-site context [refer to Section 10.2.1] the conditional posteriors for the multi-

site state rainfall parameters will then be derived. Finally, the equivalent conditional posteriors for the P3 prior in the single site context (μ and σ jointly unknown) will be given.

Similar to the transitions probabilities the conditional posteriors for the rainfall parameters of both the wet and dry states are equivalent. Hence in the following sections μ, σ^2, $\grave{\imath}$ and $\acute{O}$ represent a generic state mean, variance, mean vector and covariance matrix respectively for either the wet or dry state rainfall distributions and Y_N (or $\mathbf{Y_N}$) refers to the n points (or vectors) of data classified in either the wet or dry states.

D.5.1   Sampling the state mean conditioned on the state variance

The likelihood function for the state mean, given a known variance and Y_N data which is assumed to follow a Gaussian distribution is written as:

$$
\begin{aligned}
p(Y_N|\mu,\sigma^2) &= \prod_{i=1}^{n} p(y_i|\mu,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\
&\propto \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right]
\end{aligned}
\tag{D.14}
$$

Now, if a conjugate prior distribution with the following parameterization is used:

$$
p(\mu) \propto \exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right]
\tag{D.15}
$$

such that $\mu \sim N(\mu_0,\tau_0^2)$, with prior mean μ_0 and variance τ_0^2 the resulting posterior density for the state mean, conditioned on the data and the variance is:

$$
\begin{aligned}
p(\mu|Y_N,\sigma^2) &\propto p(Y_N|\mu,\sigma^2)p(\mu) \\
&\propto exp\left[-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right]\prod_{i=1}^{n} exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\
&\propto exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\mu-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]\right)
\end{aligned}
\tag{D.16}
$$

Algebraic simplification of this expression leads to the result that the posterior depends only the data Y_N through the average of the state data, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. This results in the following expression for the posterior:

$$p(\mu|\bar{y},\sigma^2) \propto exp\left[-\frac{1}{2\tau_n^2}(\mu - \mu_n)^2\right] \qquad (D.17)$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_k^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_k^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \qquad (D.18)$$

Again, the property of a conjugate prior can be seen because the posterior given in Equation (D.17) is a Gaussian distribution with:

$$\mu|\bar{y},\sigma^2 \sim N(\mu_n, \tau_n^2) \qquad (D.19)$$

Using these updated values, the posterior mean μ_n and posterior variance τ_n^2, the state mean can be sampled from a Gaussian distribution. When working with Gaussian distributions the inverse of the variance, termed the *precision*, plays an important role. In a parameter estimation context it represents the degree of uncertainty about the true value for that parameter. If the *posterior precision* $\left(\frac{1}{t_n^2}\right)$ is high (low variance) then the uncertainty is low, and conversely if the precision is low (high variance) then the uncertainty is high. Equation (D.18) demonstrates that the *posterior precision* equals the *prior precision* plus the *data precision.* The posterior mean is a expressed as a weighted average of the prior mean and the average of the observed data, with weights proportional to the precision.

An intuitive interpretation of the respective roles of the prior and the data can be seen by considering some extreme cases using Equation (D.18). As the number of data is increased $(n \to \infty)$ or the prior precision is decreased $(\tau_0 \to \infty)$ then the posterior will be largely dependent on the data $\left(\mu_n \to \bar{y}, \tau_n^2 \to \sigma_n^2/n\right)$. If the prior precision is increased $(\tau_0 \to 0)$, or the number of data is decreased $(n \to 0)$ then the prior and the

posterior will become identical $\left(\mu_n \to \mu_0, \tau_n \to \tau_0 \to 0\right)$. If $\tau_0 = \sigma$ then the prior distribution has the same weight as one extra data point with the value μ_0. For all other cases in between the posterior represents a compromise between the prior and the observed data.

The values used for the prior hyperparameters μ_0 and τ_0^2 are given in Appendix E.

D.5.2 Sampling the state variance conditioned on the state mean

The derivation of the conditional posterior for the state variance σ^2 follows a similar line as the results for the state mean. The likelihood function of the state variance σ^2 given a known mean μ and vector of state data Y_N which is assumed to follow a Gaussian distribution is written as:

$$
\begin{aligned}
p\left(Y_N \big| \sigma^2, \mu\right) &= \prod_{i=1}^{n} p\left(y_i \big| \sigma^2, \mu\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\
&\propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \\
&\propto \left(\sigma^2\right)^{-n/2} \exp\left[-\frac{ns^2}{2\sigma^2}\right]
\end{aligned}
\tag{D.20}
$$

where s^2 is the average squared deviation of the data, such that $s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$, which is constant for a given μ and Y_N.

A suitable conjugate prior density is the scaled inverse $-\chi^2$ distribution, with prior scale s_0^2 and n_0 prior degrees of freedom, such that:

$$
\sigma^2 \sim Inv - \chi^2\left(v_0, \sigma_0^2\right)
$$
$$
p\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-(v_0/2+1)} exp\left[-\frac{v_0 \sigma_0^2}{2\sigma^2}\right]
\tag{D.21}
$$

This prior distribution can be thought of as providing information equivalent to ν_0 observations with average squared deviation σ_0^2. When this prior is updated with the likelihood given in (D.20), the posterior is:

$$
\begin{aligned}
p\left(\sigma^2 | Y_N, \mu\right) \ &\propto \ p\left(Y_N | \sigma^2, \mu\right) p\left(\sigma^2\right) \\
&\propto \ \left(\sigma^2\right)^{-n/2} \exp\left[-\frac{ns^2}{2\sigma^2}\right] \left(\sigma^2\right)^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right] \\
&\propto \ \left(\sigma^2\right)^{-((n+\nu_0)/2+1)} \exp\left[-\frac{1}{2\sigma^2}\left(ns^2 + \nu_0 \sigma_0^2\right)\right] \\
&\propto \ \left(\sigma^2\right)^{-(\nu_n/2+1)} \exp\left[-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right]
\end{aligned}
$$

$$(D.22)$$

where

$$
\nu_n = \nu_0 + n \quad \text{and} \quad \sigma_n^2 = \frac{ns^2 + \nu_0 \sigma_0^2}{\nu_0 + n}
\tag{D.23}
$$

which is a scaled $\text{inverse} - \chi^2$ distribution, with posterior degrees of freedom, ν_n and posterior scale σ_n^2, such that:

$$
\sigma^2 | Y_N, \mu \ \sim \ Inv - \chi^2\left(\nu_n, \sigma_n^2\right)
\tag{D.24}
$$

The roles of the prior and the data are similar to the case for the state mean. From Equation (D.23) it can be seen that the posterior scale is the weighted average of the prior scale and the data scale, with the weights proportional to the prior and data degrees of freedom. The posterior degrees of freedom is the sum of the prior and data degrees of freedom. Thus similar results also hold for the extreme cases, as given for the state mean, e.g, as the number of data increases $(n \rightarrow \infty)$ the posterior becomes dominated by the data, $\sigma_n \rightarrow s$. The values used for the hyperparameters ν_0 and σ_0^2 are given in Appendix E.

D.5.3   Sampling the state multivariate mean and covariance matrix jointly

The conditional posteriors for the multi-site state distribution parameters are derived assuming that both the mean vector $\mathbf{\mu}$ and covariance matrix $\mathbf{\Sigma}$ are jointly unknown. This setup is used because of the chosen prior specification, refer to Section 10.2.1 for

more details. It is possible to derive the conditional posteriors for the univariate case where both μ and σ jointly unknown. However, in this thesis the multivariate derivation will be given and then the univariate equivalent of the conditional posteriors will be outlined.

Now, given n vectors of dimension r of observed data $\mathbf{Y_N} = \{\mathbf{y}_1, \ldots, \mathbf{y_n}\}$ that are assumed to follow a r-dimensional multivariate Gaussian distribution with a $r \times r$ covariance matrix $\mathbf{\acute{O}}$ and r-dimensional mean vector, $\mathbf{ì}$, the joint likelihood of $(\mathbf{ì}, \mathbf{\acute{O}^{-1}})$ can be written as:

$$
\begin{aligned}
p(\mathbf{Y_N}|\mathbf{ì}, \mathbf{\acute{O}^{-1}}) \; &= \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{ì}, \mathbf{\acute{O}^{-1}}) \\
&= \prod_{i=1}^{n} (2\pi)^{-r/2} |\mathbf{\acute{O}^{-1}}|^{1/2} \exp\left(-\tfrac{1}{2}(\mathbf{y_i}-\mathbf{ì})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{y_i}-\mathbf{ì})\right) \\
&\propto \prod_{i=1}^{n} |\mathbf{\acute{O}^{-1}}|^{1/2} \exp\left(-\tfrac{1}{2}(\mathbf{y_i}-\mathbf{ì})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{y_i}-\mathbf{ì})\right) \\
&\propto |\mathbf{\acute{O}^{-1}}|^{n/2} \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{ì})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{y_i}-\mathbf{ì})\right) \\
&\propto |\mathbf{\acute{O}^{-1}}|^{n/2} \exp\left(-\tfrac{1}{2}\left[n(\mathbf{ì}-\overline{\mathbf{y}})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{ì}-\overline{\mathbf{y}}) + \mathrm{tr}(\mathbf{SS} \cdot \mathbf{\acute{O}^{-1}})\right]\right)
\end{aligned}
\qquad \text{(D.25)}
$$

where $\overline{\mathbf{y}}$ is the average vector of the observed data, $\overline{\mathbf{y}} = \tfrac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$. The notation $tr(\mathbf{A})$ refers to the trace of the matrix $\mathbf{A}$, which is sum of the diagonal elements of the matrix, and $\mathbf{SS}$ is the sum of squares matrix about the sample mean, where

$$
\mathbf{SS} = \sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^{\mathrm{T}}
\qquad \text{(D.26)}
$$

The last step of the derivation of the joint likelihood uses a result given by *DeGroot* [1970], that

$$
\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{ì})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{y_i}-\mathbf{ì}) = n(\mathbf{ì}-\overline{\mathbf{y}})^{\mathrm{T}} \mathbf{\acute{O}^{-1}}(\mathbf{ì}-\overline{\mathbf{y}}) + \mathrm{tr}(\mathbf{SS} \cdot \mathbf{\acute{O}^{-1}})
\qquad \text{(D.27)}
$$

The reasons for rearranging the joint likelihood into an expression of the form given in (D.25) will become clear further on.

The conjugate joint prior distribution for the joint likelihood given in (D.25) is known as the multivariate Gaussian-Wishart distribution. This is derived using the relationship

that $p(\mathbf{ì},\acute{\mathbf{O}}^{-1}) = p(\mathbf{ì}|\acute{\mathbf{O}}^{-1})p(\acute{\mathbf{O}}^{-1})$. For the $p(\mathbf{ì}|\acute{\mathbf{O}}^{-1})$ prior a multivariate Gaussian distribution is parameterized in the following form:

$$\mathbf{ì}|\acute{\mathbf{O}}^{-1} \sim N_r(\mathbf{ì}_0, \acute{\mathbf{O}}/\kappa_0)$$
$$p(\mathbf{ì}|\acute{\mathbf{O}}^{-1}) \propto \exp\left(-\tfrac{1}{2}\kappa_0(\mathbf{ì}-\mathbf{ì}_0)^T\acute{\mathbf{O}}^{-1}(\mathbf{ì}-\mathbf{ì}_0)\right) \tag{D.28}$$

where $\mathbf{ì}_0$ is the prior mean vector and κ_0 represents the number of prior measurements on the $\acute{\mathbf{O}}$ scale.

For the prior $p(\acute{\mathbf{O}}^{-1})$ the following parameterisation is used:

$$\acute{\mathbf{O}}^{-1} \sim W_r(\nu_0, \mathbf{W_0})$$
$$p(\acute{\mathbf{O}}^{-1}) \propto |\acute{\mathbf{O}}^{-1}|^{(\nu_0-r-1)/2} \exp\left(-\tfrac{1}{2}tr(\mathbf{W_0}^{-1}\acute{\mathbf{O}}^{-1})\right) \tag{D.29}$$

where $W_r(\nu, \mathbf{W})$ represents a Wishart distribution in r dimensions with ν degrees of freedom and scale matrix $\mathbf{W}$. The Wishart distribution is the multivariate generalisation of the scaled inverse $-\chi^2$ distribution. The prior parameter ν_0 represents the number of prior degrees of freedom and because $\mathbf{W_0}$ controls the form of the precision matrix $\acute{\mathbf{O}}^{-1}$ it is known as the prior precision matrix.

Using (D.28) and (D.29) the joint prior density can be written as:

$$
\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) \quad &= p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}^{-1}) p(\boldsymbol{\Sigma}^{-1}) \\
&\propto \exp\left(-\tfrac{1}{2}\kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right) \cdot \left|\boldsymbol{\Sigma}^{-1}\right|^{(v_0 - r - 1)/2} \exp\left(-\tfrac{1}{2}\mathrm{tr}\left(\mathbf{W_0}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right) \\
&\propto \left|\boldsymbol{\Sigma}^{-1}\right|^{(v_0 - r - 1)/2} \exp\left(-\tfrac{1}{2}\left[\kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \mathrm{tr}\left(\mathbf{W_0}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right]\right)
\end{aligned}
\tag{D.30}
$$

Now it can be seen why the likelihood was rearranged in the form given in (D.25), because it gives a similar form to the prior density. When this prior is updated using the likelihood the following posterior results:

$$
\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathbf{Y_N}) \quad &\propto p(\mathbf{Y_N} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) \\
&\propto \left|\boldsymbol{\Sigma}^{-1}\right|^{n/2} \exp\left(-\tfrac{1}{2}\left[n(\boldsymbol{\mu} - \overline{\mathbf{y}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \overline{\mathbf{y}}) + \mathrm{tr}(\mathbf{SS} \cdot \boldsymbol{\Sigma}^{-1})\right]\right) \cdot \left|\boldsymbol{\Sigma}^{-1}\right|^{(v_0 - r - 1)/2} \exp\left(-\tfrac{1}{2}\left[\kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \mathrm{tr}\left(\mathbf{W_0}^{-1}\boldsymbol{\Sigma}^{-1}\right)\right]\right) \\
&\propto \left|\boldsymbol{\Sigma}^{-1}\right|^{(n + v_0 - r - 1)/2} \exp\left(-\tfrac{1}{2}\left[n(\boldsymbol{\mu} - \overline{\mathbf{y}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \overline{\mathbf{y}}) + \kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \mathrm{tr}\left((\mathbf{W_0}^{-1} + \mathbf{S})\boldsymbol{\Sigma}^{-1}\right)\right]\right)
\end{aligned}
\tag{D.31}
$$

Now *DeGroot* [1970] gives the result that:

$$
n(\boldsymbol{\mu} - \overline{\mathbf{y}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \overline{\mathbf{y}}) + \kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = (\kappa_0 + n)(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n) + \frac{\kappa_0 n}{(\kappa_0 + n)}(\boldsymbol{\mu}_0 - \overline{\mathbf{y}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \overline{\mathbf{y}})
\tag{D.32}
$$

where an expression for $\boldsymbol{\mu}_n$ will be given later. Furthermore as the rightmost term of (D.32) can be written as follows:

$$\frac{\kappa_0 n}{(\kappa_0 + n)}(\grave{\imath}_0 - \overline{\mathbf{y}})^{\mathrm{T}}\acute{\mathbf{O}}^{-1}(\grave{\imath}_0 - \overline{\mathbf{y}}) = tr\left(\frac{\kappa_0 n}{(\kappa_0 + n)}(\grave{\imath}_0 - \overline{\mathbf{y}})^{\mathrm{T}}(\grave{\imath}_0 - \overline{\mathbf{y}})\acute{\mathbf{O}}^{-1}\right)$$

(D.33)

then the posterior from (D.31) can be rewritten as:

$$
\begin{aligned}
\mathrm{p}(\grave{\imath}, \acute{\mathbf{O}}^{-1}|\mathbf{Y_N}) \quad &\propto \left|\acute{\mathbf{O}}^{-1}\right|^{(n+v_0-r-1)/2} \exp\left(-\tfrac{1}{2}\left[(\kappa_0 + n)(\grave{\imath} - \grave{\imath}_n)^{\mathrm{T}}\acute{\mathbf{O}}^{-1}(\grave{\imath} - \grave{\imath}_n) + tr\left(\left(\mathbf{W_0}^{-1} + \mathbf{SS} + \frac{\kappa_0 n}{(\kappa_0 + n)}(\grave{\imath}_0 - \overline{\mathbf{y}})^{\mathrm{T}}(\grave{\imath}_0 - \overline{\mathbf{y}})\right)\acute{\mathbf{O}}^{-1}\right)\right]\right) \\
&\propto \left|\acute{\mathbf{O}}^{-1}\right|^{(v_n-r-1)/2} \exp\left(-\tfrac{1}{2}\left[\kappa_n(\grave{\imath} - \grave{\imath}_n)^{\mathrm{T}}\acute{\mathbf{O}}^{-1}(\grave{\imath} - \grave{\imath}_n) + tr\left(\mathbf{W_n}^{-1}\acute{\mathbf{O}}^{-1}\right)\right]\right)
\end{aligned}
$$

(D.34)

As expected this posterior has the same parametric form as the conjugate prior density given in (D.30). The full list of updated parameter values is given:

$$
\begin{aligned}
\grave{\imath}_n &= \frac{\kappa_0 \grave{\imath}_0 + n\overline{\mathbf{y}}}{\kappa_0 + n} && \kappa_n = \kappa_0 + n \\
\mathbf{W_n} &= \left(\mathbf{W_0}^{-1} + \mathbf{SS} + \frac{\kappa_0 n}{(\kappa_0 + n)}(\grave{\imath}_0 - \overline{\mathbf{y}})^{\mathrm{T}}(\grave{\imath}_0 - \overline{\mathbf{y}})\right)^{-1} && v_n = v_0 + n
\end{aligned}
$$

(D.35)

To sample from the joint posterior of $\left(\grave{\mathbf{i}}, \acute{\mathbf{O}}^{-1}\right)$, one must use the following procedure: first draw $\acute{\mathbf{O}}^{-1}|\mathbf{Y_N} \sim W(\nu_n, \mathbf{W_n})$, and then draw $\grave{\mathbf{i}}\left|\acute{\mathbf{O}}^{-1}, \mathbf{Y_N} \sim N_r(\grave{\mathbf{i}}_n, \acute{\mathbf{O}}/\kappa_n)\right.$.

Again, the respective role of the priors and data which was shown in the univariate cases, similarly follows for the multivariate case. The posterior mean is again the weighted sum of the prior mean and data average, with weights proportional to the number of prior measurements and data observations. If $\mathbf{n}_0 = 1$, then the prior information is equivalent to one prior observation with squared deviation matrix, $\mathbf{W_0}^{-1}$. As the number of prior measurements on the $\acute{\mathbf{O}}$ scale is increased, then the posterior of the mean will be concentrated at the prior mean, $\grave{\mathbf{i}}_n \rightarrow \grave{\mathbf{i}}_0$.

An important thing to note with these multivariate distributions is that if improper uninformative priors are chosen e.g. $\kappa_0, \nu_0 = 0$, then the posteriors can be become improper if the number of observations is less than the dimension of the distributions $n < r$. Improper posteriors can also result if the priors are chosen such that $\nu_n < r$. This has important implications for the choice of hyperparameter values and is further discussed in Section 10.2.1

D.5.4   Sampling the state mean and variance jointly

The univariate version of the multivariate posterior density given in (D.34) is the conditional posterior of the state mean and variance assuming they are jointly unknown. For the univariate case $r = 1$, $\grave{\mathbf{i}}$ becomes $\grave{\imath}$, $\acute{\mathbf{O}}$ becomes σ^2, and hence

$$
\begin{aligned}
p\left(\mu, \sigma^2 | Y_N\right) &\propto \left(\sigma^2\right)^{-(\nu_n/2+1)} exp\left(-\frac{1}{2}\left[\kappa_n \frac{\left(\grave{\imath} - \grave{\imath}_n\right)^2}{\sigma^2} + \frac{\nu_n \sigma_n^2}{\sigma^2}\right]\right) \\
&\propto exp\left(-\frac{\left(\grave{\imath} - \grave{\imath}_n\right)^2}{2\sigma^2/\kappa_n}\right)\left(\sigma^2\right)^{-(\nu_n/2+1)} exp\left(-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right)
\end{aligned}
\tag{D.36}
$$

with

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{y}}{\kappa_0 + n} \qquad\qquad\qquad \kappa_n = \kappa_0 + n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + ns^2 + \frac{\kappa_0 n}{(\kappa_0 + n)}(\mu_0 - \bar{y})^2 \quad \nu_n = \nu_0 + n \qquad\qquad (D.37)$$

where $\nu_n \sigma_n^2$ has replaced $\mathbf{W_n^{-1}}$, and $\nu_0 \sigma_0^2$ has replaced $\mathbf{W_0^{-1}}$. The posterior density given in (D.36) for the case where μ and σ are jointly unknown can be seen to have similar form to the posterior densities given in (D.17) and (D.22), for the cases where μ is unknown conditioned on known σ and σ is unknown conditioned on known μ respectively. Hence sampling from this univariate posterior is equivalent to sampling $\sigma^2 | Y_N \sim Inv - \chi^2(\nu_n, \sigma_n^2)$ and $\mu | Y_N, \sigma^2 \sim N(\mu_n, \sigma^2/\kappa_n)$, with equivalent priors:

$$\mu \sim N(\mu_0, \sigma^2/\kappa_0)$$
$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2) \qquad\qquad (D.38)$$

Note that the equation for calculating σ_n^2 is slightly different from that given in (D.23). This is because in this case the mean and the variance are assumed to jointly unknown. The third term in the equation for σ_n^2 represents the additional uncertainty conveyed by the difference between the prior mean and the sample mean.

Appendix E - Original Single Site Prior Specification

E.1   Introduction

In the results published in *Thyer and Kuczera* [2000a] where the single site HSM model was calibrated to the rainfall data from Sydney, Brisbane and Melbourne the prior specification used for the state rainfall parameters was different to the one given in Section 5.2.4.a. In this Appendix it will be shown why it was deemed necessary to change the prior specification for the state rainfall parameters in the single site HSM model.

E.2   Original Prior Specification

In the original prior specification, deemed the P1 prior specification, a conjugate prior Gaussian distribution with prior mean μ_0 and prior variance τ_0^2 was used. Hence:

$$\mu_k \sim N(\mu_0, \tau_0^2) \tag{E.1}$$

For both the wet and dry state the values $\mu_0 = 1000.0$ and $\tau_0 = 1000.0$ were used for the prior as simulations revealed that it provided a suitably diffuse proper prior. For the state variance a scaled $inverse - \chi^2$ distribution with prior scale $\boldsymbol{s}_0^2$ and $\boldsymbol{n}_0$ prior degrees of freedom, was used:

$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2) \tag{E.2}$$

As explained in Section 5.2.4.a when $\nu_0 = 0$ this represents an uninformative improper prior. In the P1 prior this improper prior was used when at least two data points were sampled in a particular state, as a proper posterior still results. However, when no data was sampled in a particular state, then the posterior also becomes improper. Also when only one data point is sampled in a particular state, this can result in what is termed by *Gelman et al.* [1995] as an "uninteresting" mode where there is one single data point with no variance. For these cases a different set of hyperparameter values ($\nu_0 = 2, \sigma_0 = 300.0$) were used which resulted in a relatively diffuse proper prior. The motivation for using this scheme was to alleviate the problem that improper priors could not be used because the posterior becomes improper when no data was sampled in a particular state. By changing to an uninformative improper prior when there was enough data to ensure a proper posterior the aim was to let the data dominate the inferences.

Further investigations undertaken during the development of the calibration procedure for the multi-site HSM model revealed that a different prior formulation, called the P3 prior specification, was required for successful implementation of the Gibbs sampler in a multi-site context. Refer to Chapter 10 for further details and an explanation of what happened to the P2 prior specification. The single site equivalent of this P3 prior specification is given in Equations (5.10) and (5.11). It is repeated here for convenience:

$$\mu \sim N(\mu_0 = \bar{y}, \sigma^2)$$
$$\sigma^2 \sim Inv - \chi^2(\nu_0 = 2, \sigma_0^2 = s^2) \tag{E.3}$$

where $\bar{y}$ is the empirical mean and s^2 is empirical variance (average squared deviation from the empirical mean) of the entire data series Y_N.

It is worth noting that the P1 prior slightly changes the conditional posteriors given in Equation (5.9) for sampling the state rainfall parameters in the single site context. The P3 prior assumes μ and σ are jointly unknown, hence their conditional posterior is $p(\mu,\sigma|Y_N)$. This can be further broken down using the relationship $p(\mu,\sigma|Y_N) = p(\mu|\sigma,Y_N)\,p(\sigma|Y_N)$. Hence the conditional posterior for σ is only conditioned on the data Y_N. In contrast using the P1 prior formulation the conditional posterior for σ is conditioned on known μ and the data, $p(\sigma|\mu,Y_N)$. For μ the conditional posterior is still $p(\mu|\sigma,Y_N)$. Therefore using the P1 prior the conditional posteriors for the state rainfall parameters given in Equation (5.9) change to:

$$\begin{aligned}
\mu_k^i &\leftarrow p\left(\,\mu_k \mid S_N^i, \sigma_k^{i-1}, Y_N\,\right) \\
\sigma_k^i &\leftarrow p\left(\,\sigma_k \mid S_N^i, \mu_k^i, Y_N\,\right)
\end{aligned} \tag{E.4}$$

These are the conditional posteriors given in *Thyer and Kuczera* [2000a]. Either one of the formulations given in Equations (5.9) or (E.4) is allowable in the Gibbs sampler framework. The methodology for sampling from the conditional posteriors given in (E.4) is described in Appendix D.

E.3 Comparison of Results

The Sydney rainfall data was reanalyzed with the P3 prior and compared to the results given in *Thyer and Kuczera* [2000a] for the original P1 prior to determine whether the results were sensitive to the choice of the prior. A comparison between the posteriors that resulted from the P1 and P3 prior for all the single site HSM model parameters for the Sydney annual (June to May water year) rainfall data is given in Figure E.1. The June to May water year was used because *Thyer and Kuczera* [2000a] stated it had the strongest wet and dry state signal.

The results indicate that the inferences are sensitive to the choice of the prior, especially for the wet state rainfall parameters, μ_W and σ_W, and the wet to dry state transition

probability, p_{WD}. It is believed this is mainly due to the scheme used in the P1 prior for the state variance. The σ_W posterior using the P1 prior is shown to be bimodal, whereas using the P3 prior it is unimodal [Figure E.1(f)]. The reasons for this bimodality are believed to be because the prior on the state variance is changed from an informative prior to an uninformative prior depending on the number of data in the state with the P1 prior. In the σ_W posterior using the P1 prior the mode that corresponds to the larger parameter value is similar to the mode for σ_W posterior using the P3 prior and is therefore believed to be caused by the data. The other mode in the σ_W posterior using the P1 prior (corresponding to the smaller parameter value) is believed to be due to the P1 prior scheme for the informative proper, which is denoted as SD prior P1 in Figure E.1(f). The difference between the location of the mode is because the curve denoted as SD prior P1 is not actually the true SD prior as it does not include the uninformative case where $p(\sigma^2) \propto 1/\sigma^2$.

(a) p_{DW} **(b)** p_{WD}

(c) μ_D **(d)** μ_W

(e) σ_D **(f)** σ_W
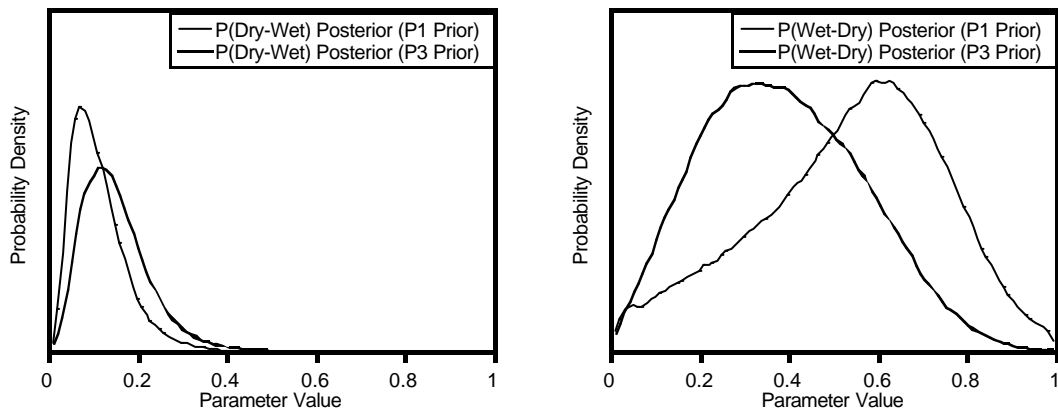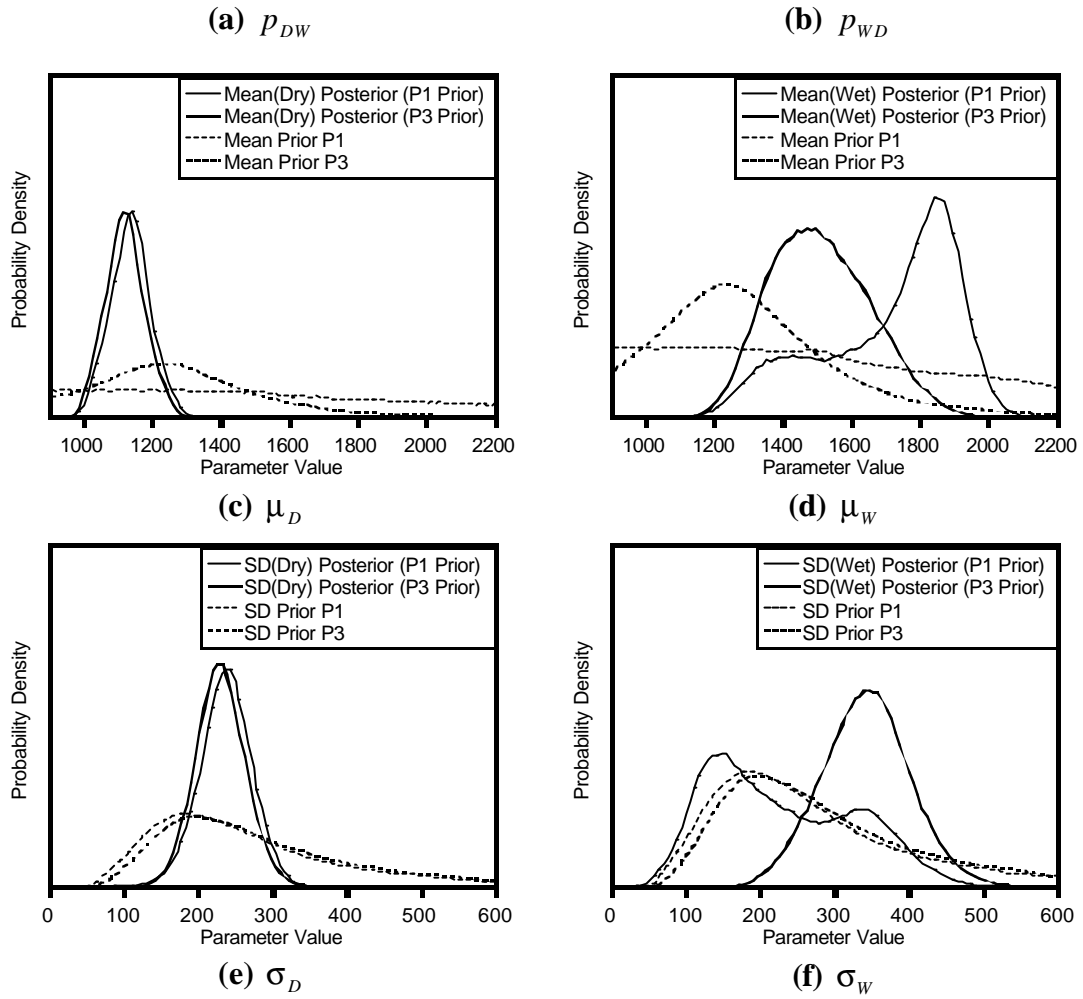
Figure E.1 – Comparisons of the posterior densities for the single site HSM model parameters for the Sydney annual rainfall data (June to May water year) for the P1 and P3 prior specifications. Also shown are the P1 and P3 prior densities (SD stands for standard deviation).

E.4   Implications

Sensitivity of the inferences to the choice of prior is not desirable in a Bayesian framework. Given the above result the P1 prior specification was re-examined and it was realized that it is not permitted in the Gibbs sampler framework to change the prior depending on the hidden state time series S_N. The hidden states are part of the data for the conditional posteriors of μ and σ. In Bayesian inference it is not allowed to have the prior change dependent on the data. The prior should remain constant and be independent of the data. Because the priors used were diffuse, this flaw only affected

the inferences when the number of data in a particular state was generally low. As Figure E.1 illustrates, the posteriors for the dry state parameters changed very little whether prior P1 or P3 was used because the number of data classified in the dry state was quite high.

It is important note that the overall conclusions from the inferences remained the same. Therefore the important findings given in *Thyer and Kuczera* [2000a] still stand. A two-state persistence structure was identified for the Sydney rainfall data. It was only the strength of the persistence structure and the wet state rainfall parameters that changed slightly.

E.5 Conclusion

Given the discovery of this violation in the P1 prior framework it was considered unacceptable for use in the Gibbs sampler and the P3 prior was therefore adopted for the use in the single site HSM model.

Appendix F - Verification of Single Site HSM Model Calibration Procedure

F.1 Introduction

To verify that the Gibbs sampler had been correctly formulated and the computer code used to implement the Gibbs sampler was free of any gross errors, synthetic calibration runs were used. Synthetic data was generated using the HSM model and then the Gibbs sampler was used to determine if it could recover the true parameter values. The results of this analysis are presented in this appendix for two different sets of parameters. The first set of parameters (denoted as set S1) had wet and dry rainfall distributions which were considered very well separated, i.e. examination of the marginal distribution of the entire time series clearly showed two separate distributions [Figure F.1(a)]. The transition probability values corresponded to an expected state residence of around 5 years. The second set of parameters (denoted as set S2) had wet and dry rainfall distributions which were much close together, compared to set S1 [Figure F.1(b)]. For this set the transition probability values were decreased slightly to correspond to an expected state residence time of 10 years for the wet state and 7 years for the dry state. Synthetic time series with 100, 1000, and 10,000 data points were applied because it would be expected that as the number of data points increases the posterior should converge to the true parameter value. The synthetic parameter values for set S1 and S2 are given in Table F.1.

Table F.1 – Synthetic parameter values.

Parameter Set	Wet State		Dry State		Transition Probabilities	
	μ_W	σ_W	μ_D	σ_D	p_{WD}	p_{DW}
S1	1000	5.0	500	5	0.2 (5)	0.2 (5)
S2	1200	240	800	160	0.1 (10)	0.15 (7)
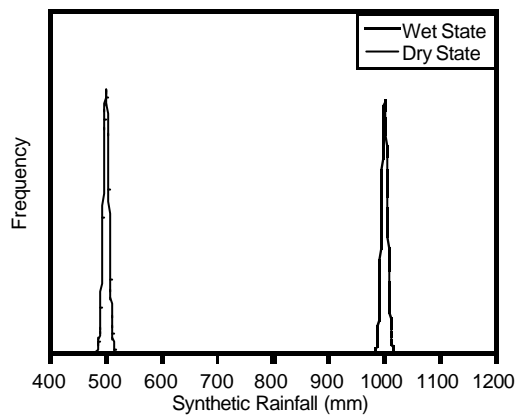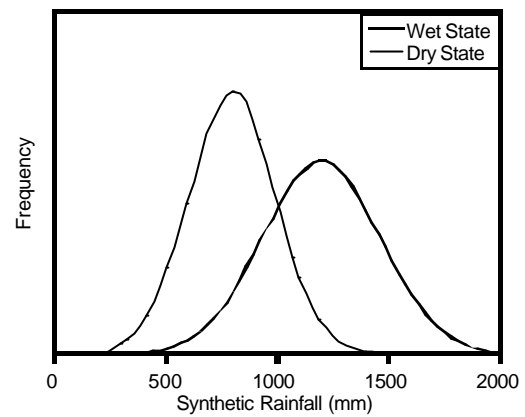
(a) S1 parameter set **(b) S2 parameter set**

Figure F.1 – Wet and dry state rainfall distributions for the synthetic eainfall series.

F.2   Results

The posteriors for p_{WD} and μ_W are shown in Figure F.2 for the S1 set and Figure F.3 for the S2 set. The posteriors are compared using percentile box plots. The bottom and top of the box correspond to the 5th and 95th percentiles respectively, the middle solid line is the median (50th percentile) and the two dashed lines are the 25th and 75th percentiles.   It is clearly seen that as the number of the data points increases the posterior of each parameter converges to the true parameter value. A similar result was found for all the other HSM model parameters.

(a) p_{WD} **(true value=0.2)**

(b) μ_W **(true value=1000.0)**
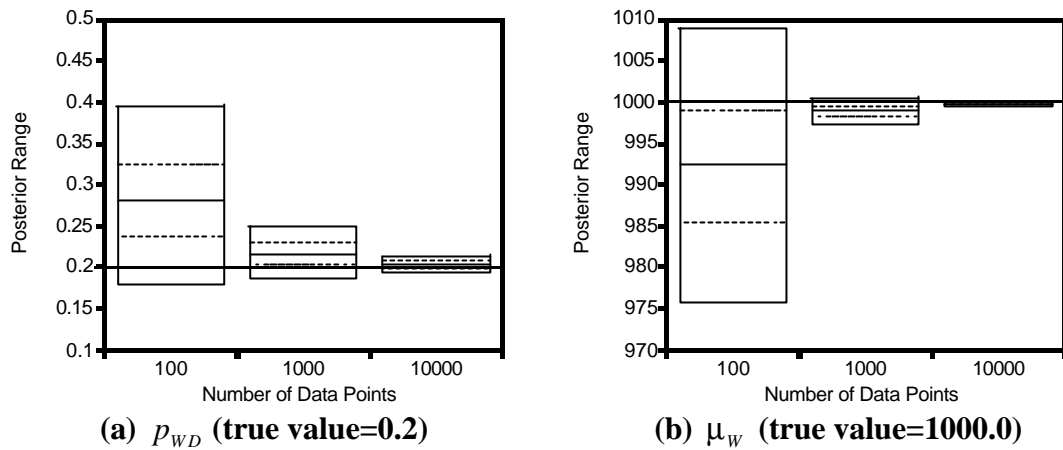
Figure F.2 – Posteriors for selected HSM model parameters for synthetic series set S1 with varying number of data points. Dark line indicates true parameter value.

(a) p_{WD} **(true value=0.1)**

(b) μ_W **(true value=1200.0)**
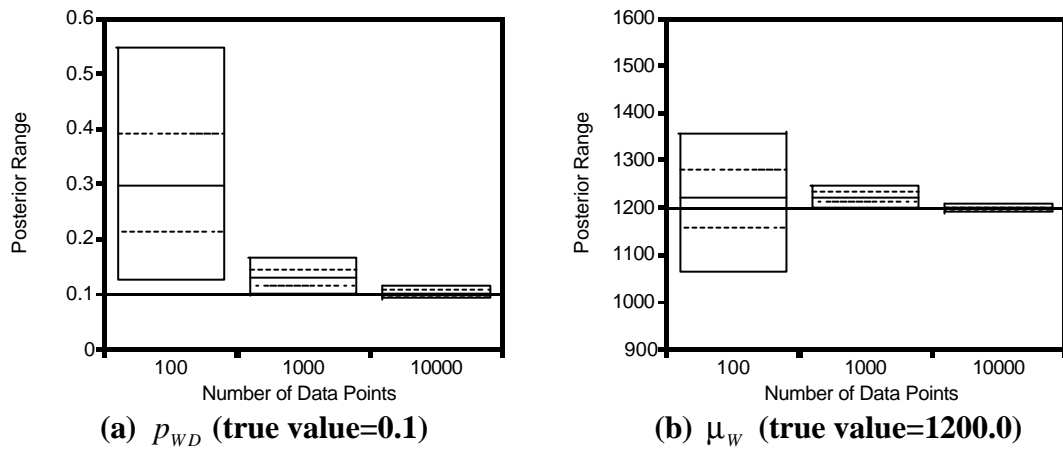
Figure F.3 – Posteriors for selected HSM model parameters for synthetic series set S2 with varying number of data points. Dark line indicates true parameter value.

F.3 Conclusion

The results indicated that as expected the posteriors shrunk towards the true parameter values as the number of data points in the synthetic series was increased. These synthetic calibration runs verify that the application of the Gibbs sampler to the HSM model has been correctly formulated and the computer code used to generate these results is working correctly.

Appendix G - Comparison of Priors to Posteriors for Single Site HSM Model

G.1 Introduction

In the calibration procedure for the single site HSM model informative prior distributions were used for the state rainfall parameters. The prior parameter values were chosen to ensure the priors were as diffuse as possible. An important part of post-calibration analysis is to ensure these priors were actually diffuse compared to the posteriors. In this Appendix the priors are compared to the posteriors to verify this for each of the sites.

G.2   Results

For each of data sets used to calibrate the single site HSM model the prior and posteriors for the state rainfall parameters are compared in Figure G.1 to Figure G.16. As the same prior was used for both the wet and dry state only one prior is shown in the diagrams, while both the posteriors for the wet and dry state means and standard deviations are shown. These plots confirm that, as intended, the priors are relatively diffuse compared to the posteriors. It is worth commenting that for the state standard deviations for some of the sites, especially Melbourne, Adelaide, Perth, Clarence Town and Dungog, the priors are not as diffuse compared to the results for the other sites. However, as they do not dominate the posteriors they are not considered to have a large influence on the inferences. The priors for the state transition probabilities were not compared because the prior parameter values chosen resulted in a uniform prior [refer to Appendix D for details].
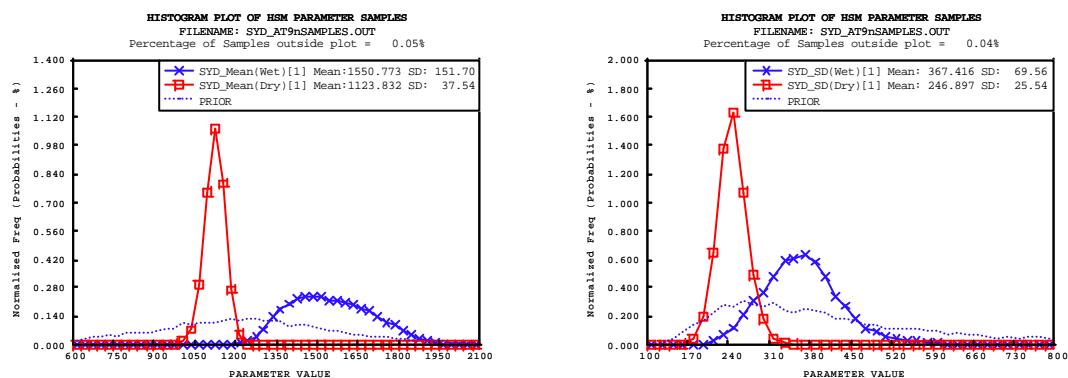
Figure G.1 – Sydney annual (September to August) rainfall – state mean and standard deviation.
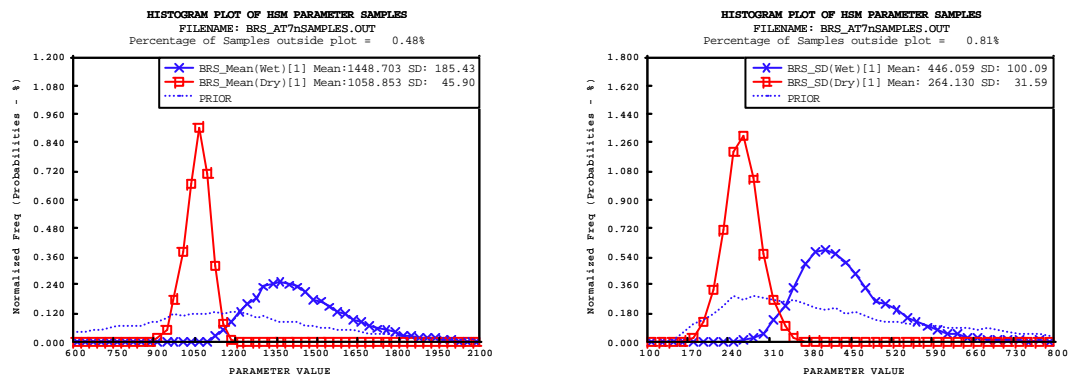
Figure G.2 – Brisbane annual (July to June) rainfall data – state mean and standard deviation.
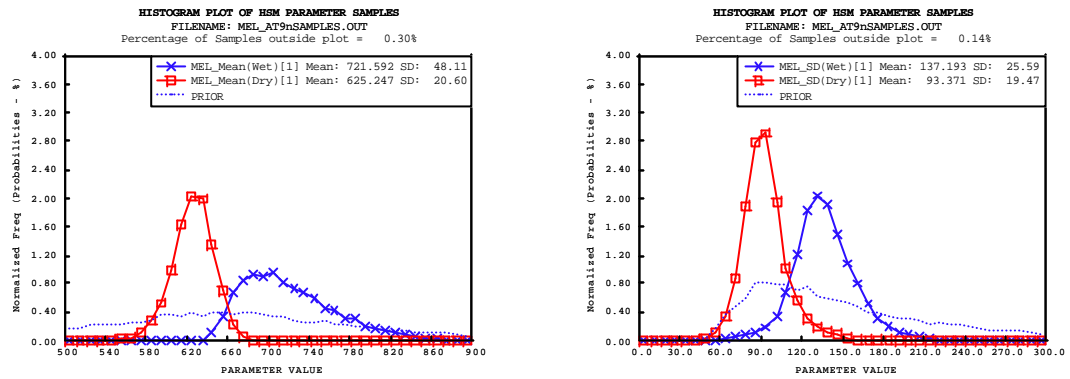
Figure G.3 – Melbourne annual (September to August) rainfall data - state mean and standard deviation.
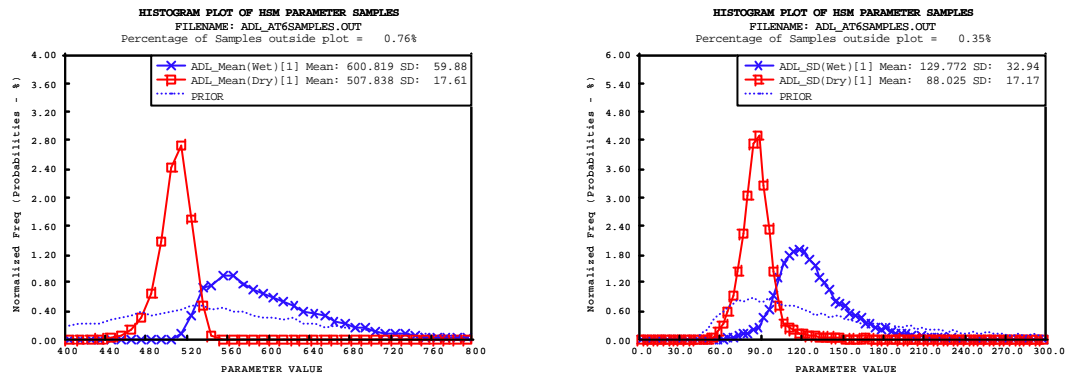
Figure G.4 – Adelaide annual (June to May) rainfall data - state mean and standard deviation.
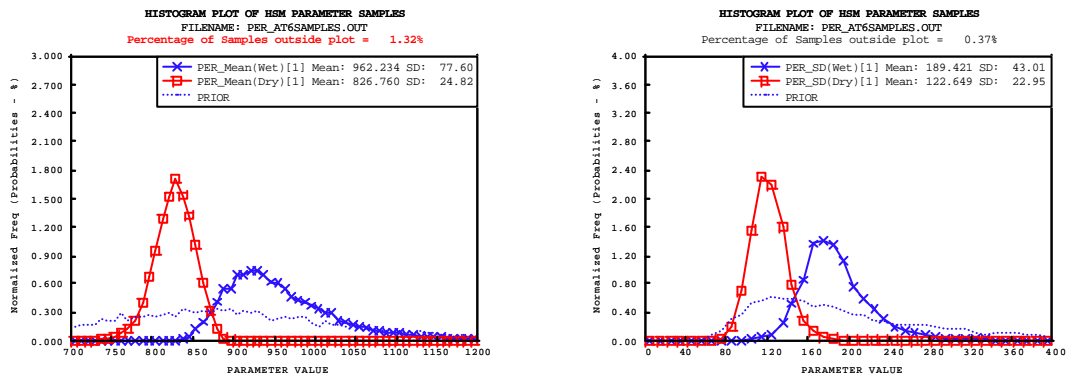
Figure G.5 – Perth annual (June to May) rainfall data - state mean and standard deviation.
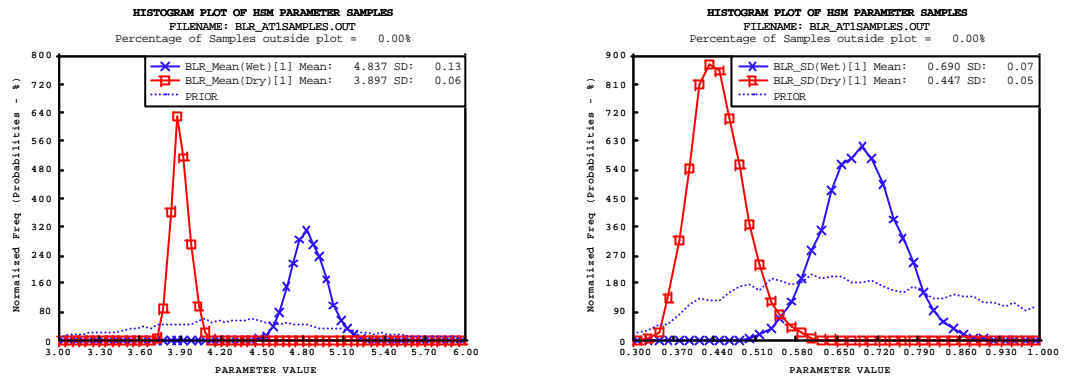
Figure G.6 – Burdekin River reconstructed runoff – transformed state mean and standard deviation.
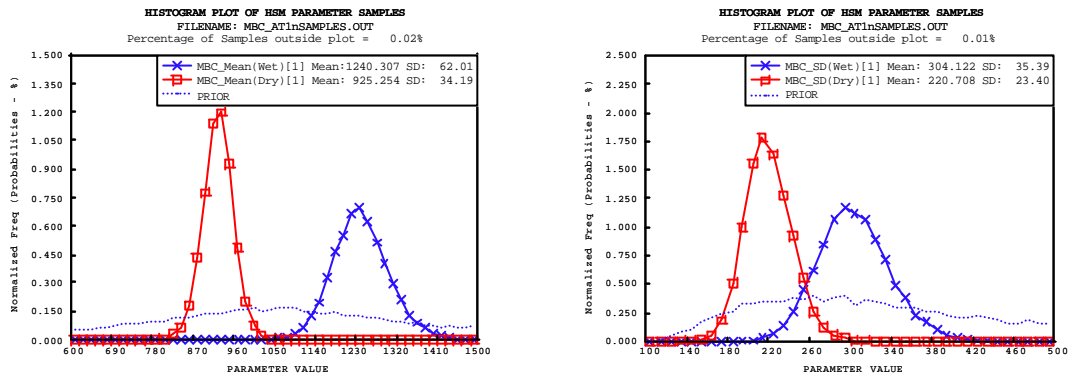
Figure G.7 – Mt. Victoria composite annual (January to December) rainfall data - state mean and standard deviation.
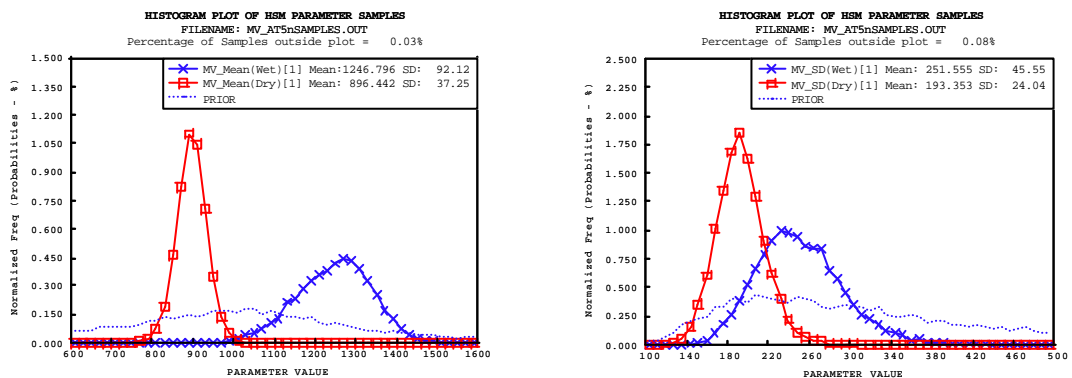
Figure G.8 – Moss Vale annual (May to April) rainfall data - state mean and standard deviation.
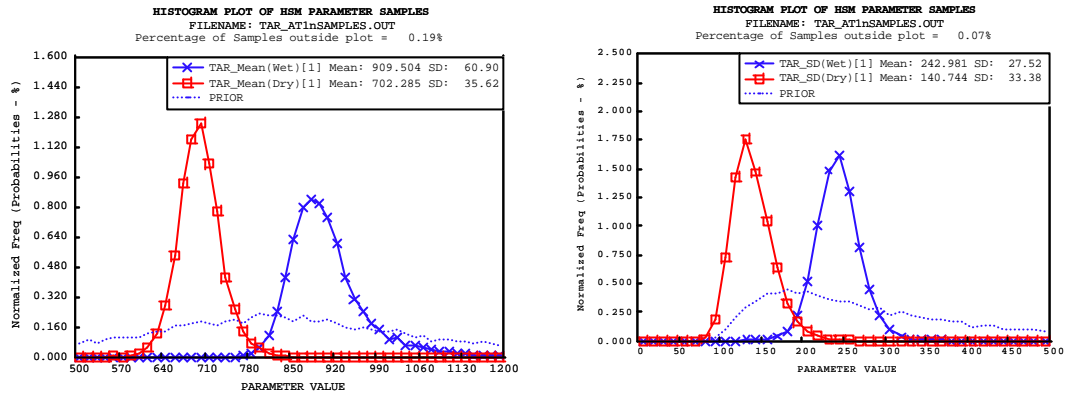
Figure G.9 – Taralga annual (January to December) rainfall data - state mean and standard deviation.
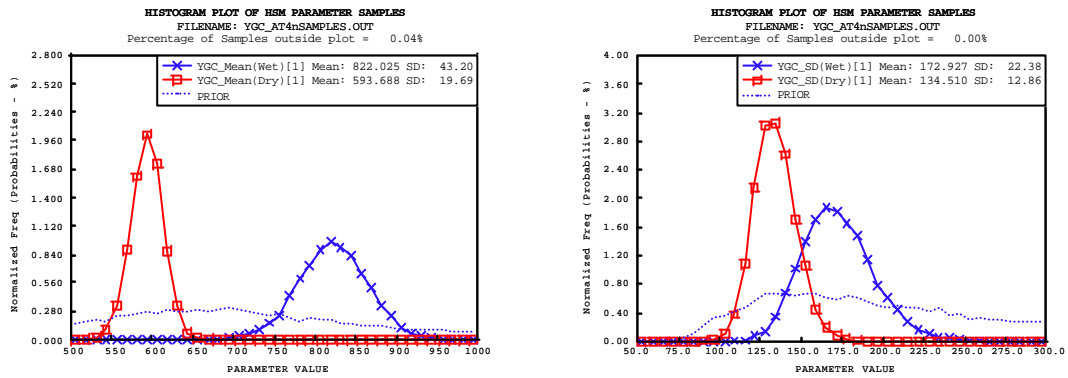
Figure G.10 – Yarra composite annual (April to March) rainfall data - state mean and standard deviation.
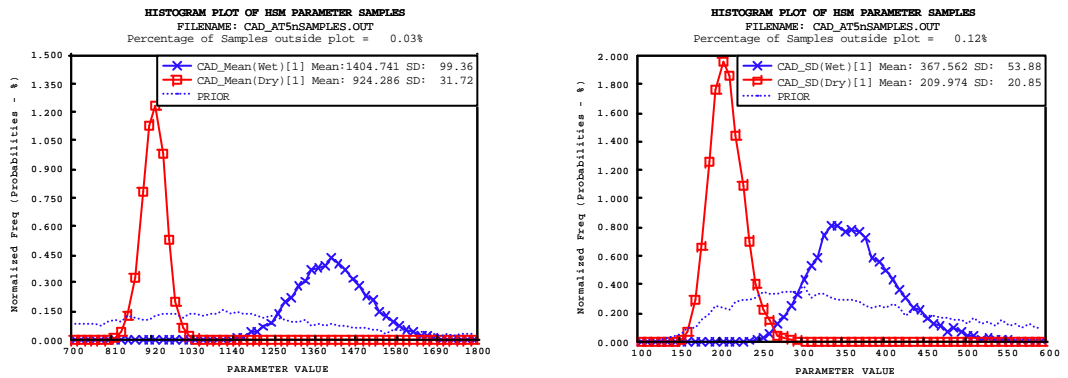
Figure G.11 – Cataract Dam annual (May to April) rainfall data - state mean and standard deviation.
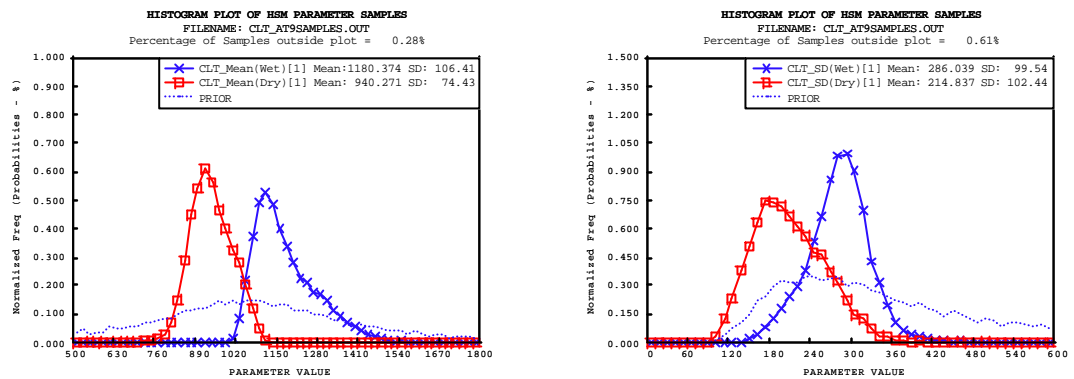
Figure G.12 – Clarence Town annual (September to August) rainfall data - state mean and standard deviation.
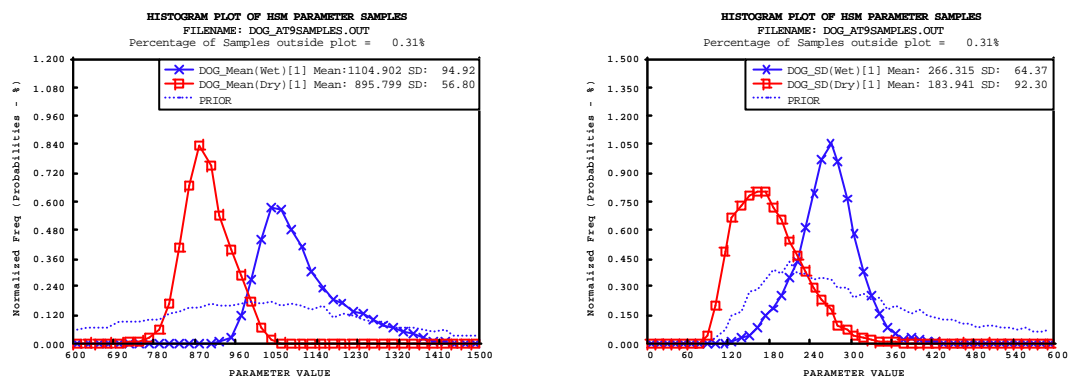
Figure G.13 – Dungog annual (September to August) rainfall data - state mean and standard deviation.
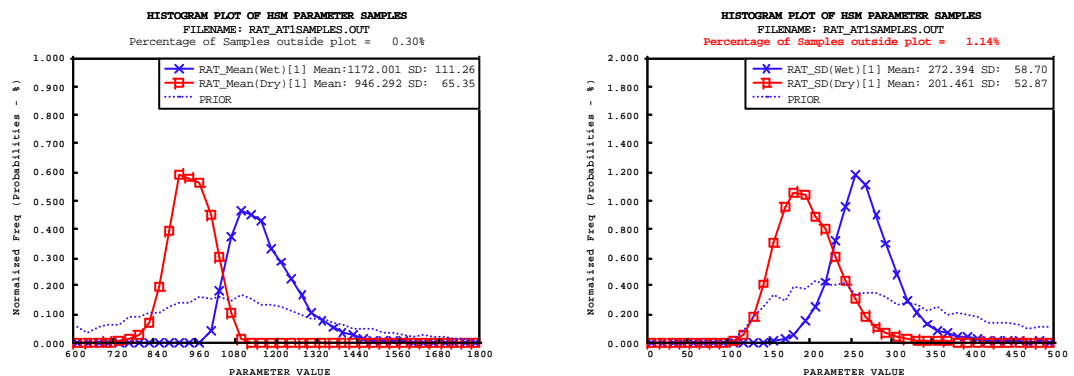
Figure G.14 – Raymond Terrace annual (September to August) rainfall data - state mean and standard deviation.
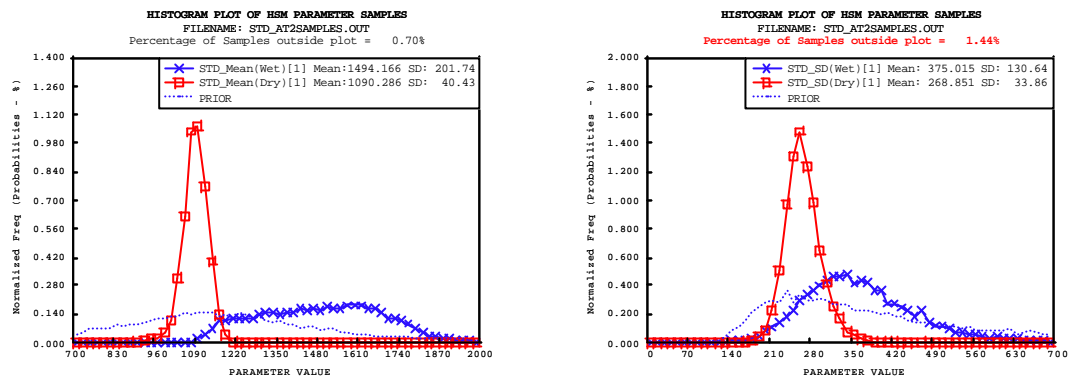
Figure G.15 – Stroud annual (February to January) rainfall data - state mean and standard deviation.
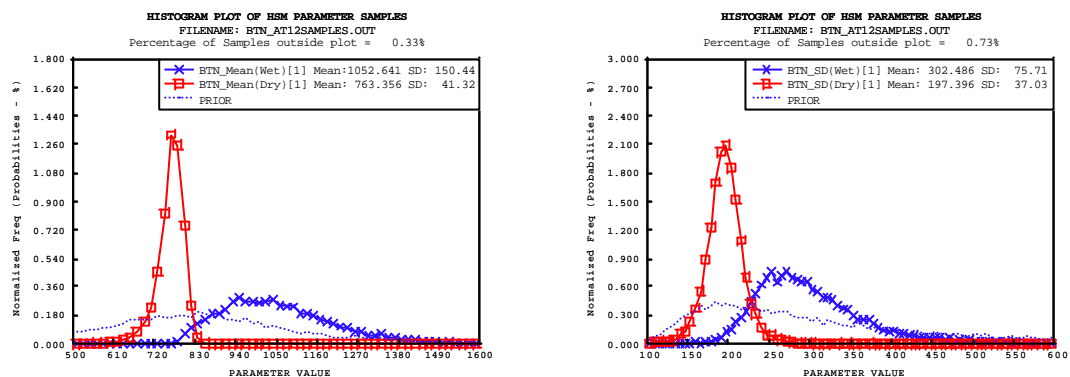
Figure G.16 – Branxton annual (December to January) rainfall data - state mean and standard deviation.

G.3   Conclusion

The results given in this appendix confirm that the informative priors for the state rainfall parameters in the calibration procedure for the single site HSM model were diffuse compared to the posteriors. Hence they exert only a minor influence on the inferences.

Appendix H - Simulated Rainfall Results for Single Site HSM Model

H.1   Introduction

The posterior predictive distribution of the replicated data $p\left(y^{rep}\mid Y_N\right)$, as defined in Section 6.4.1, is simulated for the single site HSM model using the posteriors as given in Chapter 7. Comparison of the sampling distribution of drawing N samples from this posterior predictive distribution to the observed data distribution provides an indication whether the model is a good fit to the observed data. The results presented here are for all the rainfall data sets that are not shown in the main body of this thesis. Inspection of Figure H.1 to Figure H.13 shows that the observed data was within the 5% and 95% confidence limits of the simulated data for all the data sets. This is considered to be a good fit to the data.
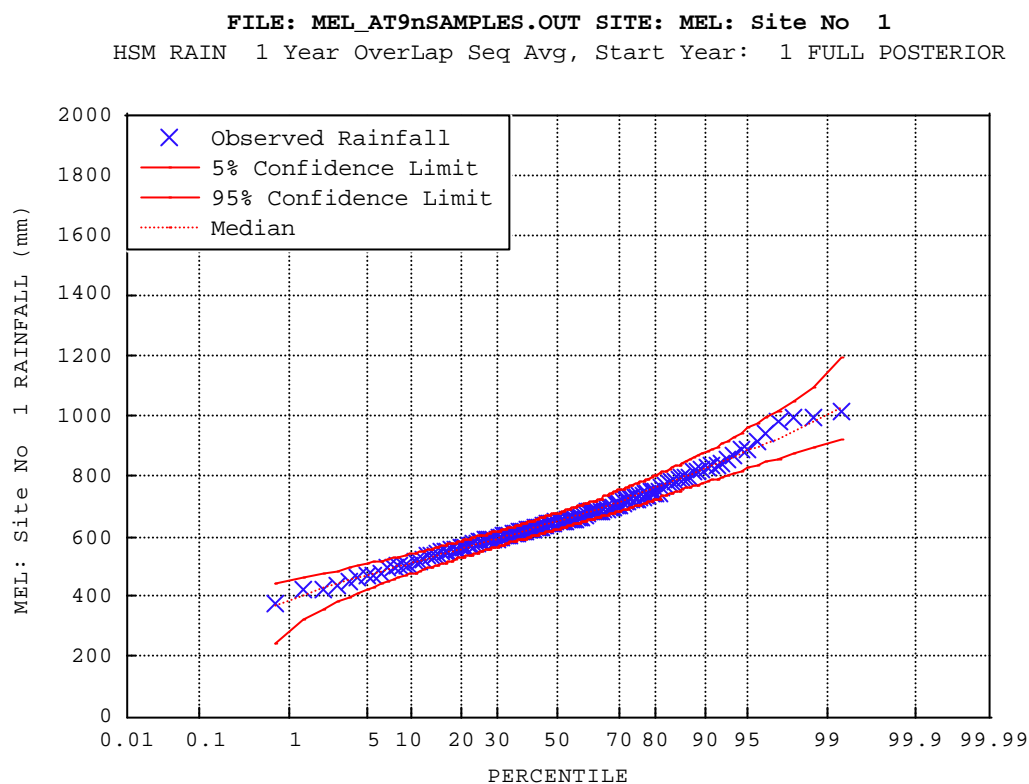
Figure H.1 – Melbourne annual (Sep. to Aug.) rainfall data.
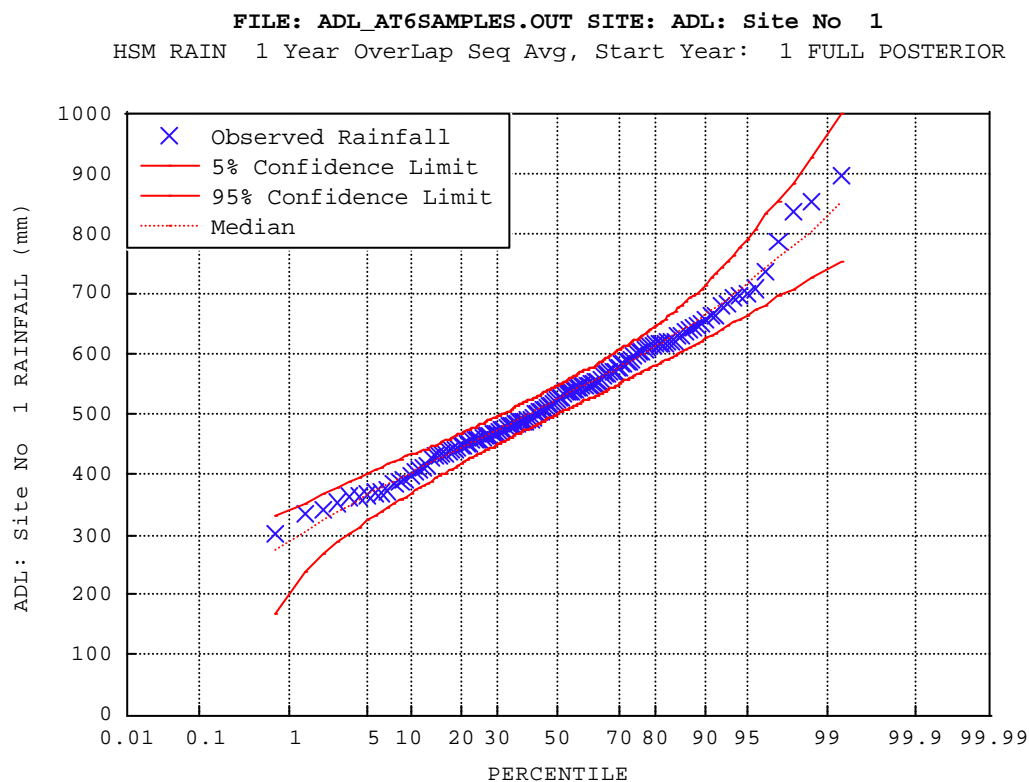
Figure H.2 – Adelaide annual (June to May) rainfall data.
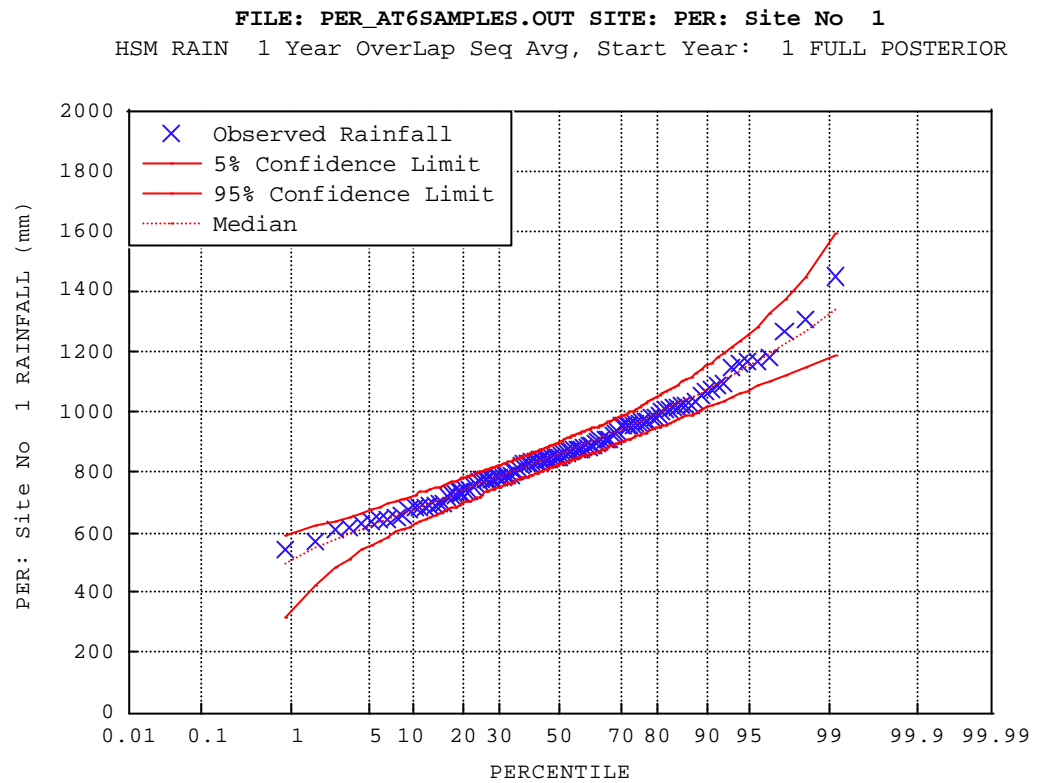
Figure H.3 – Perth annual (June to May) rainfall data.
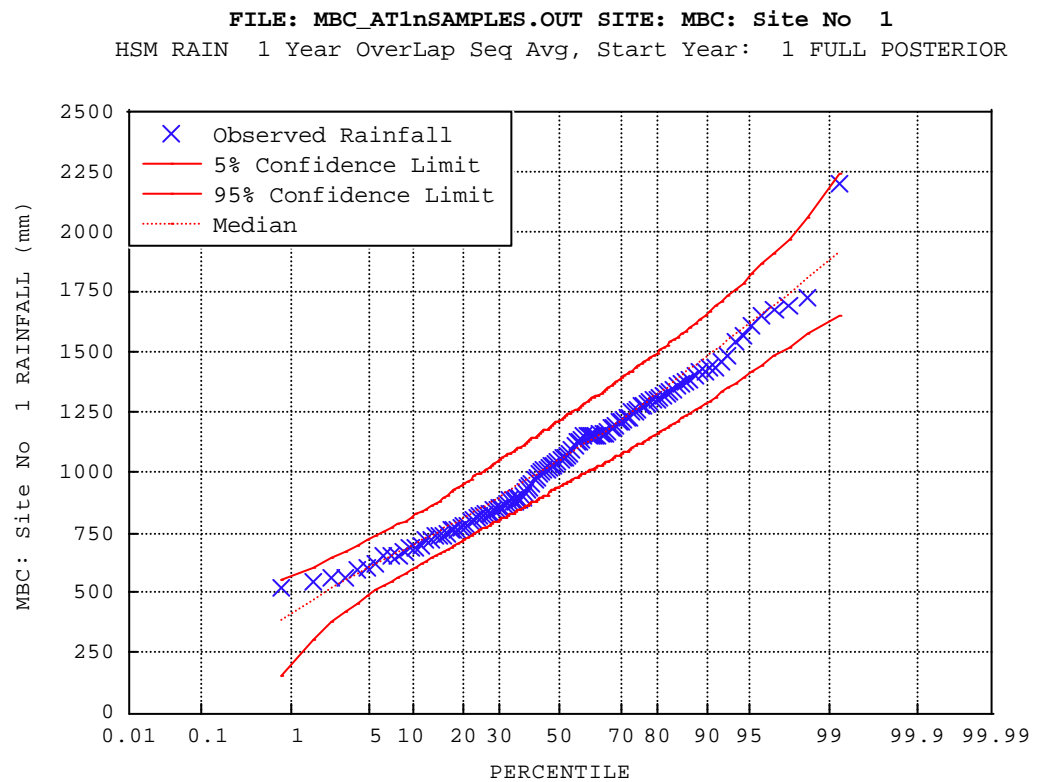
Figure H.4 – Mt. Victoria composite annual (Jan. to Dec.) rainfall data.
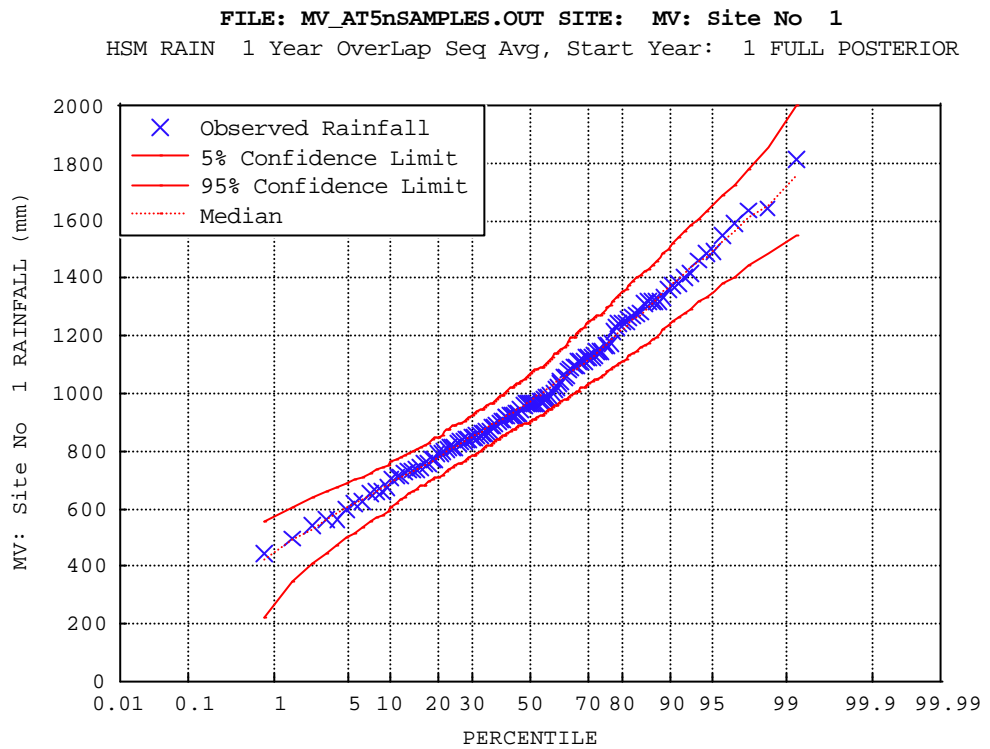
Figure H.5 – Moss Vale annual (May to April) rainfall data.
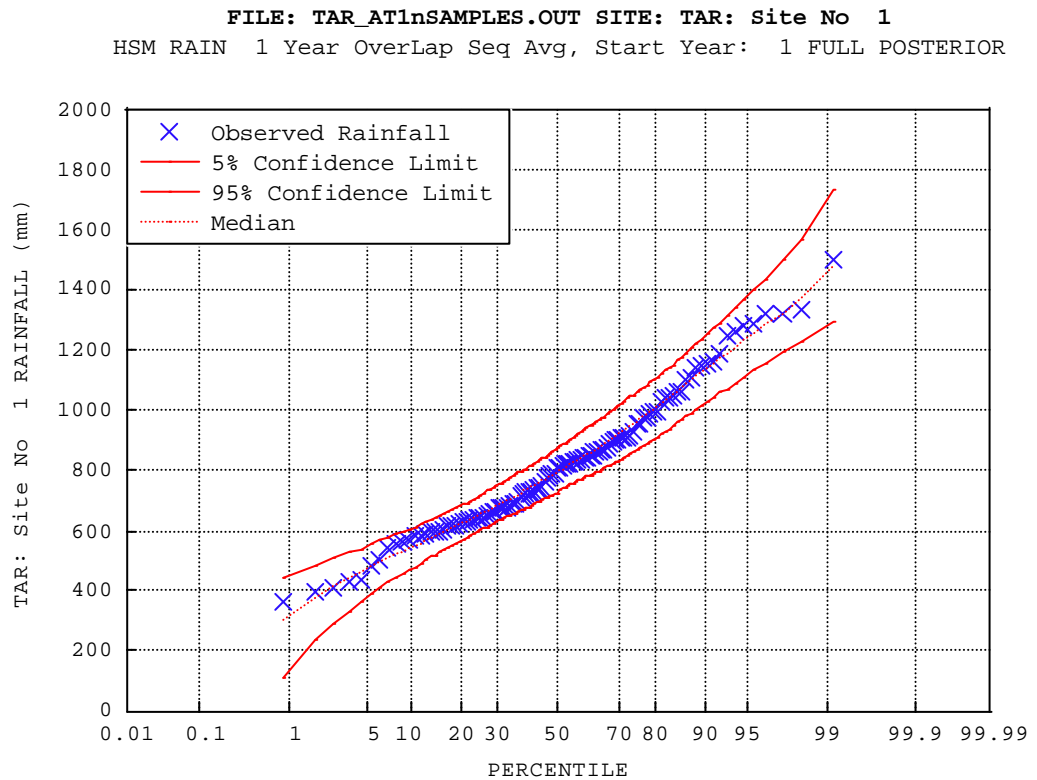
Figure H.6 – Taralga annual (Jan. to Dec.) rainfall data.

FILE: YGC_AT4nSAMPLES.OUT SITE: YGC: Site No 1
HSM RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
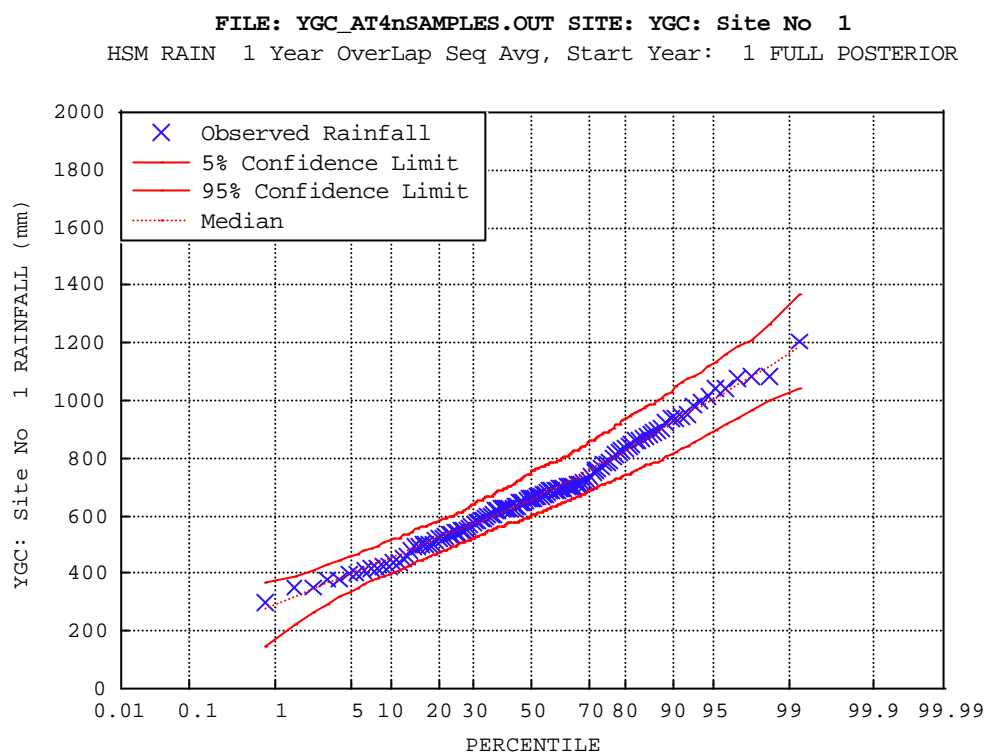
Figure H.7 – Yarra composite annual (April to March) rainfall data.

FILE: CAD_AT5nSAMPLES.OUT SITE: CAD: Site No 1
HSM RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
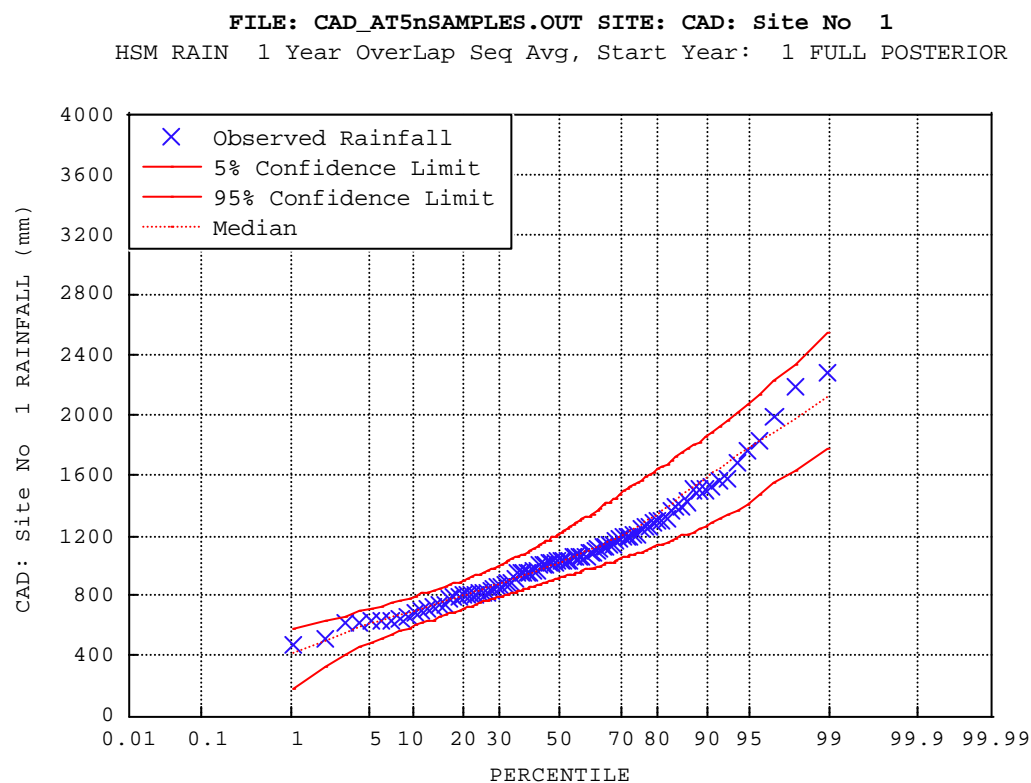
Figure H.8 – Cataract Dam annual (May to April) rainfall data.

FILE: CLT_AT9SAMPLES.OUT SITE: CLT: Site No 1
HSM RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
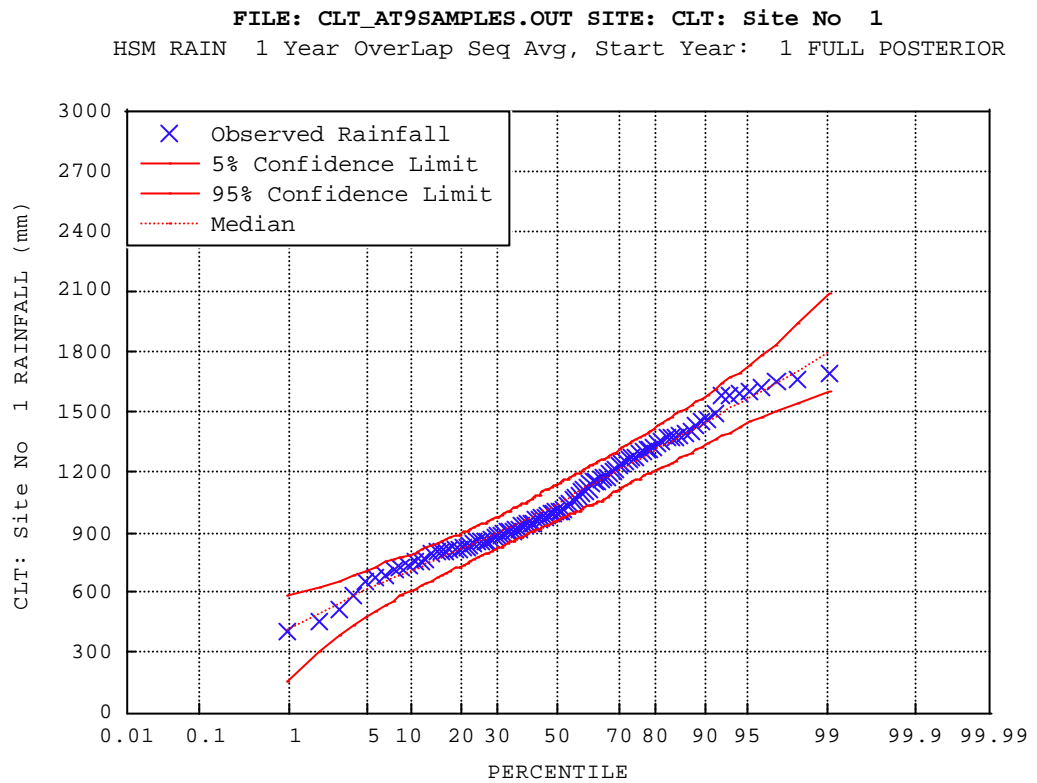
Figure H.9 – Clarence Town annual (Sep. to Aug.) rainfall data.

FILE: DOG_AT9SAMPLES.OUT SITE: DOG: Site No 1
HSM RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
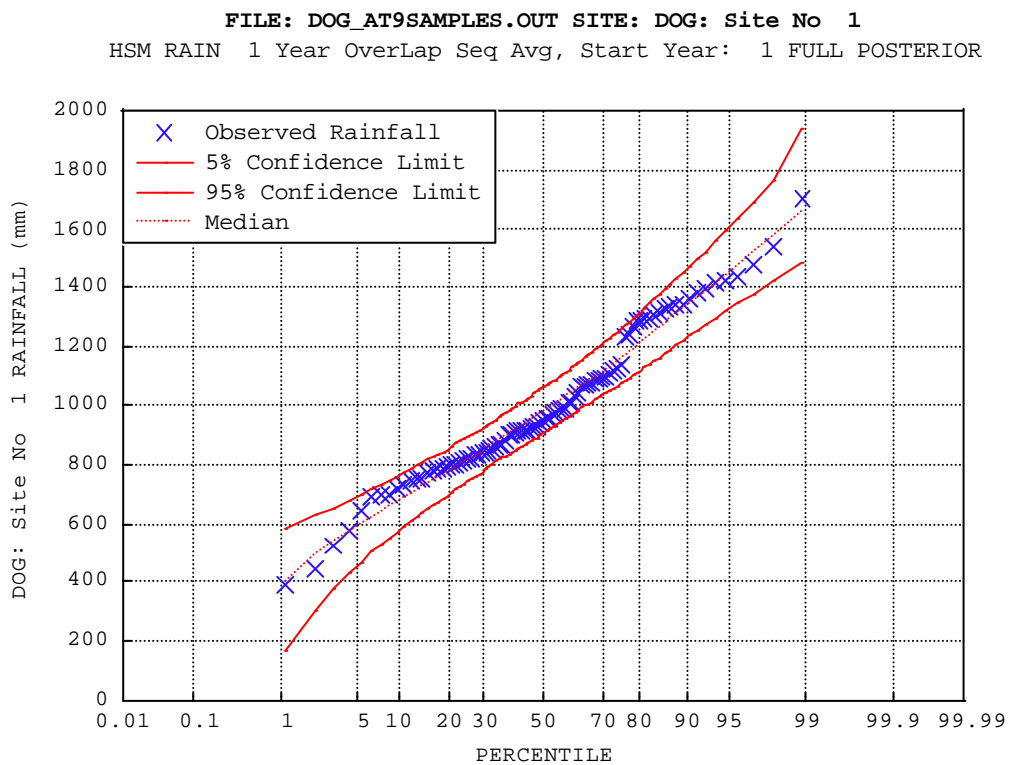
Figure H.10 – Dungog annual (Sep. To Aug.) rainfall data.
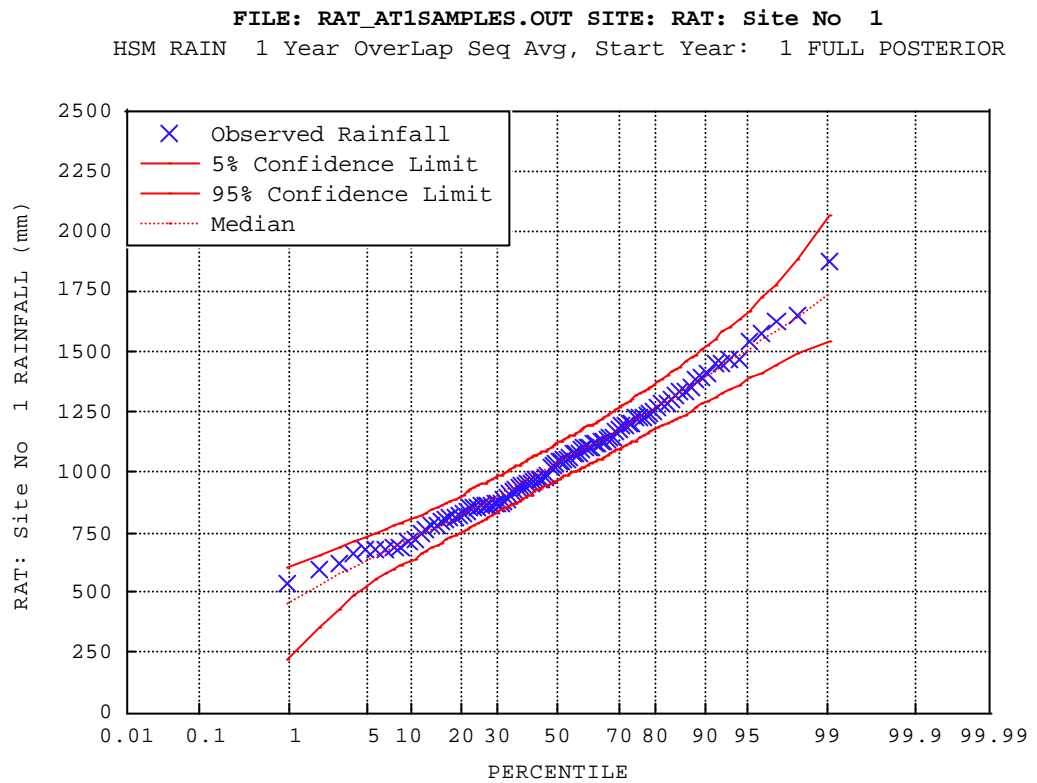
Figure H.11 – Raymond Terrace annual (Jan. to Dec.) rainfall data.
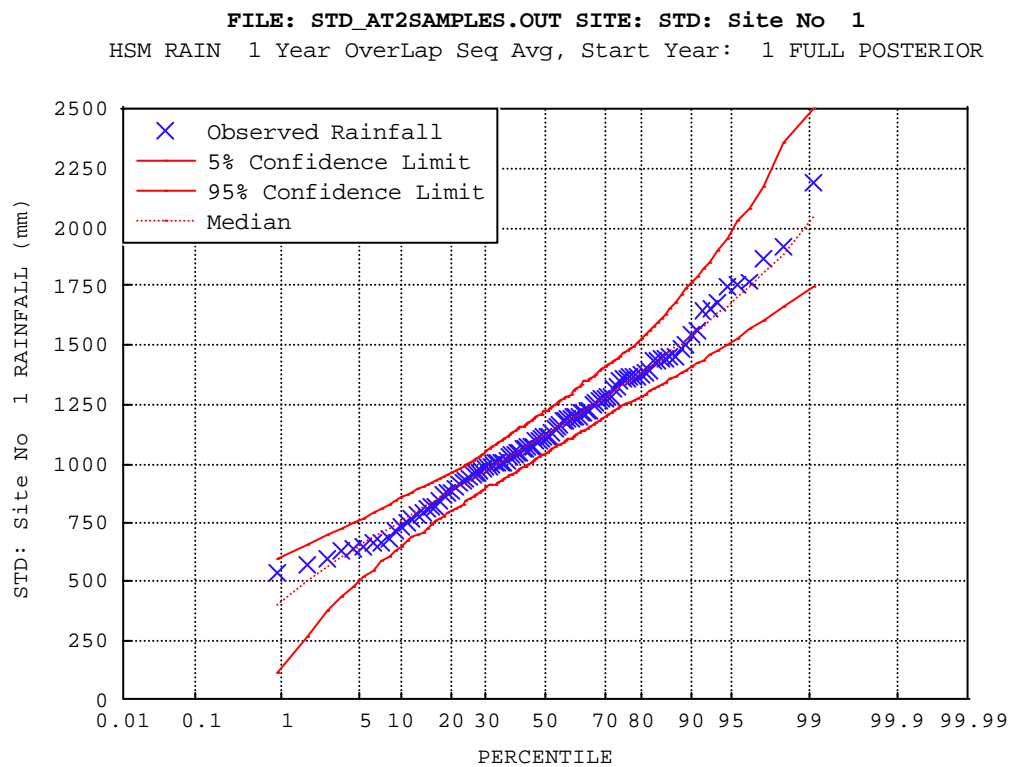
Figure H.12 – Stroud annual (Feb. to Jan.) rainfall data.
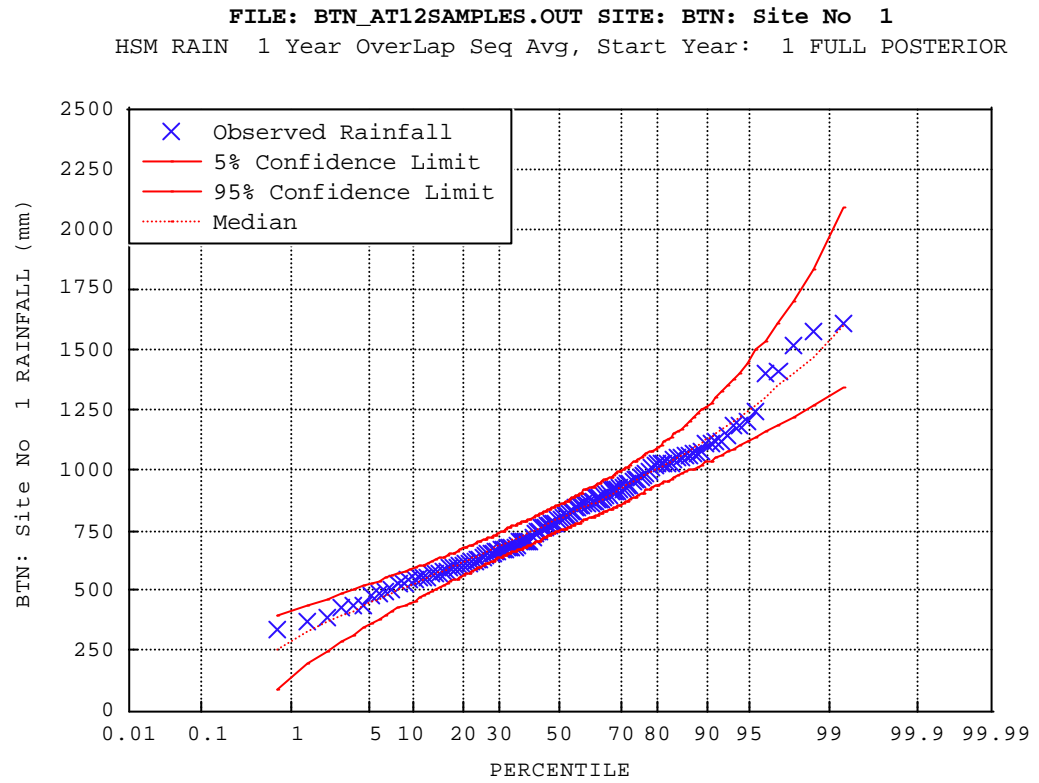
Figure H.13 – Branxton annual (Dec. to Jan.) rainfall data.

Appendix I - Derivation of Likelihood Function for AR(1) Model

I.1   Introduction

In this Appendix the likelihood function for the AR(1) model given a time series of rainfall data Y_N, $p(Y_N|\theta)$ will be derived. This is not straightforward because of the complications caused by the Box-Cox transformation. To motivate this derivation the explantion of the AR(1) modelling framework given in Section 8.2 will be reiterated.

As given in Equation (8.1) the AR(1) model has the form:

$$z_t = \mu + \phi_1(z_{t-1} - \mu) + \varepsilon_t \qquad\qquad (I.1)$$

where z_t is the value of the time series at time step, t, μ is the mean of the time series, ϕ_1 is the lag-one autoregressive parameter ε_t is an uncorrelated Gaussian random variable, with zero mean and variance, σ_ε^2, such that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$.

From Equation (I.1) it can be seen that given the value z_{t-1} the z_t also follows a Gaussian distribution, such that $z_t|z_{t-1} \sim N(\mu + \phi_1(z_{t-1} - \mu), \sigma_\varepsilon^2)$. Hence, to use the

AR(1) model to simulate a rainfall time series $Y_N = \{y_1, \ldots, y_n\}$ it is first necessary to ensure that the rainfall data follow a Gaussian distribution. If this is not the case a Box-Cox transformation [*Box and Cox*, 1964] is commonly applied to the rainfall data such that:

$$z_t = \begin{cases} \dfrac{y_t^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log \ y_t & \lambda = 0 \end{cases} \qquad (I.2)$$

where the transformation parameter λ is usually chosen to ensure the z_t's follow a Gaussian distribution.

In this analysis the transformation parameter is treated as unknown, hence the vector of unknown model parameters for the AR(1) model is:

$$\theta' = (\mu, \sigma_\varepsilon, \phi_1, \lambda) \qquad (I.3)$$

I.2 Derivation of the Likelihood

The derivation of the likelihood function for the AR(1) model parameters for a time series of rainfall data $p(Y_N | \theta)$ will first be derived in terms of the transformed data $p(Z_N | \theta)$.

To derive the likelihood function for a single tranformed data point z_t it must be realized that there are complications caused by the Box-Cox transformation. If Equation (I.2) is rearranged in terms of z_t, then:

$$y_t = \begin{cases} (z_t \lambda + 1)^{1/\lambda} & \lambda \neq 0 \\ \exp(z_t) & \lambda = 0 \end{cases} \qquad (I.4)$$

This places a constraint on the transformed rainfall values that $z_t \lambda + 1 > 0$. Hence the z_t's actually follow a truncated Gaussian distribution, such that:

$$z_t | z_{t-1} \sim TN\left(\mu + \phi_1(z_{t-1} - \mu), \sigma_\varepsilon^2\right) \qquad (I.5)$$

where $TN(\mu, \sigma^2)$ denotes a truncated Gaussian distribution subject to some constraint, in this case $z_t \lambda + 1 > 0$.

The corresponding probability density of a single observation z_t assumed to follow this trunated Gaussian distribution, is written as:

$$p(z_t | z_{t-1}, \theta) \propto \begin{cases} \dfrac{1}{P_t^\lambda} \sigma_\varepsilon^{-1} \exp\left[-\dfrac{1}{2}\left(\dfrac{z_t - \mu - \phi_1(z_{t-1} - \mu)}{\sigma_\varepsilon}\right)^2\right] & \text{if } z_t\lambda + 1 > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (I.6)$$

where P_t^λ is a normalising probability introduced to compensate for the truncation of the distribution. For a distribution to be a proper probability distribution the integral of its density must sum to 1. When a distribution is truncated the integral will not sum to 1. Therefore P_t^λ represents the cumulative probability of the region of the distribution that is not truncated, such that:

$$P_t^\lambda = \begin{cases} \int_{-1/\lambda}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left[-\frac{1}{2}\left(\frac{x - \mu - \phi_1(z_{t-1} - \mu)}{\sigma_\varepsilon}\right)^2\right] dx & \text{if } \lambda > 0 \\ \int_{-\infty}^{-1/\lambda} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left[-\frac{1}{2}\left(\frac{x - \mu - \phi_1(z_{t-1} - \mu)}{\sigma_\varepsilon}\right)^2\right] dx & \text{if } \lambda < 0 \end{cases} \qquad (I.7)$$

When the probability density is normalized by this P_t^λ factor then the integral will sum to 1 and it becomes a proper probability distribution.

To determine the probability density of a single rainfall data point y_t the following change of variable transformation is applied to the density given in Equation (I.6):

$$\begin{aligned} p(y_t | y_{t-1}, \theta) &= \frac{dz_t}{dy_t} p(z_t | z_{t-1}, \theta) \\ &= y_t^{\lambda-1} p(z_t | z_{t-1}, \theta) \\ &\propto \begin{cases} (y_t)^{\lambda-1}(P_t^\lambda \sigma_\varepsilon)^{-1} \exp\left[-\frac{1}{2}\left(\frac{z_t - \mu - \phi_1(z_{t-1} - \mu)}{\sigma_\varepsilon}\right)^2\right] & \text{if } z_t\lambda + 1 > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \qquad (I.8)$$

Now to calculate the full likelihood function for a time series of rainfall data Y_N the following relationship is used (as adopted by *Chib* [1996] the notation $Y_t = \{y_1, \dots y_t\}$ is used):

$$p(Y_N|\theta) = p(y_n|Y_{n-1},\theta)\,p(Y_{n-1}|\theta)$$
$$= p(y_n|y_{n-1},\theta)\,p(Y_{n-1}|\theta) \tag{I.9}$$

In the first line the conditional probability theorem is applied and in the second the assumed Markovian property of the rainfall data is used. By repeatedly applying these two theorems to the right hand term, it can be seen that a recursive expression will result. The summary of this recursion is:

$$p(Y_N|\theta) = p(y_n|y_{n-1},\theta)\ldots p(y_t|y_{t-1},\theta)\ldots p(y_2|y_1,\theta)\,p(y_1|\theta) \tag{I.10}$$

The probability density of the typical term in this recursion $p(y_t|y_{t-1},\theta)$ is given in Equation (I.8).

When all the probability densities for each of the terms given in (I.9) are multiplied together the following expression for the full likelihood results:

$$p(Y_N|\theta) \propto \begin{cases} \left\{ \prod_{t=2}^{n}\left(\dfrac{y_t^{\lambda-1}}{P_t^{\lambda}\sigma_\varepsilon}\right)\right\}\exp\left[-\tfrac{1}{2}\sum_{t=2}^{n}\left(\dfrac{z_t-\mu-\phi_1(z_{t-1}-\mu)}{\sigma_\varepsilon}\right)^2\right]p(y_1|\theta) & z_t\lambda+1>0\ \forall\,t=2,n \\ 0 & otherwise \end{cases} \tag{I.11}$$

where P_t^{λ} and z_t are as given above. It is important to note that the normalizing factor P_t^{λ} changes for each data point z_t because it is dependent on $\mu+\phi_1(z_{t-1}-\mu)$.

In general terms, the likelihood for the terminal point $p(y_1|\theta)$ could be calculated based on the marginal density:

$$p(y_1|\theta) = \int p(y_1|y_0,\theta)p(y_0|\theta)\,dy_0 \tag{I.12}$$

However, this marginal density is not easily derived. Instead, in this analysis the following expression is used:

$$p(y_1|\theta) = \frac{dz_1}{dy_1}\,p(z_1|\theta)$$
$$= y_1^{\lambda-1}\,p(z_1|\theta) \tag{I.13}$$

which applies the change of variable used in Equation (I.8) to $p(y_1|\theta)$. The density for the terminal point in transformed space $p(z_1|\theta)$ can be calculated because if the

truncation is ignored then the marginal density for z_t, as given by *Box and Jenkins* [1970], is:

$$p(z_t|\theta) \sim N\left(\mu, \frac{\sigma_\varepsilon^2}{1-\phi_1^2}\right)$$

(I.14)

Ignoring the truncation for only the terminal point is not expected to have a major impact on the inferences. Hence, this method was used to calculate the likelihood function for the terminal point, $p(y_1|\theta)$.

I.3 Relationship Between Parameters in Transformed and Untransformed Space

Using first-order approximations it is possible to derive a relationship between the parameters of the AR(1) model, μ and σ_ε which are in transformed space to their equivalents in untransformed space, μ_y and σ_y. This derivation begins by using the knowledge that $z_t = f(y_t)$, as defined in Equation (I.2). Applying a Taylor series expansion to this function gives:

$$z_t = f(\mu_y) + \frac{dz_t}{dy_t}\bigg|_{z_t=\mu_y}(y_t - \mu_y)$$

$$E[z_t] = E[f(\mu_y)] + \frac{dz_t}{dy_t}\bigg|_{z_t=\mu_y} E[(y_t - \mu_y)]$$

$$E[z_t] = E[f(\mu_y)]$$

$$\mu = \frac{\mu_y^\lambda - 1}{\lambda}$$

(I.15)

where in the second line expectations are taken, in the third line because $E[(y_t - \mu_y)] = E[y_t] - E[\mu_y] = 0$ that term drops out and the fourth line follows because of the relationship given in Equation (I.2). Hence, Equation (I.15) gives an expression that relates the mean in transformed space μ to its equivalent first order approximation in untransformed space μ_y.

To derive a similar expression for σ_ε the Taylor series expansion is also utilised, such that:

$$z_t - f(\mu_y) = \left.\frac{dz_t}{dy_t}\right|_{z_t = \mu_y} (y_t - \mu_y)$$

$$E\left[(z_t - f(\mu_y))^2\right] = \left(\left.\frac{dz_t}{dy_t}\right|_{z_t = \mu_y}\right)^2 E\left[(y_t - \mu_y)^2\right]$$

$$Var(z_t) = \left(\mu_y^{(\lambda-1)}\right)^2 Var(y_t)$$

$$\sigma_z^2 = \mu_y^{2(\lambda-1)} \sigma_y^2 \qquad\qquad (I.16)$$

where expectations are taken to produce the second line and the third line again uses the relationship given in Equation (I.2). Using the knowledge that $\sigma_z^2 = \dfrac{\sigma_\varepsilon^2}{1 - \phi_1^2}$ [*Box and Jenkins,* 1970] then Equation (I.16) can rearranged to give an expression in terms of σ_ε, where:

$$\sigma_\varepsilon^2 = \sigma_z^2 \left(1 - \phi_1^2\right)$$

$$= \mu_y^{2(\lambda-1)} \sigma_y^2 \left(1 - \phi_1^2\right)$$

$$\sigma_\varepsilon = \mu_y^{(\lambda-1)} \sigma_y \sqrt{\left(1 - \phi_1^2\right)} \qquad\qquad (I.17)$$

Hence an expression that relates the standard deviation in transformed space σ_ε to its equivalent first order approximation in untransformed space σ_y again results.

For an explanation of why this derivation was undertaken refer to Section 8.3.2.

Appendix J - Verification of AR(1) Model Calibration Procedure

J.1   Introduction

To verify that the Metropolis algorithm and the likelihood function for the AR(1) model had been correctly formulated and the computer code used was free of any errors synthetic calibration runs were used. Initially in the first batch of synthetic calibration runs synthetic data generated from a bivariate Gaussian distribution was used to verify that the Metropolis algorithm had been correctly coded. Because the jump distribution used in the Metropolis algorithm was multivariate Gaussian, if the Metropolis algorithm has been correctly formulated it should have no trouble recovering the true synthetic parameter values for bivariate Gaussian data. The second batch of calibration runs undertaken was to ensure that the likelihood function for AR(1) model was formulated correctly. Synthetic data was generated using the AR(1) modelling structure with a Box-Cox transformation as given in Section 8.2. The posteriors for both sets were examined for time series of length 100, 1000 and 10,000 data points to verify that they converged to the true parameter values as the number of data points increased.

J.2 Results

J.2.1 Bivariate Gaussian synthetic data

To generate the synthetic bivariate Gaussian data the following parameter values were used:

$$N_2 \left(\begin{bmatrix} 10.0 \\ 10.0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix} \right) \tag{J.1}$$

It can be seen that both components have the same parameter values. Therefore the posteriors for the bivariate Gaussian synthetic data are shown in Figure J.1 only for one component and the correlation. Percentile plots were again used [refer Appendix F] to compare the posteriors for varying length data series. It can be clearly seen that as the number of data points increases the posteriors converge towards the true parameter values. In addition the acceptance rate was within the optimal range [as given in Section 8.3.1.a]. These results verified that the Metropolis algorithm has been correctly formulated.

(a) Mean (true value =10.0) (b) Stan. Dev. (true value = 1.0)

(c) Correlation (true value=0.5)
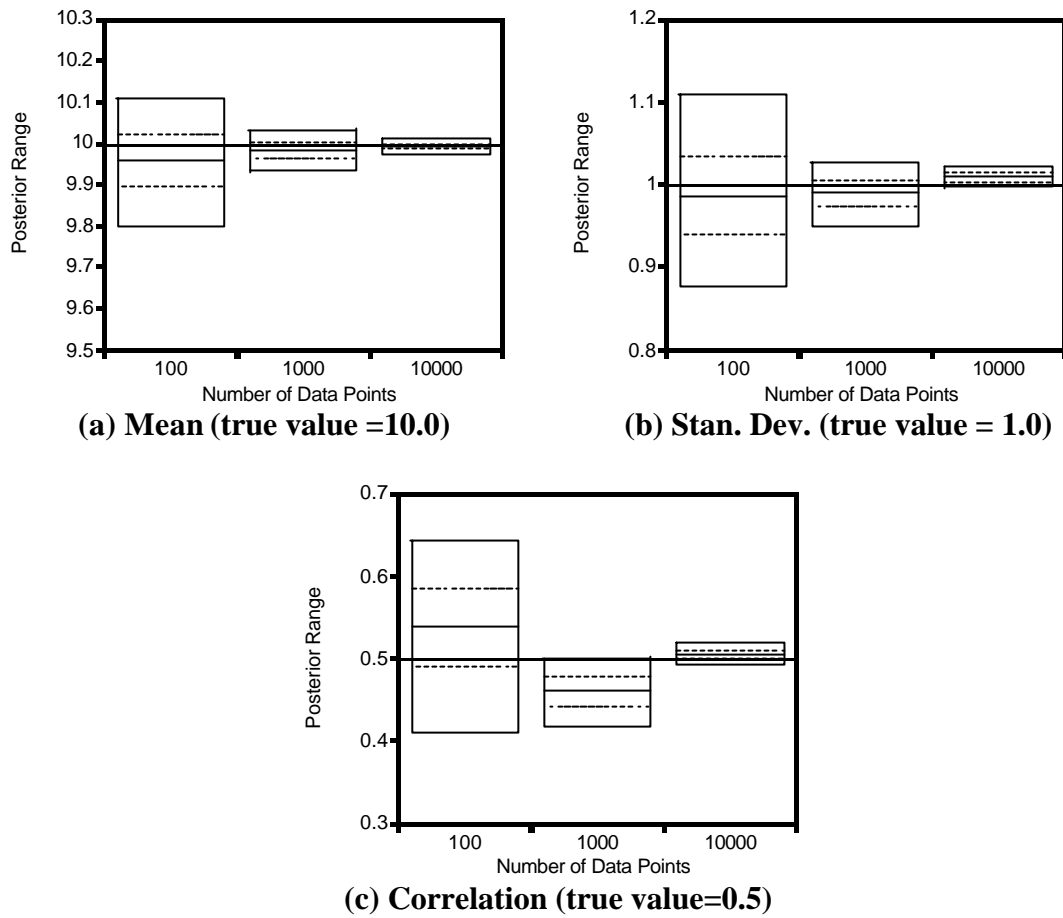
Figure J.1 – Posteriors for the parameters of one component of the bivariate Gaussian model for a varying number of data points. Dark line indicates true parameter value.

J.2.2   AR(1) synthetic data

To test that the likelihood function for the AR(1) model with Box-Cox transformation had been formulated correctly synthetic data was generated using two different parameter sets, denoted as the S3 and S4 set. Their values are summarized in Table J.1. The procedure followed was to generate synthetic z_t values using an AR(1) model and then to transform it with a Box-Cox transformation. The Metropolis algorithm was then applied to determine if the true synthetic values of the AR(1) modelling parameters and the transformation parameter could be recovered. During the generation of the synthetic data it was checked whether any values generated by the AR(1) model violated $z_t \lambda + 1 > 0$ prior to applying the transformation. If they did, they were resampled as this is a constraint of this modelling structure - refer to Appendix I for more details. This causes a truncation in the distribution of transformed values. The likelihood function

was derived to allow for this truncation. The aim of these synthetic calibration runs is to verify that this likelihood function is correct.

Table J.1 – Synthetic Parameter Values for AR(1) model with Box-Cox transformation.

Parameter Set	Mean μ	Stan. dev. σ_ε	Lag-one autoregressive coefficient ϕ_1	Transformation parameter λ
S3	10.0	3.0	0.5	0.5
S4	3.0	4.0	0.5	0.5

It is possible to calculate the proportion of the transformed data distribution that is truncated. Using the fact that the mean of the z_t values is simply the AR(1) mean parameter, while the variance σ_z^2 is calculated by $\sigma_z^2 = \sigma_\varepsilon^2 / (1 - \phi_1^2)$ [*Box and Jenkins, 1970*] the proportion that violates the constraint $z_t \lambda + 1 > 0$ can be calculated. For the S3 set of parameter values this proporation is very close to 0, while for set S4 it is closer to 15%. Hence, set S3 verifies that the AR(1) model likelihood function is correct, while set S4 verifies that the AR(1) model likelihood function with the truncation due to the transformation is correct. This is the reason for choosing two different sets of parameter values.

The posteriors for all the parameters are shown in Figure J.2 and Figure J.3 for set S3 and S4 respectively. It can be clearly seen that in all cases the posterios converge towards the true parameter values as the number of data points increased.

(a) Mean (true value = 10.0)

(b) Stan. Dev. (true value = 3.0)

(c) Lambda (true value = 0.5)

(d) Phi (true value = 0.5)
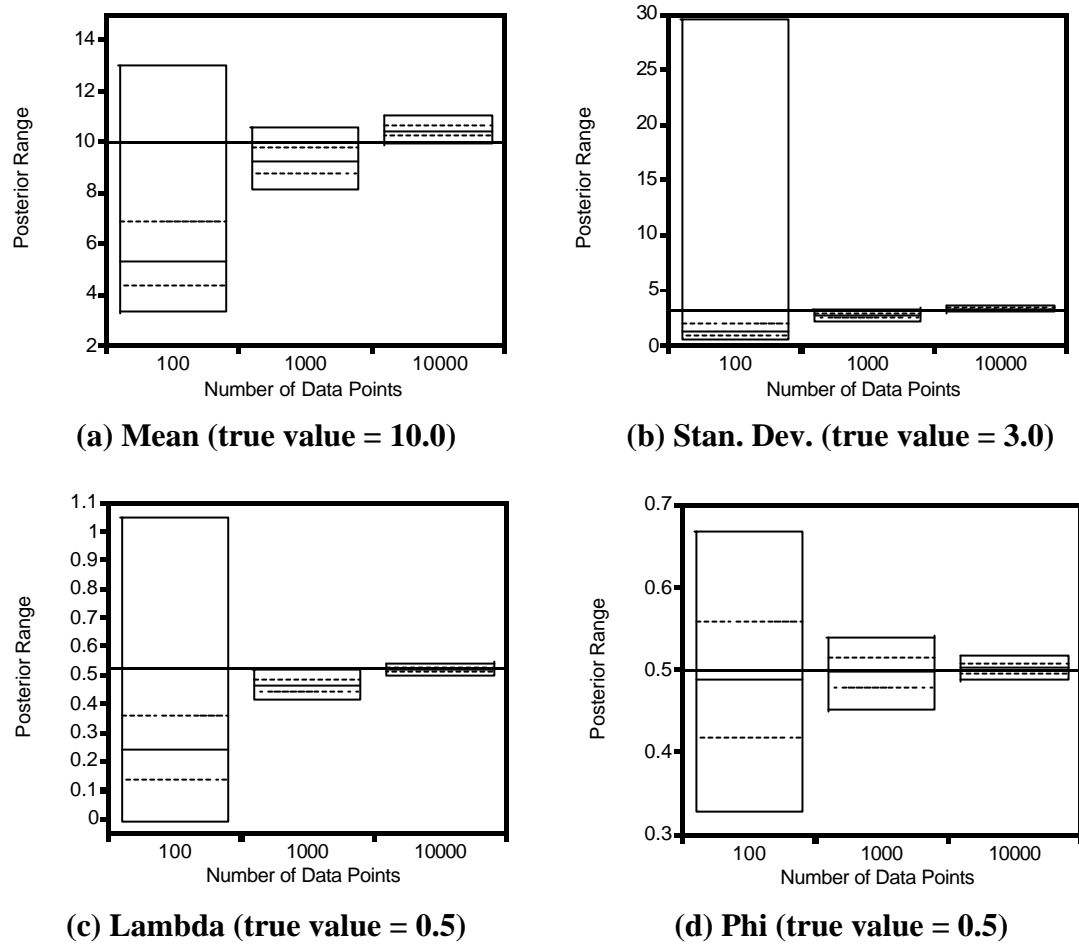
Figure J.2 – Posteriors of the AR(1) model parameters for synthetic parameter set S3 for a varying number of data points. Dark line indicates true parameter value.

(a) Mean (true value = 3.0)

(b) Stan. Dev. (true value = 4.0)

(c) Lambda (true value = 0.5)

(d) Phi (true value = 0.5)
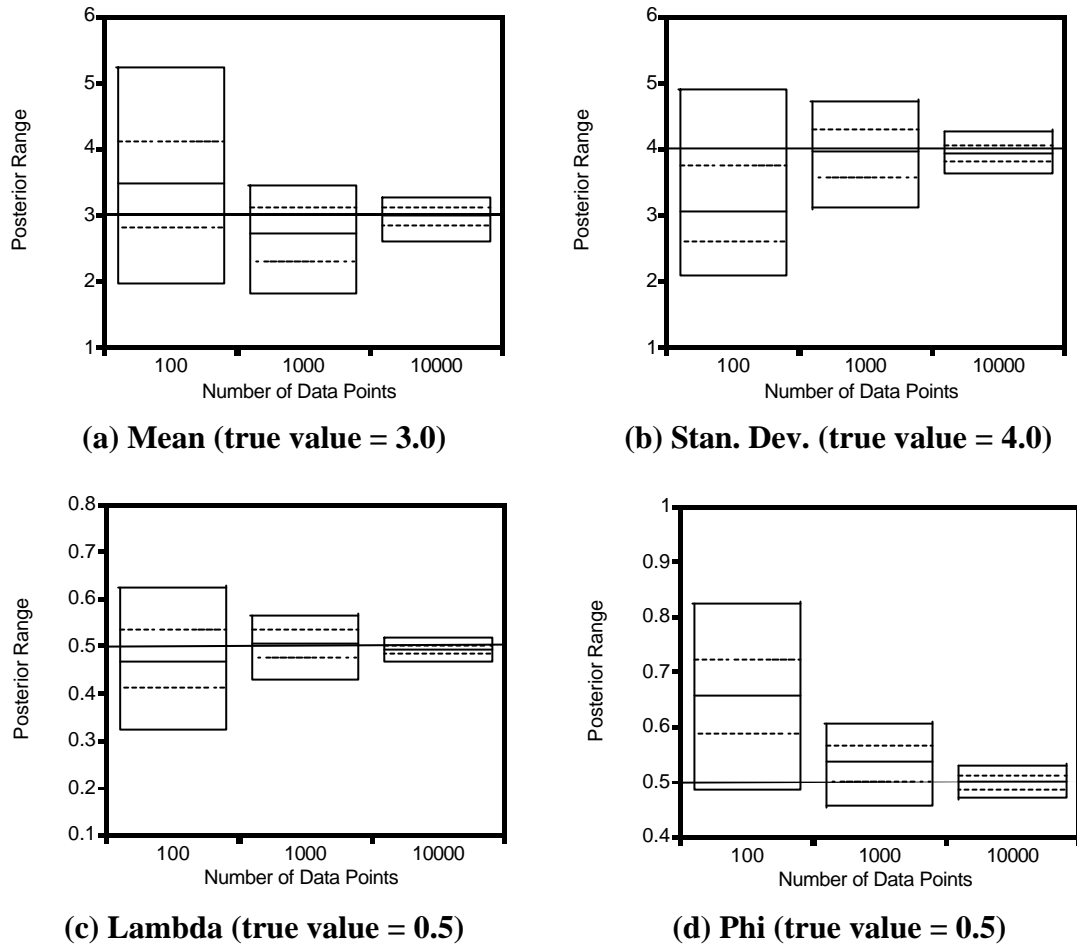
Figure J.3 – Posteriors of the AR(1) model parameters for synthetic parameter set S4 for a varying number of data points. Dark line indicates true parameter value.

J.3   Conclusion

The results indicated that, as expected, the posteriors converged towards the true parameter values as the number of data points in the synthetic series was increased. These synthetic calibration runs verify that the implementation of the Metropolis algorithm and the derivation of the likelihood function for the AR(1) model with a Box-Cox transformation was indeed correct.

Appendix K - Simulated Rainfall Results for AR(1) Model

K.1 Introduction

The posterior predictive distribution of the replicated data $p\left(y^{rep}\mid Y_N\right)$, as defined in Section 6.4.1, is simulated for the AR(1) model using the posteriors as given in Chapter 8. Comparison of the sampling distribution of drawing N samples from this posterior predictive distribution to the observed data distribution provides an indication whether the model is a good fit to the observed data. Inspection of Figure K.1 to Figure K.10 shows that the observed data was within the 5% and 95% confidence limits of the simulated data for all the data sets. This is considered to be a good fit to the data.
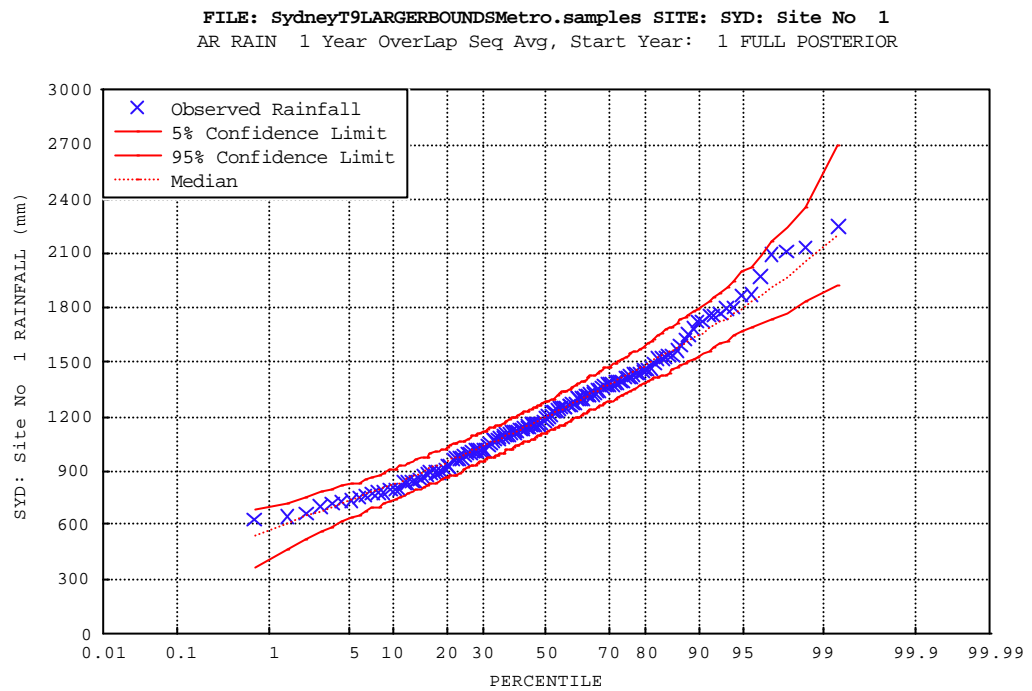
Figure K.1 – Sydney annual (Sep. to Aug.) rainfall data.
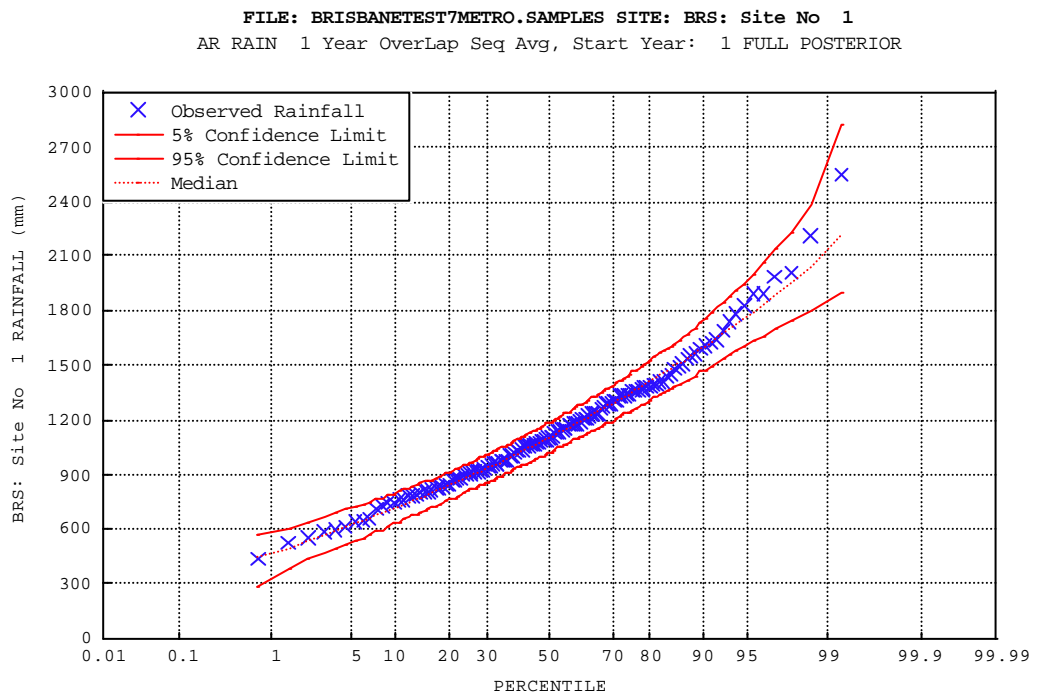
Figure K.2 – Brisbane annual (July to June) rainfall data.
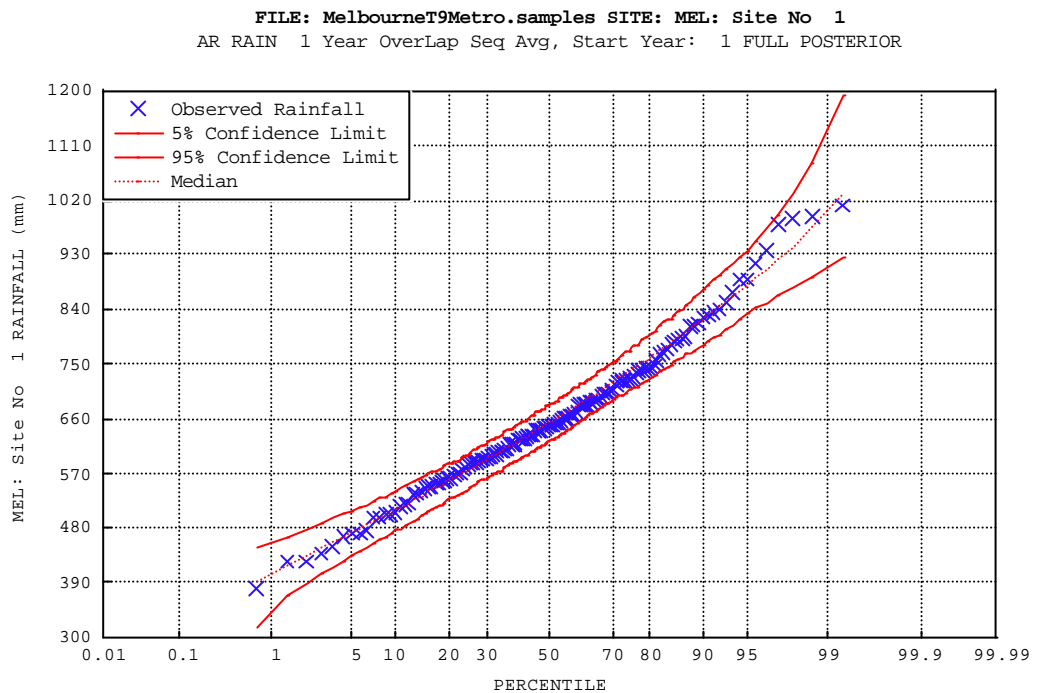
Figure K.3 – Melbourne annual (Sep. to Aug.) rainfall data.
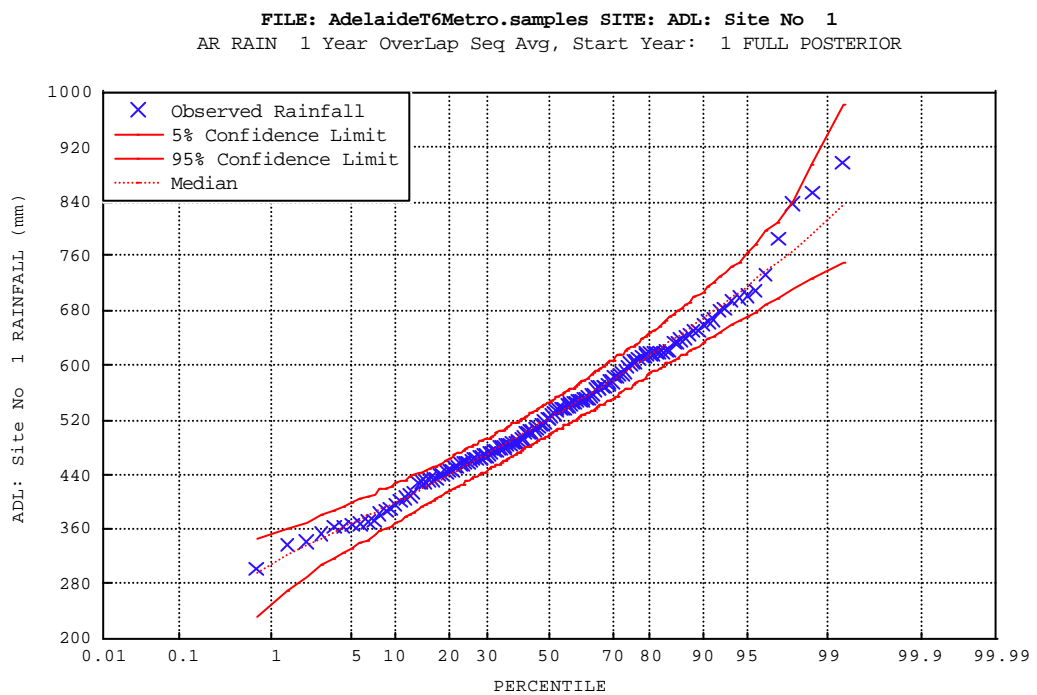
Figure K.4 – Adelaide annual (June to May) rainfall data.
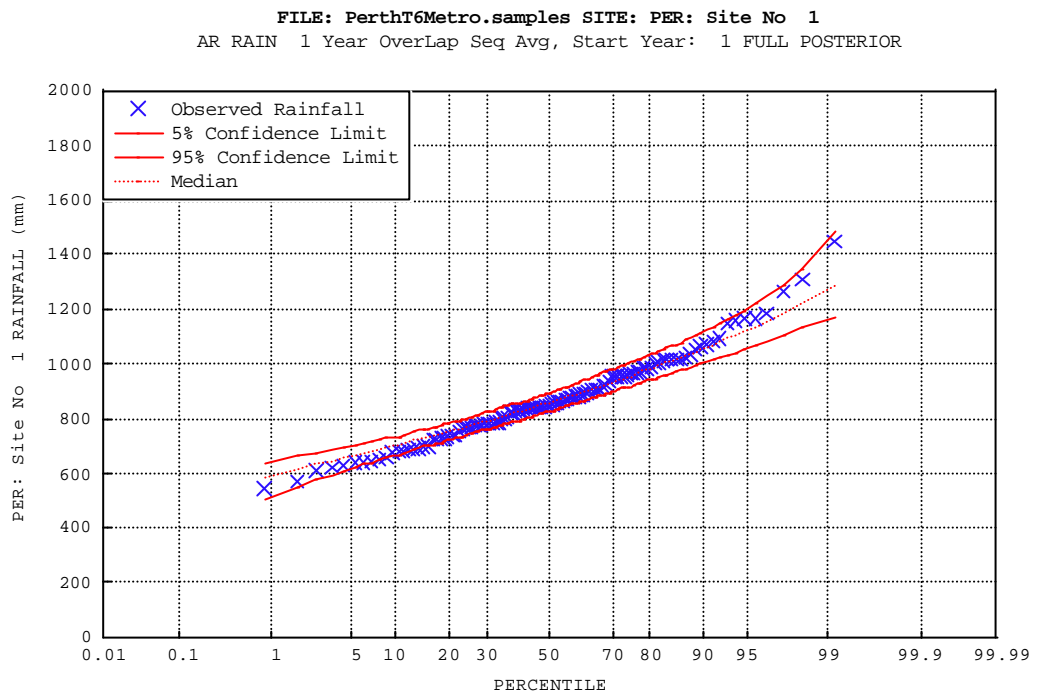
Figure K.5 – Perth annual (June to May) rainfall data.
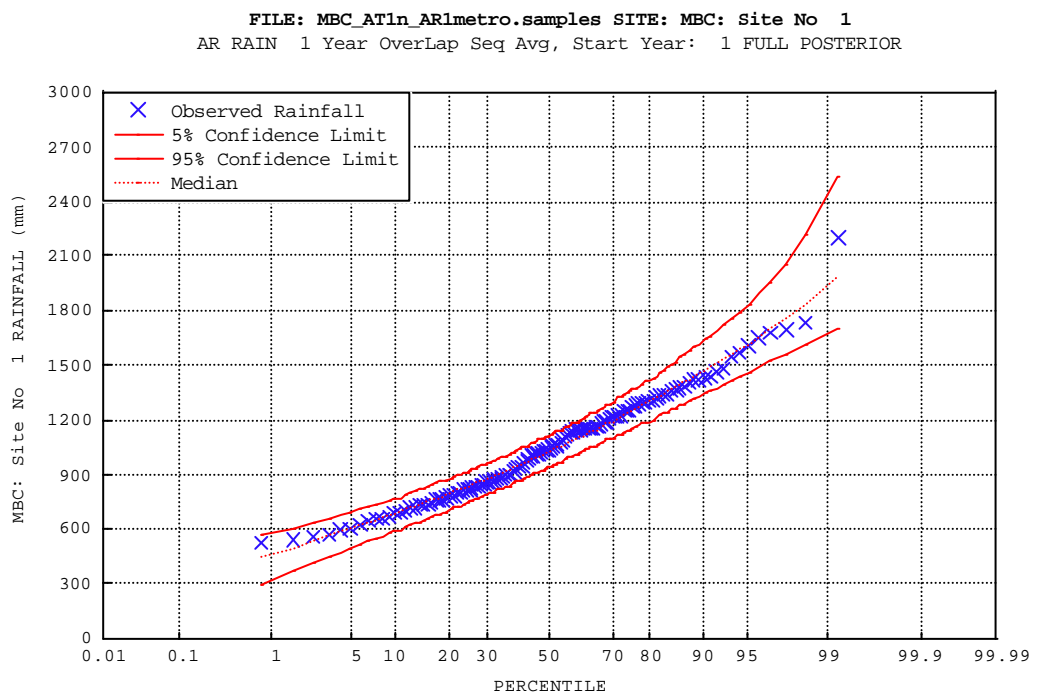
Figure K.6 – Mt. Victoria composite annual (Jan. to Dec.) rainfall data.

FILE: MV_AT5n_AR1metro.samples SITE: MV: Site No 1
AR RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
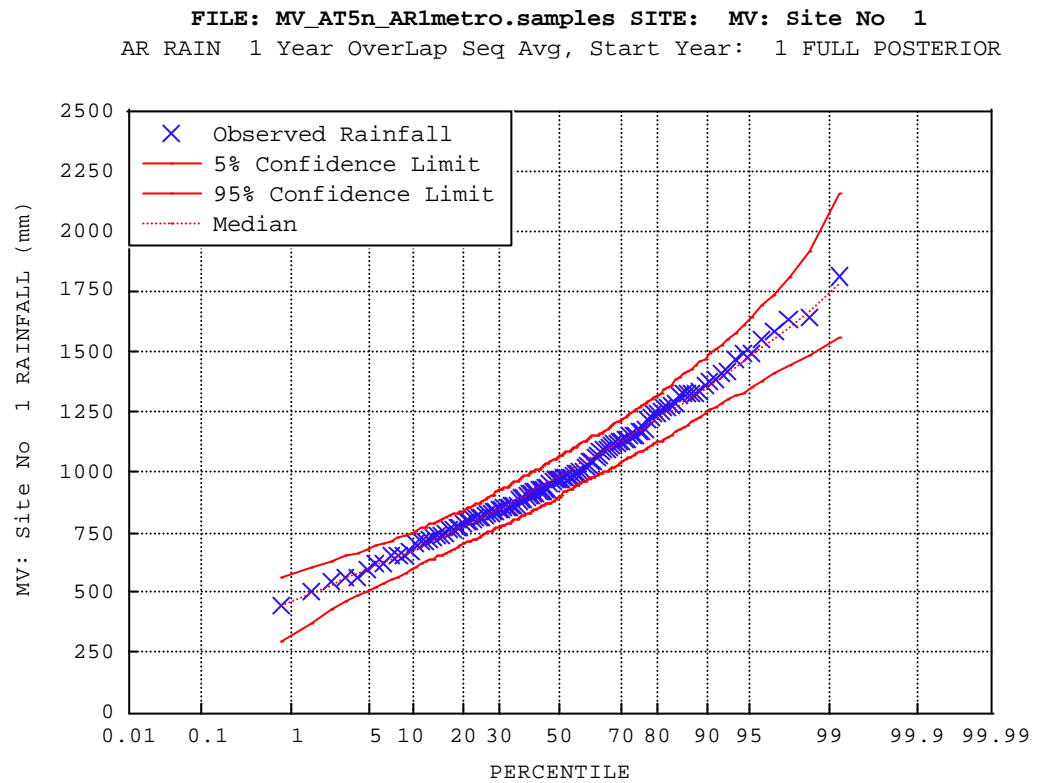
Figure K.7 – Moss Vale annual (May To April) rainfall data.

FILE: TAR_AT1n_AR1metro.samples SITE: TAR: Site No 1
AR RAIN 1 Year OverLap Seq Avg, Start Year: 1 FULL POSTERIOR
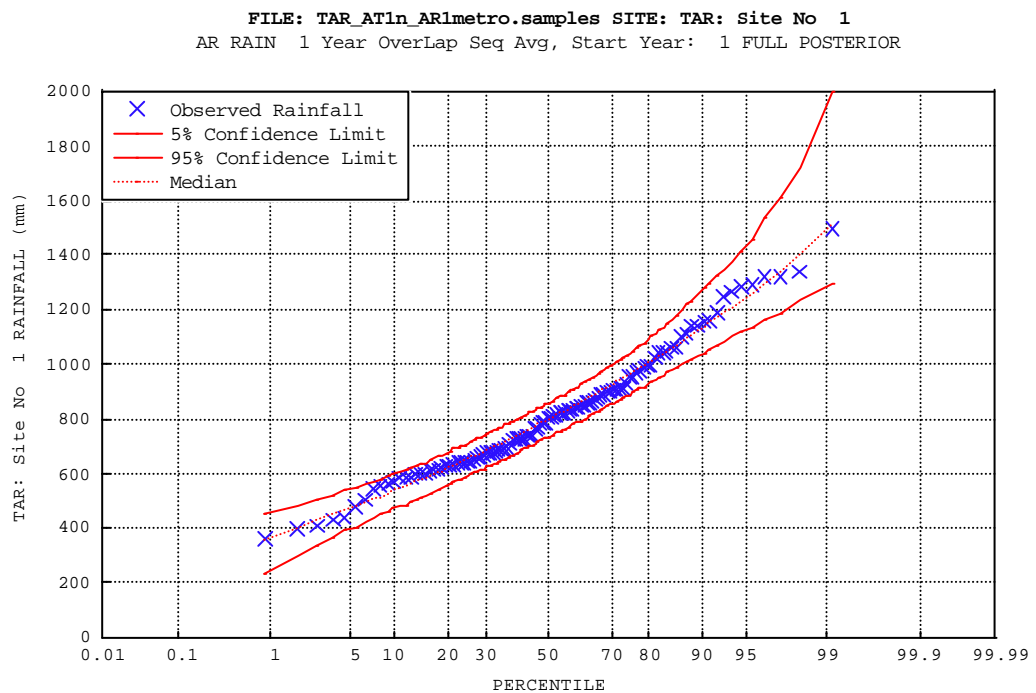
Figure K.8 – Taralga annual (Jan. to Dec.) rainfall data.
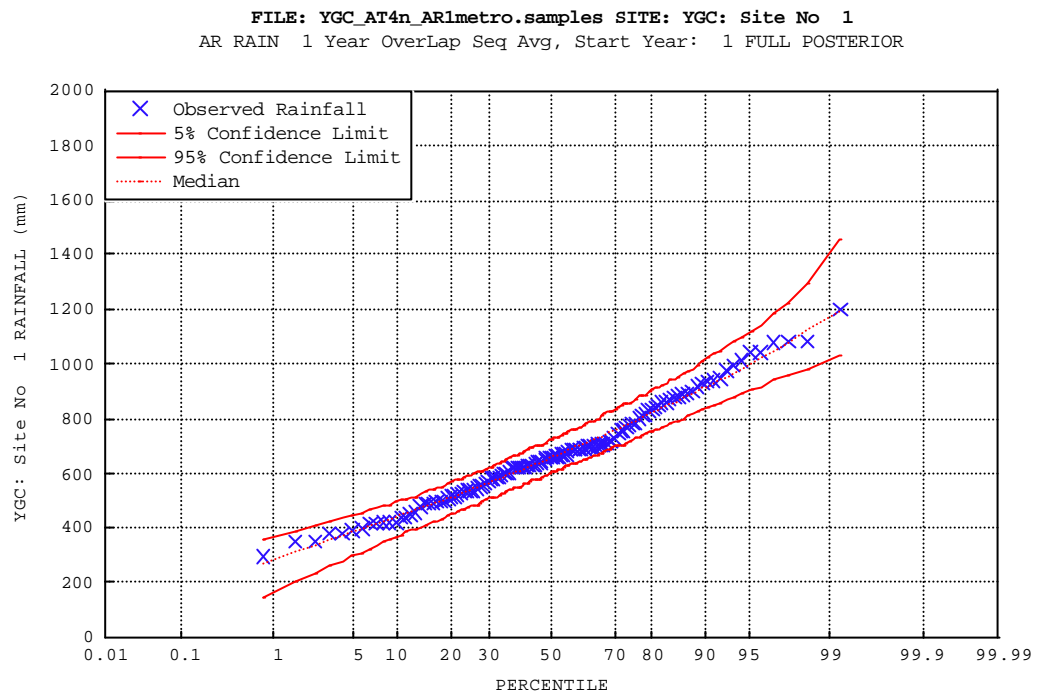
Figure K.9 – Yarra composite annual (April to March) rainfall data.
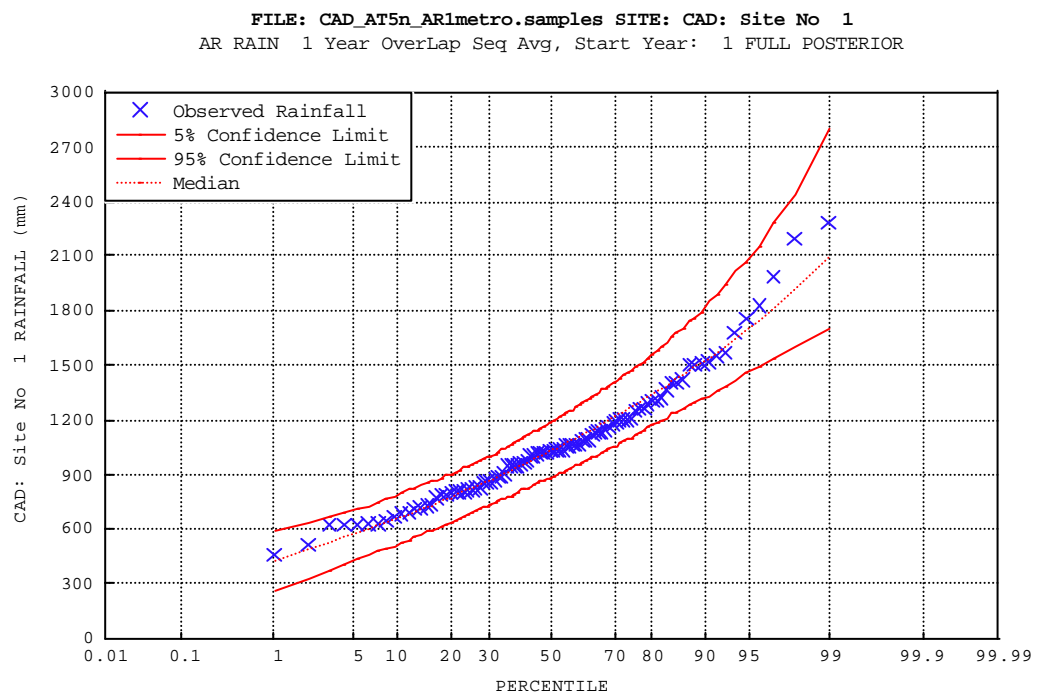
Figure K.10 – Cataract Dam annual (May to April) rainfall data.

Appendix L - Verification of Multi-site HSM Model Calibration Procedure

L.1   Introduction

To verify that the Gibbs sampler had been correctly formulated for the multi-site HSM model and the computer code used to implement the Gibbs sampler was free of any gross errors, synthetic calibration runs were used. Multi-site synthetic data was generated using the multi-site HSM model and the Gibbs sampler was used to determine if the true parameter values could be recovered. Multi-site data from five sites was generated because five is maximum of number of sites that will be used for calibration in this thesis. The synthetic parameter values used for the state rainfall distributions were similar to the expected values from the posterior for the Warragamba catchment rainfall data [refer Table 7.2] because for these sites a two-state persistence structure was identifiable. The spatial correlation between all the sites was set to 0.8 for both the wet and dry states, as this was the approximately the average value of the spatial correlation for the Warragamba catchment rainfall sites. The values for the transition probabilities were chosen which corresponded to a medium wet and dry persistence structure. Synthetic time series of 100, 1000, and 10,000 data points were used to calibrate the HSM model to determine if the posteriors converged towards the true synthetic parameter value as the number of data points increased.

In addition, it was also verified that the method for sampling missing data values was correctly implemented. The method used was to take the synthetic rainfall data for one of the sites and remove a portion of the data series to mimic that the data was missing. This was completed for the 100 and 1000 time series lengths, where 10% and 5% of the values were removed respectively. It was not done for 10,000 time series as removing 10% of the data for one site means 1000 missing data points and 1000 extra parameters. The computer code developed to implement the Gibbs sampling procedure was not designed to handle 1000 parameters.

L.2   Results

The posteriors for p_{WD} and p_{DW}, the rainfall parameters for site 1 and the correlation between site 1 and site 2 are shown in Figure L.1. Again, percentile box plots are used to compare the posteriors [refer Appendix F]. It is clearly seen that as the number of data points increases the posterior of each parameter converges to the true parameter value. A similar result was found for all the other multi-site HSM model parameters.

(a) p_{WD} **(true value=0.14)**   (a) p_{DW} **(true value=0.14)**
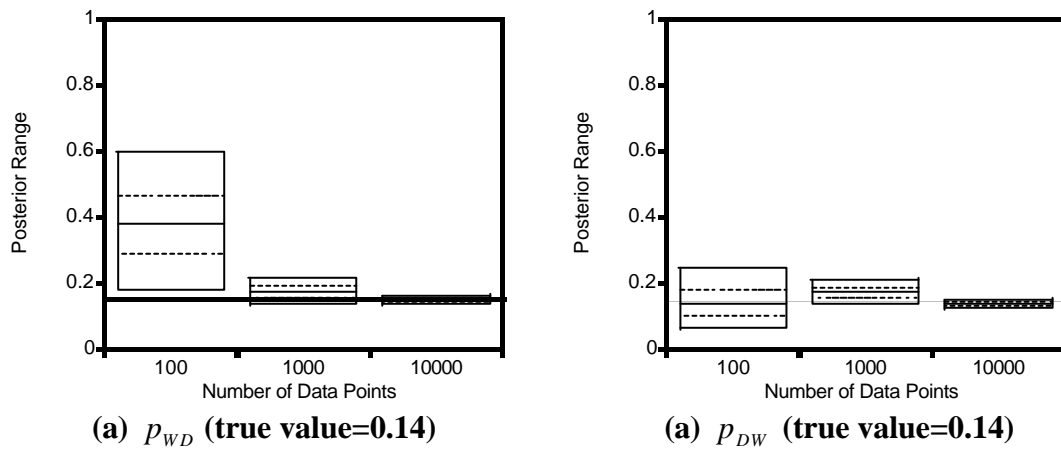
Figure L.1 – Posteriors for selected HSM model parameters for multi-site synthetic series with varying number of data points. Dark line indicates true parameter value.
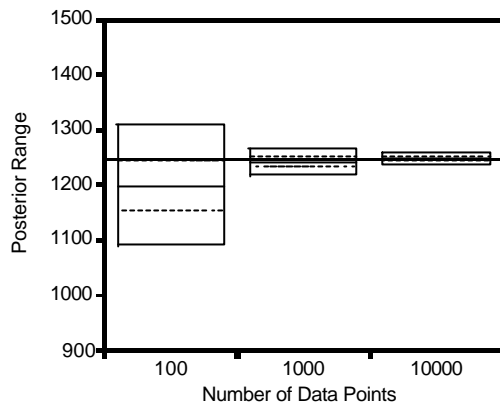
(c) Site 1: μ_W (true value=1240.0)
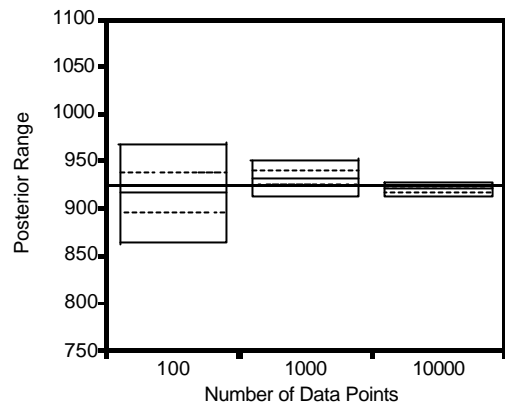
(d) Site 1 : μ_D (true value=925.0)
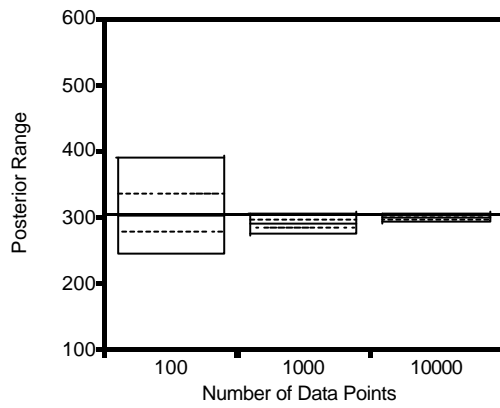
(e) Site 1: σ_W (true value=304.0)
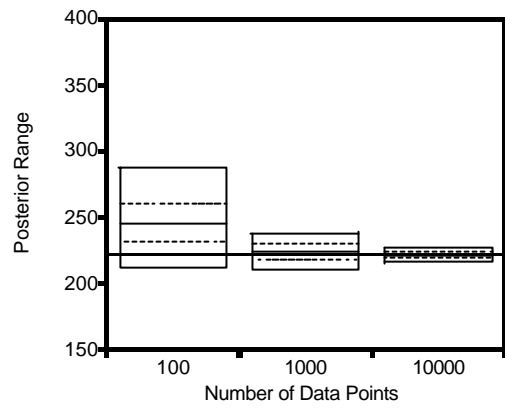
(f) Site 1: σ_D (true value=304.0)
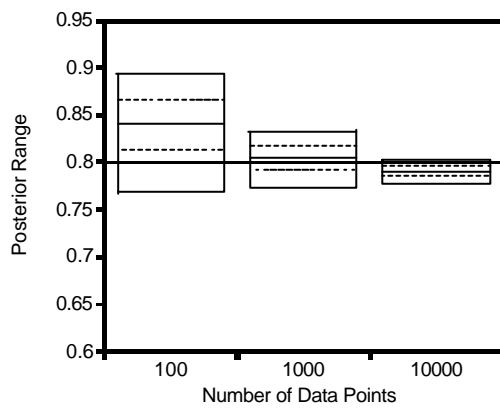
(g)   Site 1 & 2 correlation (wet) (true value=0.8)
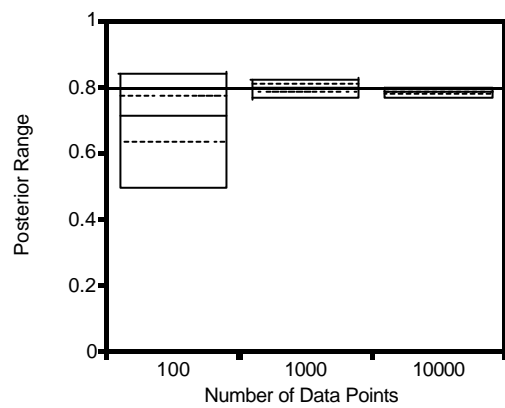
(h)   Site 1 & 2 correlation (dry) (true value=0.8)

Figure L.1(cont.) – Posteriors for selected HSM model parameters for multi-site synthetic series with varying number of data points. Dark line indicates true parameter value.

The results for the verification of the missing data procedure are shown in Figure L.2 for the 100 year case when 10% of the data was removed from one site. The posteriors of selected parameters are shown for the site with the missing data. These are compared to the case when the data was not missing for the 100 year time series. It can be seen that there was little difference in the posterior variance when the data was missing. Only the dry state mean [Figure L.2(d)] showed a slight increase in the posterior variance. This is likely to be because there was a low proportion of missing data from only one site and the rainfall information data from four other sites was used to estimate the missing rainfall data values for that one site.

(a) p_{WD} (true value=0.14)

(b) p_{DW} (true value=0.14)

(c) μ_W (true value=1240.0)

(d) μ_D (true value=925.0)
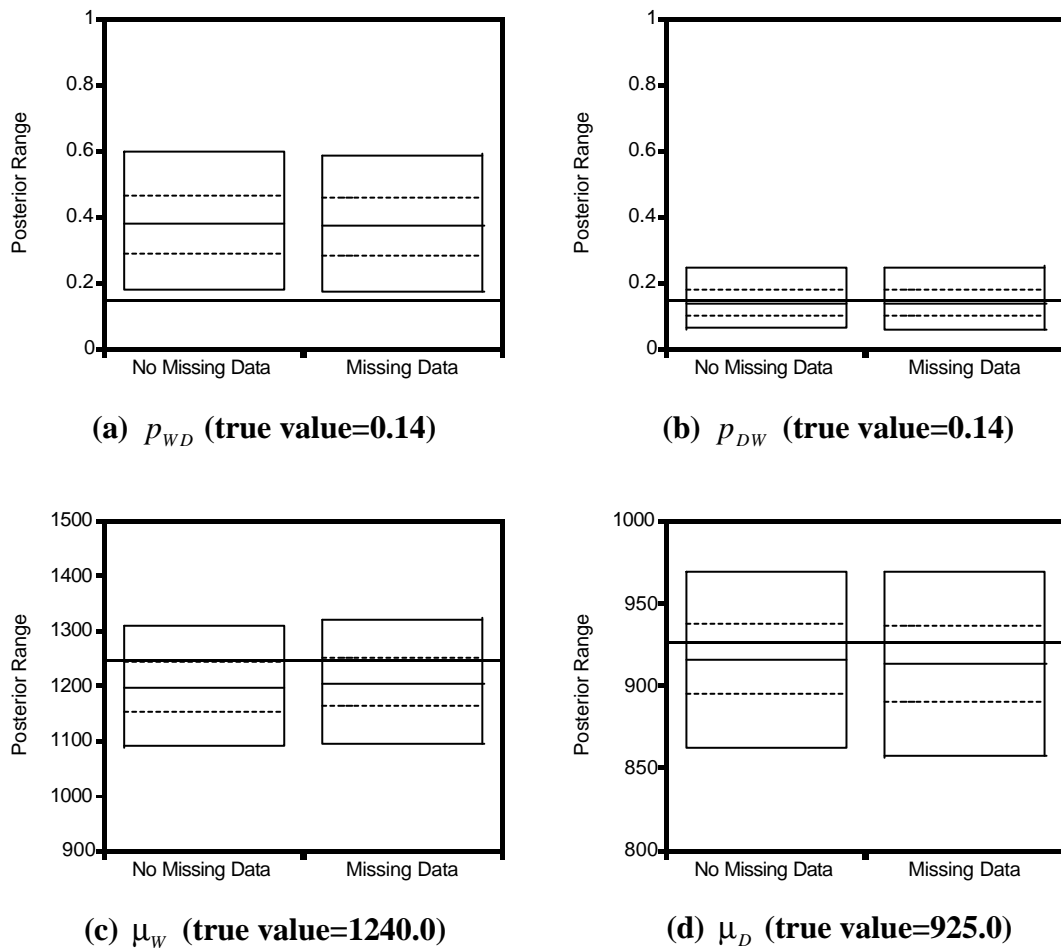
Figure L.2 – Comparison of posteriors for selected HSM model parameters for multi-site synthetic series with 10% missing data (one site) and without missing data for the 100 year time series. Dark line indicates true parameter value.

Figure L.3 shows the posterior of the missing data values compared to the actual synthetic value for the 100 year time series. This shows that the majority of the posteriors for the missing data values are in the vicinity of the actual synthetic data value. It would not be expected that the posterior should capture the true synthetic data values because these values are estimated based on correlations from four data points from four sites. However, these results do illustrate how close the missing data posteriors are to the actual synthetic parameter value.
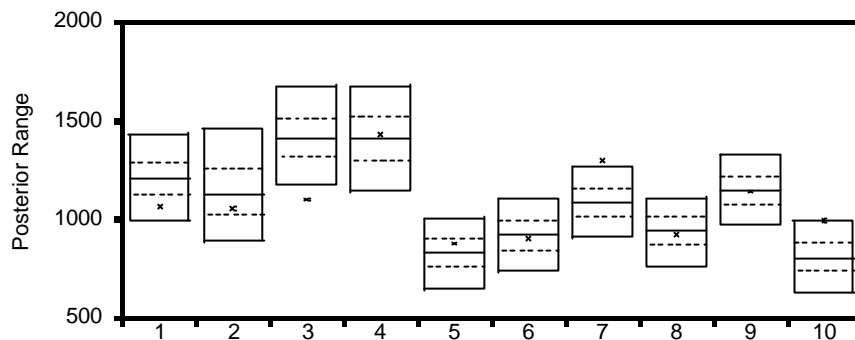
Figure L.3 – Comparison of posteriors for the missing data values to the actual synthetic values (shown as crosses) for the 100 year time series.

Similar results to the above were also found for the 1000 year time series, the posteriors did not show any distinct changes when missing data was sampled. These results verified that the method for sampling the missing data values has been correctly implemented.

L.3   Conclusion

It was found that as the number of the data points increased the posterior of every parameter would converge to the true parameter value. This was as expected and verified that the implementation of the Gibbs sampler for the multi-site HSM model, including the procedure for handling the missing data, has been successfully formulated and coded.

Appendix M - Comparison of Priors to Posteriors for Multi-site HSM Model

M.1 Introduction

Similar to Appendix G for the single site HSM model in this Appendix the priors are compared to the posteriors for the multi-site HSM model. This is to verify that these priors were actually diffuse compared to the posteriors.

M.2   Results

The priors and posteriors for the state mean and standard deviation for the Warragamba catchment rainfall data four-site analysis are compared in Figure M.1 to Figure M.4. Note that compared to the single site results given in Appendix G the priors are more diffuse. The prior and posterior comparisons for the state correlation parameters are given in Figure M.5. The priors can be seen to be diffuse. Interestingly it can be seen that the dry state correlations are generally higher than the wet state correlations. For the Warragamba catchment rainfall data five-site analysis all the priors will not be shown as the results are similar to the four-site analysis – the priors are relatively diffuse compared to the posteriors. The priors for the extra site included in the analysis, Cataract Dam, are shown in Figure M.6. The state rainfall correlations between that site and the four other sites are shown in Figure M.7. Again, the priors can be seen to be diffuse compared to the posteriors.

For the Williams River catchment three-site analysis the comparison of the priors to the posteriors for the state mean and standard deviation is shown in Figure M.8 to Figure M.10. For the state correlation this comparison is shown in Figure M.11. It can be seen that the priors are diffuse compared to the posteriors.

In these results for the Warragamba and the Williams River catchments there is a feature in the priors that warrants a comment. The priors for the state rainfall correlations have a peak at a correlation of one. This result was unexpected, and it was questioned whether the priors had been correctly formulated. *DeGroot* [1970] provides a function to calculate the distribution of a correlation for the bivariate Gaussian case. It was found to be of the same form as shown in this Appendix with the prior mode located at a correlation of one. This is a minor concern and it may have to be investigated if alternative priors are available as part of the future research. However, given that generally the priors for the correlations were relatively diffuse over the range of the posteriors this was considered adequate for this analysis.
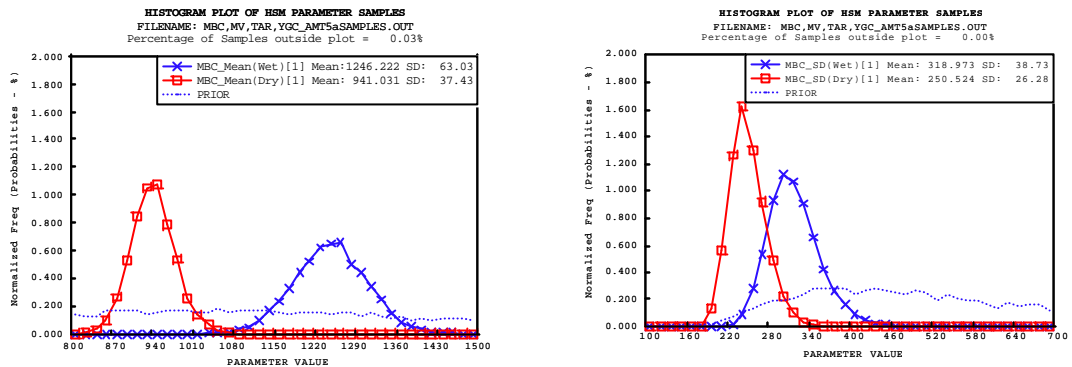
Figure M.1 – Mt. Victoria (MtV) state rainfall parameters, four-site analysis.
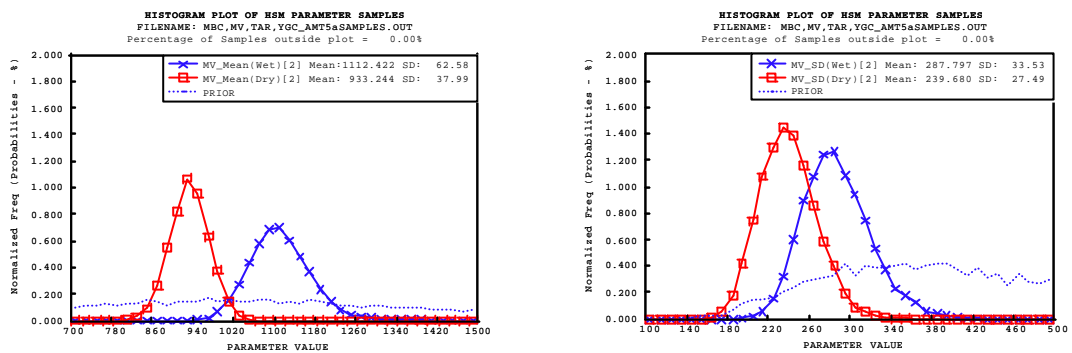
Figure M.2 – Moss Vale (MV) state rainfall parameters, four-site analysis.
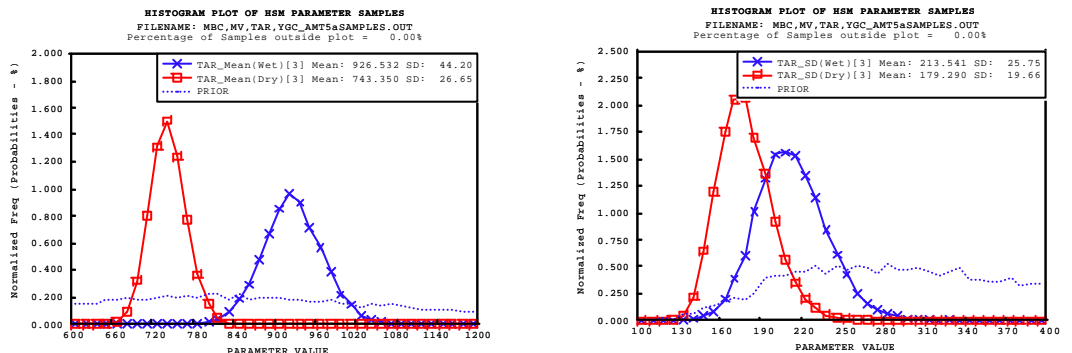
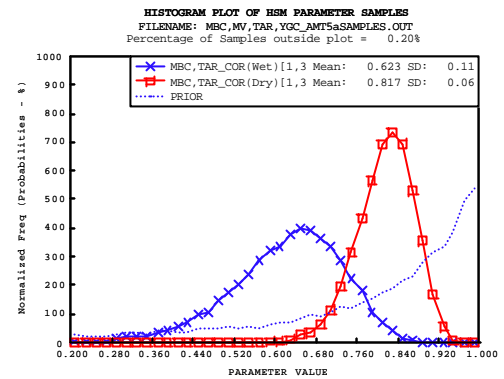**Figure M.3 – Taralga (TR) state rainfall parameters, four-site analysis.**
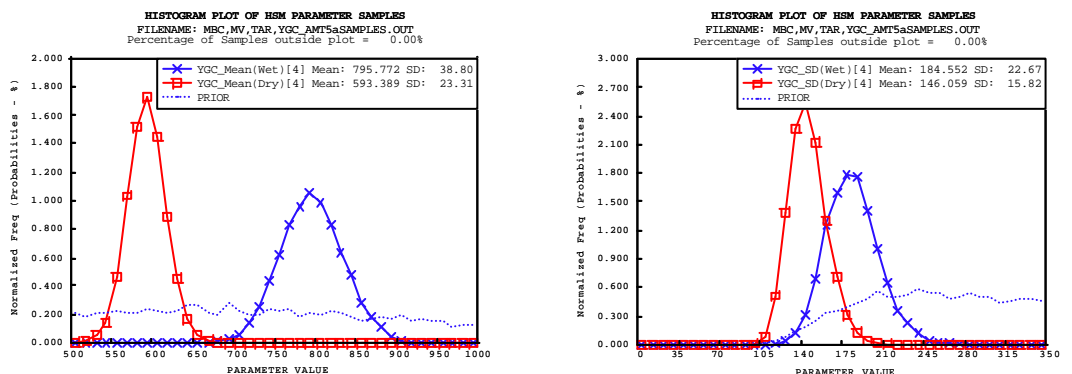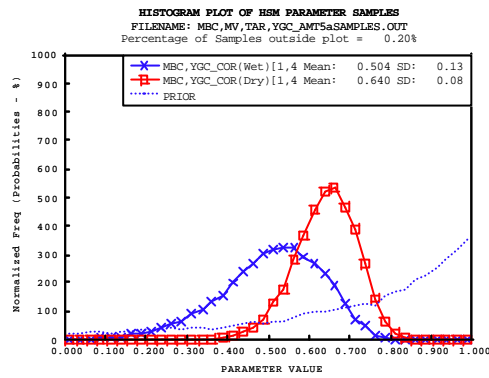

Figure M.4 – Yarra (YA) state rainfall parameters, four-site analysis.
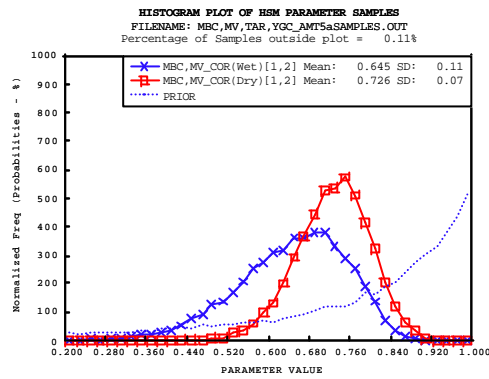
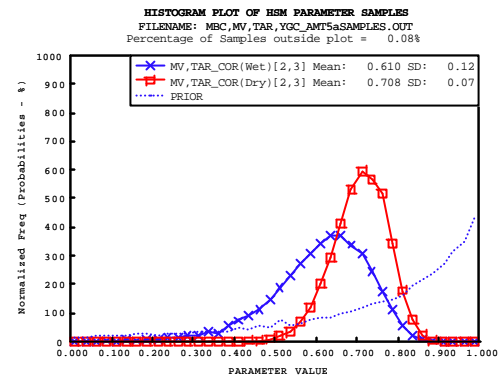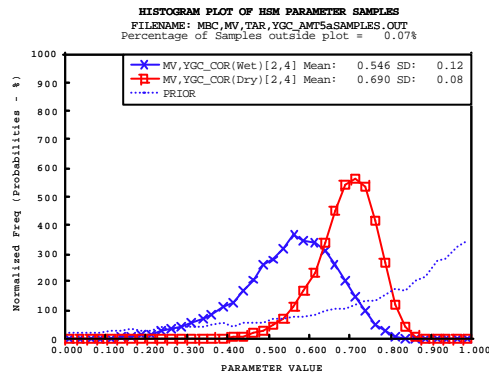**(a) MtV and MV correlation**

(b) MtV and TR correlation

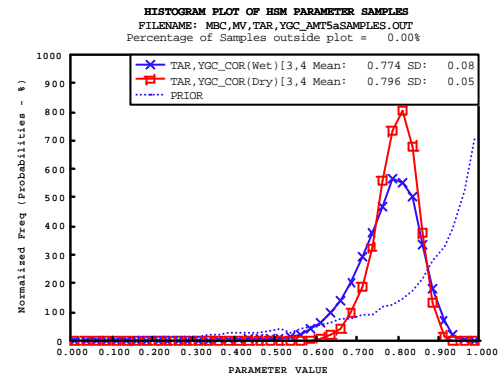**(c) MtV and YA correlation**

(d) MV and TR correlation

(e) MV and YA correlation

(f) TR and YA correlation

Figure M.5 – State rainfall correlation parameters for the Warragamba catchment four-site analysis.
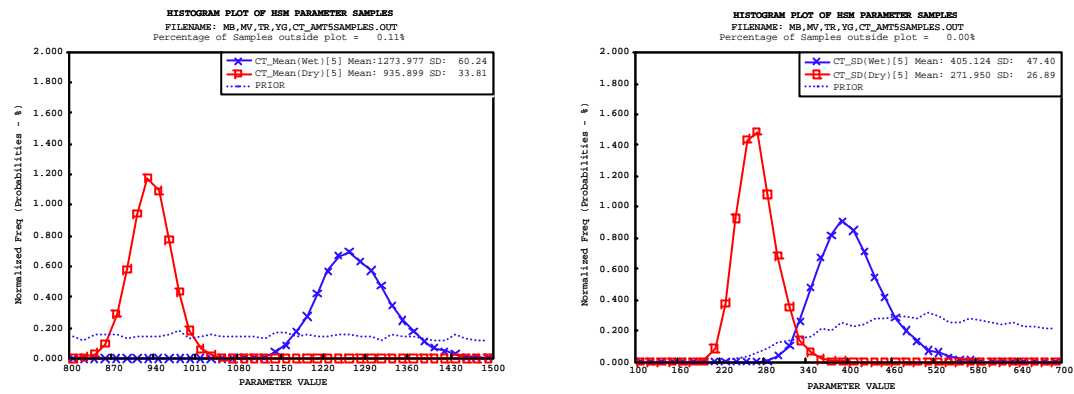
Figure M.6 – Cataract Dam (CD) state rainfall parameters, five-site analysis.

(a) CD and MtV correlation

(b) CD and MV correlation

(c) CD and TR correlation

(d) CD and YA correlation
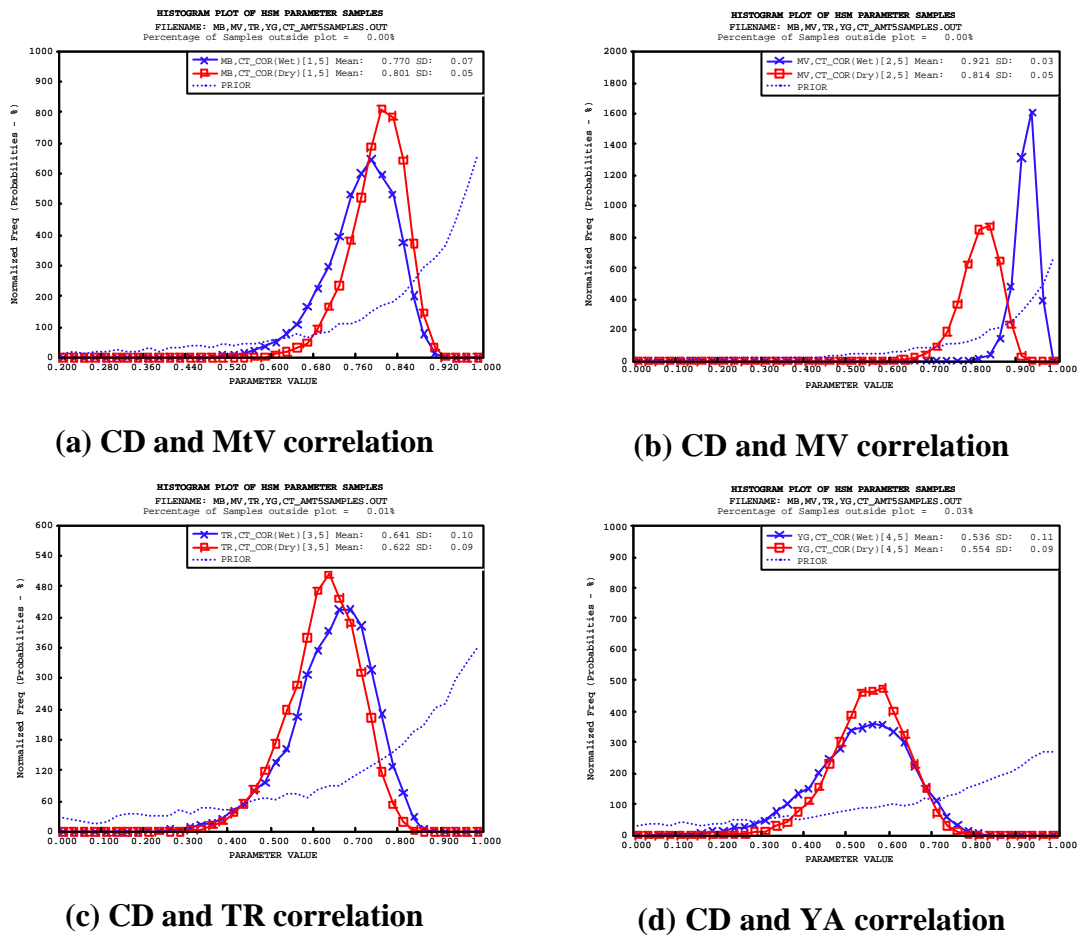
Figure M.7 – State rainfall correlation parameters between Cataract Dam and the four other Warragamba catchment sites, five-site analysis.
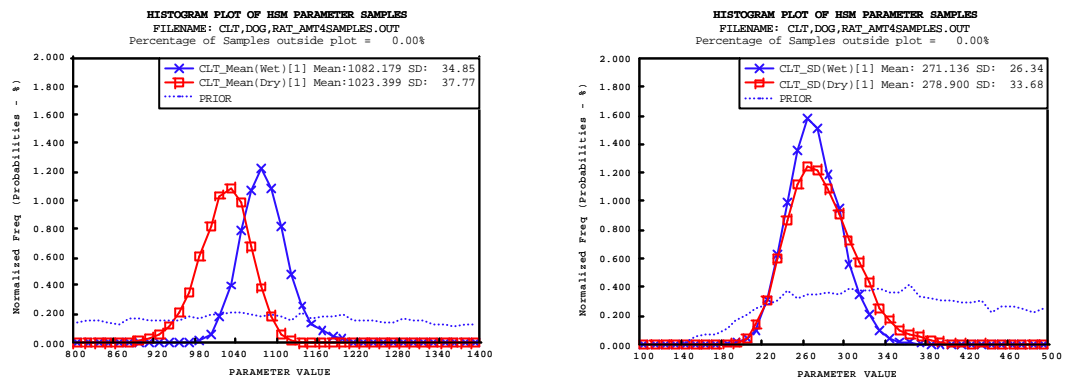
Figure M.8 – Clarence Town (CT) state rainfall parameters.
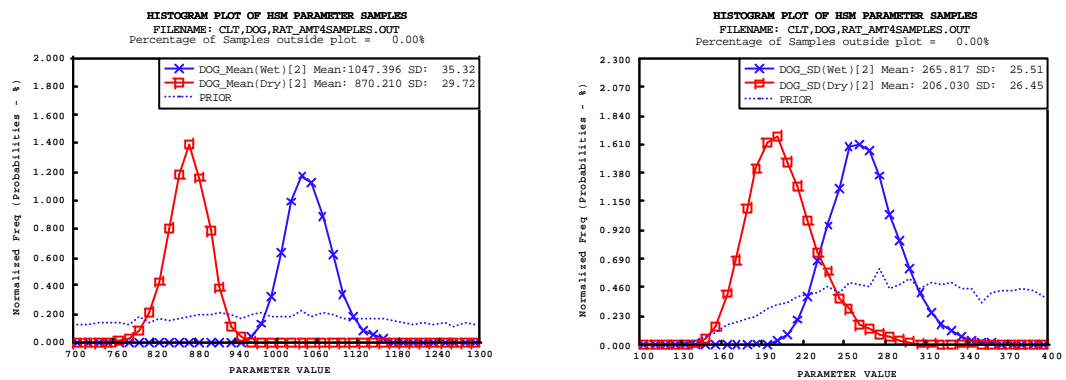
Figure M.9 – Dungog (DG) state rainfall parameters.
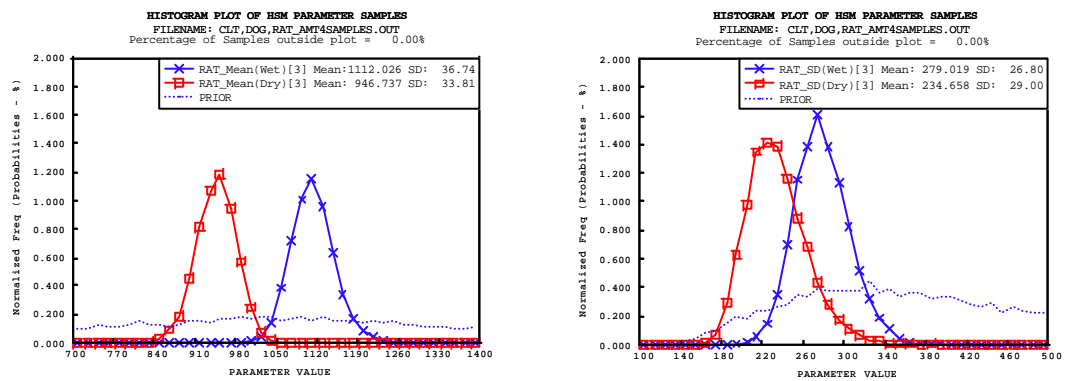
Figure M.10 – Raymond Terrace (RT) state rainfall parameters.
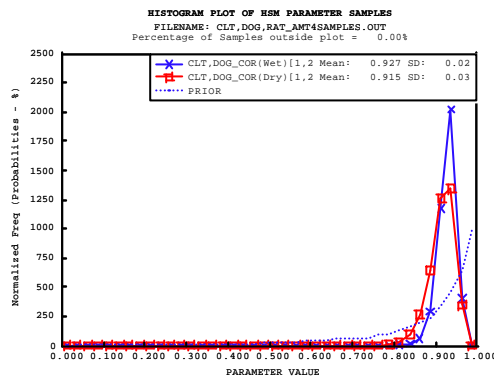
(a) CT and DG correlation
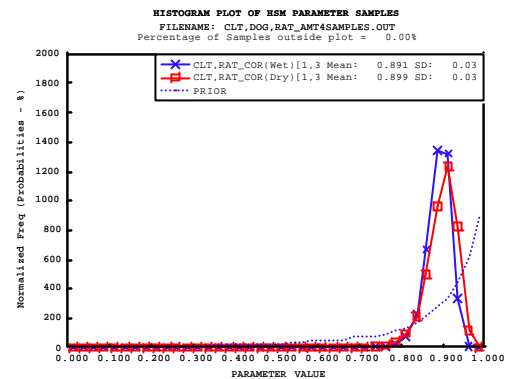
(b) CT and RT correlation
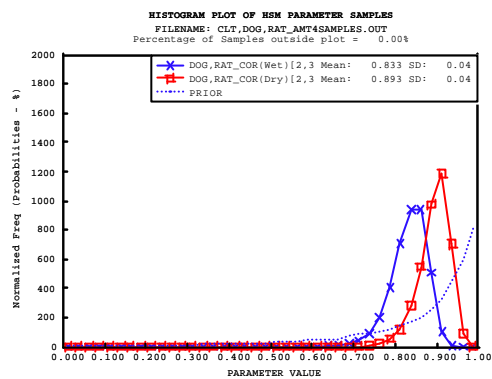
(c) DG and RT correlation

Figure M.11 – State rainfall correlation parameters for William River catchment three-site analysis.

M.3   Conclusion

The results given in this Appendix confirmed that the informative priors for the state rainfall parameters in the calibration procedure for the multi-site HSM model were diffuse compared to the posteriors. Hence they exert only a minor influence on the inferences.

Appendix N - Williams River Catchment Multi-site Results

N.1   Introduction

For the Williams River three-site analysis with Clarence Town, Dungog, and Raymond Terrace it was found that the Gibbs sampler had difficulty converging for the majority of water years. A similar result was also found for the Williams River five-site analysis, except none of the water years were able to achieve convergence. In this Appendix iterative sequences of parameter samples are used to provide a further explanation of why the Gibbs sampler could not converge.

N.2   Results

During the iterative sequence of the Gibbs sampler some of the chains would flip between what seemed to be two different modes in the posterior. The two different modes had hidden state time series which were the basically the inverse of one another. This is shown in Figure N.1 for the Williams River three-site analysis, January to December water year. Similar results were found for other water years, except for the April to March water year, which was able to achieve convergence to the equivalent of mode 1 shown in Figure N.1.

Iterative sequences of selected parameter samples show the differences between the two modes. One mode corresponded to a strong wet and dry separation for Clarence Town [Figure N.2] but not for Dungog [Figure N.3] or Raymond Terrace [Figure N.4]. The other mode had the opposite, low separation for Clarence Town and high separation for Dungog and Raymond Terrace. The standard deviation (SD) for Dungog would also flip from the wet SD being higher than the dry to the dry SD being higher than wet [Figure N.5]. There was little change in remaining parameters for the different modes.

N.3   Conclusion

As stated in Section 11.2.2 various techniques were trialled to alleviate this problem. However none were able to facilitate convergence. The reasons for nonconvergence remain unknown. It is likely that the high correlation (close to 0.9) between the sites maybe a contributing factor. It is recommended that further investigations using synthetic calibration runs be undertaken to develop a greater understanding of the multi-site HSM model and its calibration procedure.
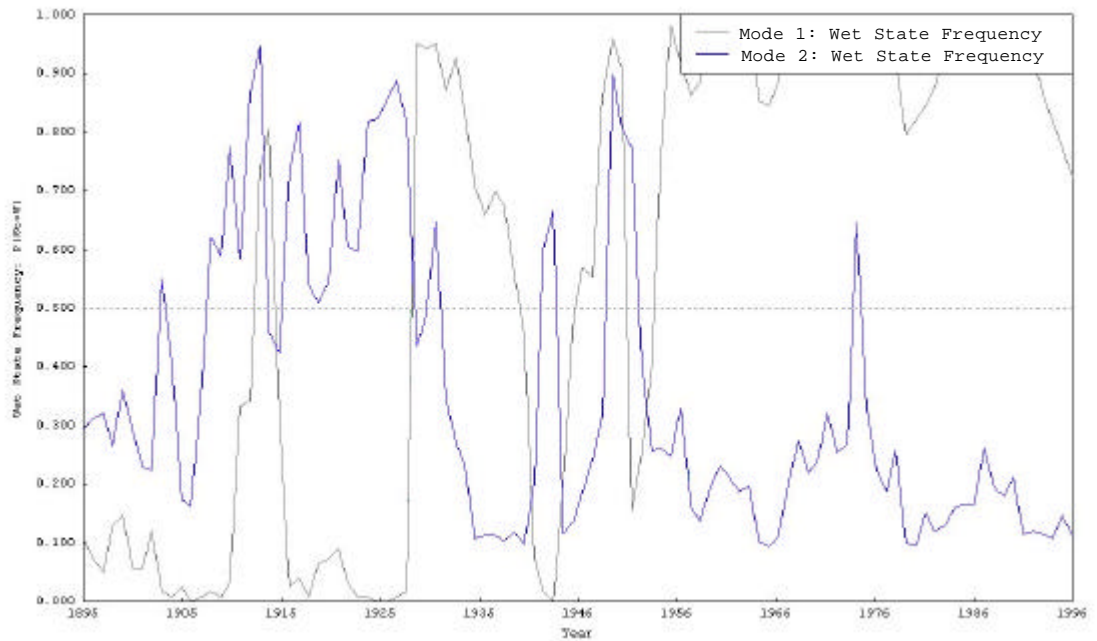
Figure N.1 – State frequency time series for the two different modes in the posterior of the Williams River three-site analysis.
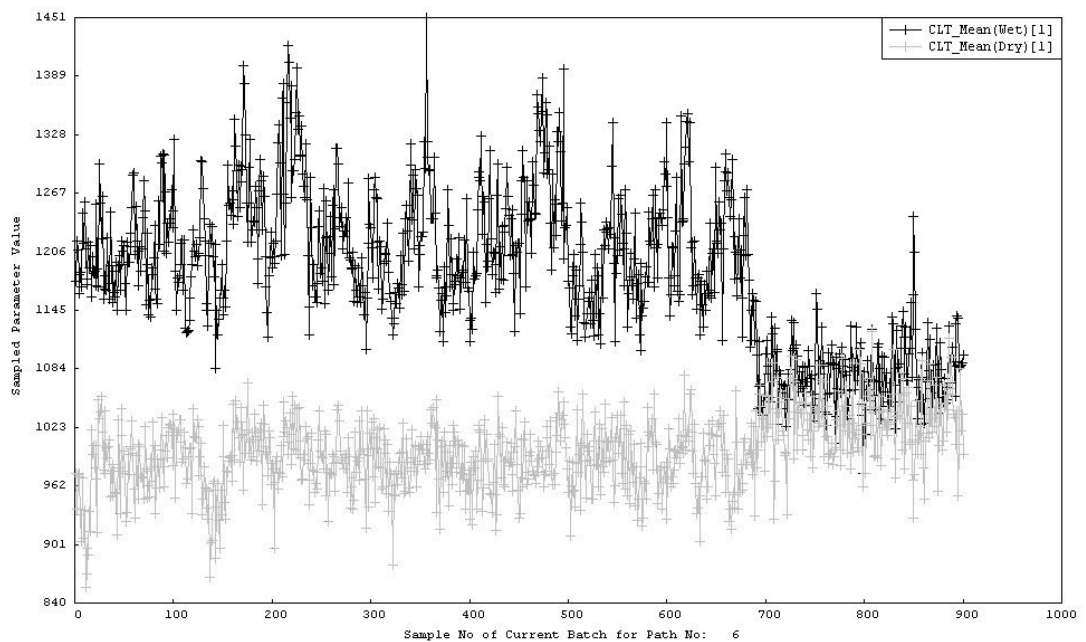

Figure N.2 – Iterative sequence of parameter samples for Clarence Town (CLT) wet and dry mean showing a transition between the two modes.
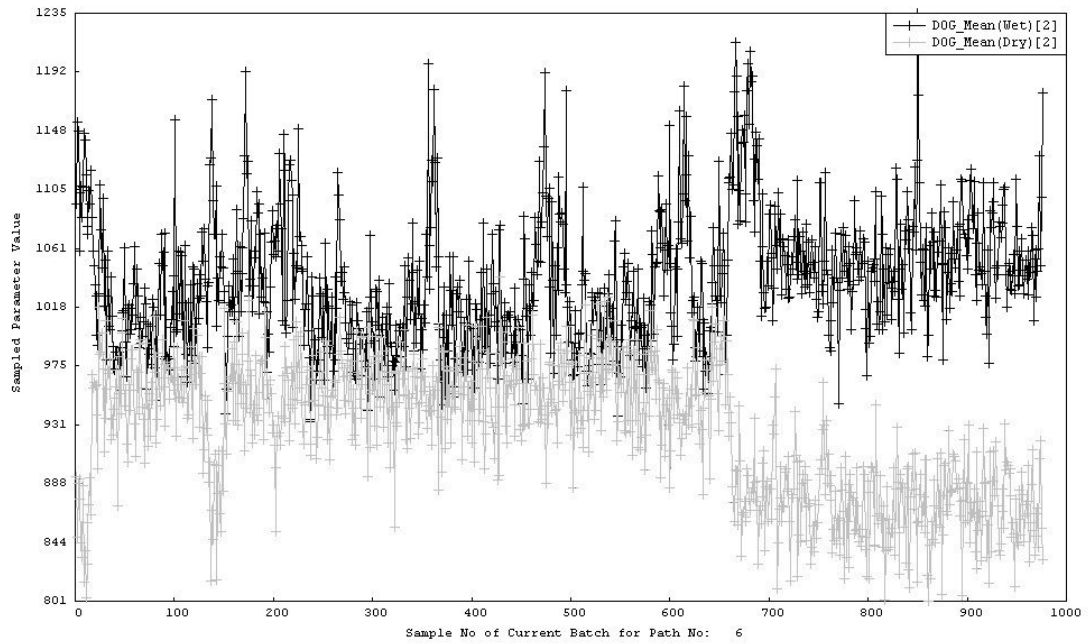
Figure N.3 – Iterative sequence of parameter samples for Dungog (DOG) wet and dry mean showing a transition between the two modes.
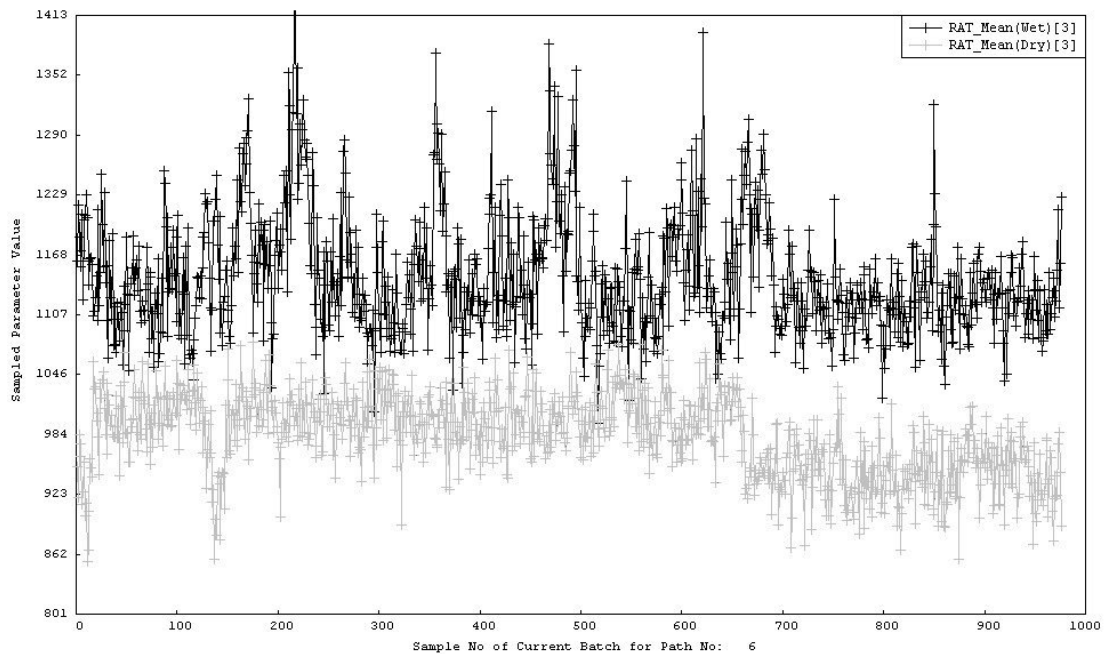

Figure N.4 – Iterative sequence of parameter samples for Raymond Terrrace (RAT) wet and dry mean showing a transition between the two modes.
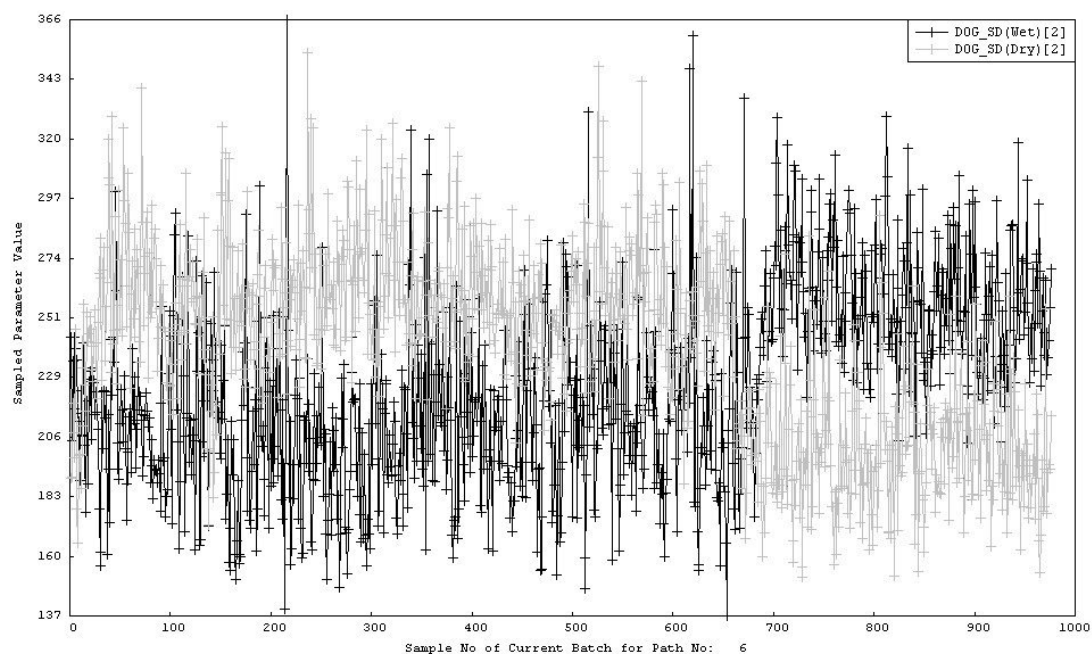
Figure N.5 – Iterative sequence of parameter samples for Dungog (DOG) wet and dry standard deviation showing a transition between the two modes.