

# Modelling Annual Rainfall Using a Hidden-State Markov Model

R Srikanthan<sup>ab</sup>, M A Thyer<sup>c</sup>, G Kuczera<sup>d</sup> and T A McMahon<sup>bc</sup>

<sup>a</sup> Bureau of Meteorology, Melbourne, Australia 3000 (r.srikanthan@bom.gov.au)

<sup>b</sup> Cooperative Research Centre for Catchment Hydrology, Monash University, Australia 3800

<sup>c</sup> Department of Forest Resources Management, University of British Columbia, BC, Canada

<sup>d</sup> Department of Civil, Surveying and Environmental Engineering, University of Newcastle, Australia 2308

<sup>e</sup> Department of Civil and Environmental Engineering, University Melbourne, Australia 3010

**Abstract:** In the past, stochastic modelling approaches generally assumed no variation in the parameters between years. There is a growing awareness of long term persistence in climatic data in the form of wetter and drier years. To take this information into account, model parameters should be varied in some way. The hidden-state Markov (HSM) model assumes that the climate is composed of two states, either a dry state (low rainfall year) or a wet state (high rainfall year). This approach provides an explicit mechanism for the HSM model to simulate the influence of quasi-periodic phenomenon such as ENSO. Separate distributions are used to model the rainfall in the two states. The states are not known *a priori* and are estimated along with the model parameters. For model calibration, a Bayesian framework is used to infer the distribution of the model parameters. A Markov Chain Monte Carlo method known as Gibbs sampler is used to estimate the parameters and their uncertainties. The HSM model was applied to 44 rainfall stations located in various parts of Australia and the results indicated that 32 stations are either highly likely to or could possibly have two states. One thousand replicates each of length equal to the historical data were generated and several model evaluation statistics computed. The HSM model satisfactorily preserved all the model evaluation statistics.

**Keywords:** Annual rainfall; Hidden-state Markov model; Parameter uncertainty; Bayesian;

## 1. INTRODUCTION

The modelling of annual rainfall data serves two purposes. Firstly, it enables the understanding of the stochastic nature of the annual rainfall data and its implications for long periods of low and high rainfall. This understanding is necessary to manage water supply systems during low rainfall periods. Secondly, any stochastic model should be able to maintain its statistical characteristics at different time scales and a good annual rainfall model allows one to disaggregate the generated annual rainfall data into monthly data. In this case, the annual data becomes the input to various disaggregation schemes.

The lag one Markov or the first order autoregressive model has been widely used to generate annual rainfall data [Srikanthan and McMahon, 1985, 2000]. The main drawback with this model is that it cannot model the long wet and

dry spells observed in the data. Thyer and Kuczera [1999, 2000] developed a hidden-state Markov (HSM) model which explicitly assumes that the climate has two states. The HSM model was applied to annual rainfall data from 5 capital cities in Australia and it was found that only two cities, namely, Brisbane and Sydney exhibited two-state persistence.

The objective of the present study is to apply the HSM model to a large set of rainfall data from different climatic conditions and determine the extent to which two-state persistence exists in Australian annual rainfall data. Based on the analysis of both the Sydney rainfall data and stochastically generated data, Thyer and Kuczera [2000] observed that long-term records of data with length in excess of 120 years are required to detect two-state persistence. Notwithstanding this observation, it was decided to utilize 44 sites

(Table 1) with long records (but shorter than that recommended by Thyer and Kuczera). The

selected sites adequately represent the different climatic conditions in Australia.

Table 1. Details of the rainfall stations used in the study.

Number	Name	Latitude	Longitude	Length (years)	Mean (mm)	Category <sup>†</sup>
1005	Wyndham Port	-15.46	128.10	79	695	b
2016	Lissadell	-16.67	128.57	105	616	a
5008	Mardie	-21.19	115.98	108	276	c
6036	Meedo	-25.66	114.62	94	216	b
9034	Perth	-34.93	138.58	115	868	a
10037	Cuttening	-31.73	117.76	96	312	a
12065	Norseman Post Office	-32.20	121.78	102	287	c
14902	Katherine Council	-14.46	132.26	111	974	a
15540	Alice Springs Post Office	-23.71	133.87	112	280	b
17031	Marree	-29.65	138.06	113	164	c
19032	Orroroo	-32.74	138.61	118	341	c
22020	Wallaroo	-33.93	137.63	135	360	c
23000	Adelaide	-31.95	115.84	139	530	a
24511	Eudunda	-34.18	139.09	118	446	c
28004	Palmerville	-16.00	144.08	109	1034	c
33035	Kalamia Estate	-19.54	147.41	112	1085	b
35027	Emerald Post Office	-23.53	148.16	108	642	b
36007	Barcaldine Post Office	-23.55	145.29	112	496	b
39023	Cape Capricorn Lighthouse	-23.48	151.23	87	801	c
39082	Rockhampton Post Office	-23.40	150.50	96	946	b
40043	Cape Moreton Lighthouse	-27.03	153.47	129	1550	c
40214	Brisbane	-27.48	153.03	133	1154	c
41082	Pittsworth Post Office	-27.71	151.63	112	703	a
42023	Miles Post Office	-26.66	150.18	114	661	c
44026	Cunnamulla Post Office	-28.07	145.68	120	374	b
47053	Wentworth Post Office	-34.11	141.91	132	288	b
49002	Balranald RSL	-34.64	143.56	121	322	c
54004	Bingara Post Office	-29.87	150.57	113	745	c
62021	Mudgee (George Street)	-32.59	149.58	122	670	c
66062	Sydney	-33.86	151.20	140	1226	b
69018	Moruya Heads Pilot Station	-35.91	150.15	123	972	b
72000	Adelong	-35.31	148.06	115	795	a
72044	Tumut	-35.30	148.22	113	822	a
75031	Hay Miller Street	-34.52	144.85	119	369	b
77030	Narraport	-36.01	143.03	112	354	a
80056	Tongala	-36.25	144.95	69	443	c
81007	Caniambo	-36.46	145.66	95	524	c
84030	Orbost	-37.63	148.46	115	855	c
86071	Melbourne	-37.81	144.97	143	657	a
86117	Toorourrong Reservoir	-37.48	145.15	106	804	a
87043	Meredith (Darra)	-37.82	144.15	124	685	a
91033	Frankford (Rossville)	-41.32	146.73	106	1069	b
92012	Fingal	-41.64	147.97	110	611	b
94061	Sandford (Maydena)	-42.93	147.52	111	578	b

<sup>†</sup> Persistence structure: a – highly unlikely to have two-state persistence  
b – highly likely to have two-state persistence  
c – possibly to have two-state persistence

## 2. HIDDEN-STATE MARKOV MODEL

The HSM model (Figure 1) assumes the climate is in one of two states: wet (W) or dry (D). Each state has an independent rainfall distribution, assumed to be Gaussian. The time spent in each state is governed by the state transition probabilities. This provides an explicit mechanism to replicate variable lengths of wet and dry cycles.

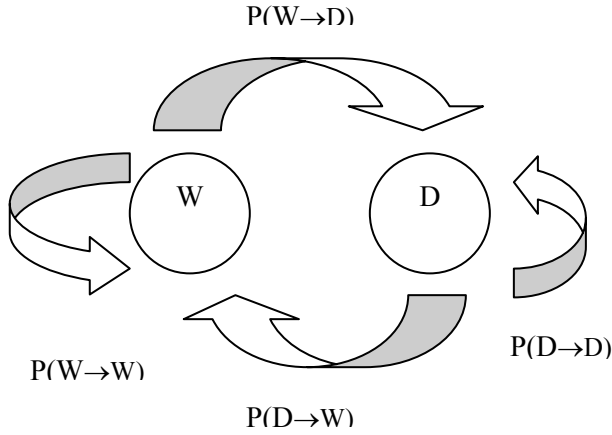


Figure 1. Schematic representation of the HSM model

The simulation of annual rainfall time series is a two-step process. In the first step the climate state at year  $t$ ,  $s_t$ , is simulated by a Markovian process:

$$s_t | s_{t-1} \sim \text{Markov}(\mathbf{P}) \quad (1)$$

where  $\mathbf{P}$  is a (2x2) state transition probability matrix whose elements are:

$$p_{ij} = \Pr(s_t = j | s_{t-1} = i) \quad i, j = W, D \quad (2)$$

Once the state for year  $t$  is known, the rainfall is simulated using:

$$y_t \sim \begin{cases} N(\mu_w, \sigma_w^2) & \text{if } s_t = W \\ N(\mu_d, \sigma_d^2) & \text{if } s_t = D \end{cases} \quad (3)$$

where  $N(\mu, \sigma^2)$  denotes a Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Therefore the vector of unknown parameters for the HSM model,  $\theta$ , is composed of the rainfall distribution parameters for each state, the transition probabilities, and the hidden-state time series,  $S_N = \{s_1, s_2, \dots, s_N\}$ , where:

$$\theta' = (\mu_w, \sigma_w, \mu_d, \sigma_d, \mathbf{P}, S_N) \quad (4)$$

Prior to model calibration the hidden-state time series is unknown. Thus it is included as a model parameter to be estimated during the calibration process.

## 3. CALIBRATION OF HSM MODEL

For model calibration a Bayesian framework is used to infer the distribution of the model parameters,  $\theta$ , for the given time series data,  $Y_N$ . This distribution is referred to as the posterior distribution of the model parameters,  $p(\theta | Y_N)$ . Since the states are not known a priori, it is not possible to derive an analytical expression for the posterior distribution. Thus Markov Chain Monte Carlo (MCMC) simulation methods are employed to draw samples from the posterior distribution. The basic idea of MCMC methods is to simulate a Markov chain iterative sequence, where at each iteration a sample of the model parameters,  $\theta$ , is generated. Given certain conditions the distribution of these samples converges to a stationary distribution which is the posterior distribution,  $p(\theta | Y_N)$ . To calibrate the HSM model, the MCMC method known as the Gibbs sampler is applied. The details of the calibration process are given in Thyer and Kuczera (2000) and it results in the expected values of the parameters with associated uncertainties.

Several indices [Thyer, 2001] are used to interpret the results and these are briefly defined below. The wet and dry separation index (WADSI) is defined as

$$WADSI = \frac{\mu_w - \mu_d}{\sqrt{(\sigma_w^2 + \sigma_d^2)}} \quad (5)$$

This index is a convenient measure of the separation between the wet and dry states. If the difference between the wet and dry means is large then the value of WADSI will be relatively high.

The state signal index (SSI) is defined as follows:

$$SSI = \frac{\sum |P(W) - 0.5|}{N} \quad (6)$$

Values of SSI close to zero indicate no persistence in the rainfall to stay in either wet or dry state. Values of SSI around 0.3 generally indicates

persistence, but this needs to be confirmed with a visual inspection of a time series plot of the posterior probability of a year being classified as wet  $\{P(s_t = W|Y_N)\}$ .

The strength of the two-state persistence is assessed using the expected state resident times (SRT) and is obtained as the reciprocal of the transition probabilities.

$$E(SRT_D) = \frac{1}{p_{DW}}; E(SRT_W) = \frac{1}{p_{WD}} \quad (7)$$

The posterior probability distributions of the transition probabilities are examined to see whether the transition probabilities are well defined.

#### 4. DISCUSSION OF RESULTS

The HSM model was calibrated to the 44 Australian rainfall stations (Table 1) and the indices described above were derived to interpret the results. As long records are needed to definitely detect the existence of two-state persistence structure, the rainfall stations were classified into the following three categories:

(a) Highly unlikely to have two-state persistence

The posterior probability density function of WADI has a mode less than or equal to zero. There is significant posterior probability mass for both  $p_{DW}$  and  $p_{WD}$  at = 0 and 1 so that the transition probabilities are not identifiable or posterior probability is fairly uniform over wide range of transition probabilities resulting in poorly identified transition probabilities.

(b) Highly likely to have two-state persistence

To have two-state persistence, wet and dry distributions must be separate (high WADSI). There is zero or very small posterior mass for WADSI less than or equal to zero. Zero posterior probability mass for both  $p_{DW}$  and  $p_{WD}$  at = 0 and 1 and well defined transition probabilities so that the annual climate can move between states.

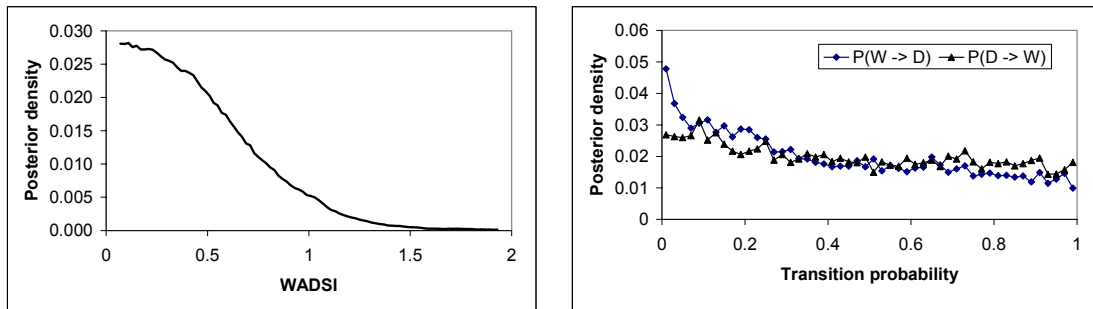
(c) Possibly have two-state persistence

All the remaining sites fall into this category meaning that the likelihood of two-state persistence is not clear, but possible.

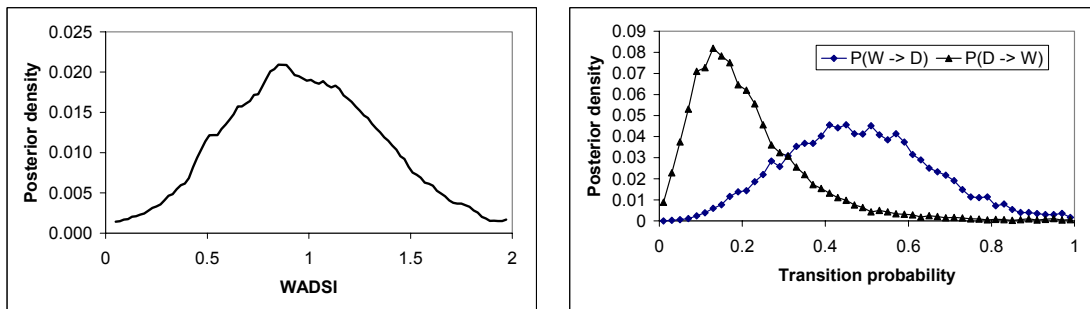
To determine the existence of two states, the empirical distributions of the difference in the means and the WADSI for the starting months having the largest WADSI were examined. The

posterior probability densities of WADSI and transition probabilities for one station from each of the three categories are shown in Figure 2. For Perth (Figure 2a), the two states are not identifiable as the posterior probability density function has its mode at zero. In addition, the two transition probabilities are not well defined and hence it is highly unlikely to have two-state persistence structure. For Meedo and Mardie, the two states are clearly identifiable as the probability of obtaining a value of zero or less for WADSI is zero or very small (Figures 2b and 2c). However, the two transition probabilities are well defined for Meedo (Figure 2b) only and it is highly likely to have two-state persistence structure. On the other hand, the two transition probabilities are not very well defined for Mardie (Figure 2c) and as a result it may not have two-state persistence structure. For a full set of results, the reader is referred to Srikanthan et al. [2002a]. By examining the expected residence times, SSI values and the posterior probability densities of WADSI and transition probabilities, 12 stations were categorised as highly unlikely to have two-state persistence structure. Of the remaining 32 stations, 15 stations indicated that it is highly likely to have two-state persistence and 17 stations could possibly have two-state persistence structure (Table 1). The locations of the three categories of stations are shown in Figure 3. The sites showing the existence of two-state persistence approximately correspond to the areas influenced by the ENSO, however we need to analyse more sites to clearly separate areas of two-state persistence from the rest.

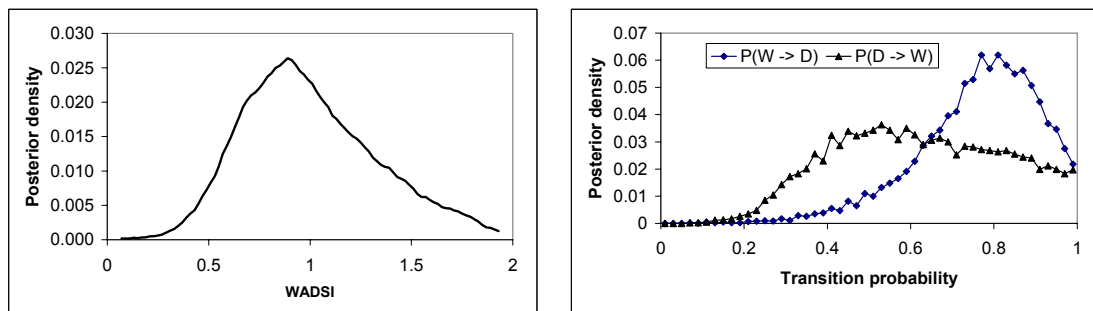
Assuming the fitted model is correct, one thousand replicates each of length equal to the historic record were generated with parameter uncertainty for the 32 stations which exhibited two-state persistence structure. The mean, standard deviation, coefficient of skewness, lag one autocorrelation coefficient, extreme rainfalls and 2, 3, 5, 7 and 10-year low rainfall totals were calculated from each of the 1000 replicates. A full set of results is presented in Srikanthan et al [2002b]. The significance of the departures of the generated values from the corresponding historical values can be objectively assessed using the ratio  $(\text{Hist} - E[\text{Gen}])/\text{sd}(\text{Gen})$  where Hist is the observed statistic, Gen is generated statistic,  $E[\text{Gen}]$  is the mean of the generated statistic and  $\text{sd}(\text{Gen})$  is the standard deviation of the generated statistic. If the ratio is less than -2 or greater than 2, then the observed statistic is not consistent with the model. It can be seen from Table 2 that none of the ratio is outside the range (-2, 2). This shows that the model is consistent with the data.



(a) Highly unlikely to have two-state persistence - Perth



(b) Highly likely to have two-state persistence - Meedo



(c) Possibly have two-state persistence - Mardie

Figure 2. The posterior probability densities of WADSI and transition probabilities.

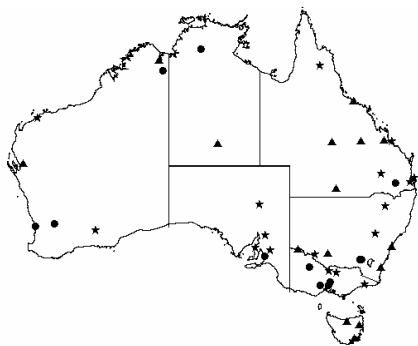


Figure 3. The locations of the stations those are highly unlikely to have two-state persistence (circles), highly likely to have two-state persistence (triangles) and possibly have two-state persistence (stars).

## 5. CONCLUSIONS

The hidden-state Markov model (HSM) which explicitly models the wet and dry states was applied to annual data from 44 rainfall stations located in various parts of Australia. Thirty-two stations exhibited two-state persistence structure. One thousand replicates were generated and several model comparison statistics were calculated. Comparison with the historical parameters shows that the HSM model preserves these statistics. It can be concluded that the HSM model is an improved model to generate annual rainfall data.

Table 2. The ratio of the model evaluation statistics.

Station	Mean	Std dev	Skew	Corre l	Extremes		Low rainfall sums			
					Max	Min	2	3	5	10
Wyndham	-0.20	-0.20	0.35	0.10	0.41	0.47	-0.23	-0.35	-0.47	-0.68
Mardie	0.22	0.22	-0.31	0.09	-0.36	0.82	-1.47	-0.44	-0.85	-0.47
Meedo	-0.09	-0.09	0.43	1.04	0.90	1.35	-0.03	-0.06	-1.31	-0.08
Norseman	-0.16	-0.16	0.54	-0.34	0.33	0.98	0.19	0.69	0.24	0.74
Alice Springs	-0.12	-0.12	0.35	1.24	-0.12	0.86	0.51	0.98	-0.70	-0.53
Marree	-0.11	-0.11	0.35	0.20	0.18	1.24	-0.03	0.44	0.57	-0.80
Orroroo	-0.25	-0.25	0.69	0.23	0.86	1.02	0.27	0.13	-1.18	-1.35
Walleroo	-0.20	-0.20	0.14	-0.26	-0.23	0.85	1.20	1.03	-0.10	0.69
Eudunda	-0.21	-0.21	0.57	0.49	0.45	1.02	0.38	0.01	0.01	0.54
Palmerville	-0.19	-0.19	0.33	-0.03	-0.10	1.05	0.72	0.16	-0.14	-0.54
Kalamia	-0.10	-0.10	0.19	-0.12	-0.64	1.74	0.87	-0.02	-0.73	0.14
Emerald	-0.20	-0.20	0.14	1.00	0.00	0.28	-0.78	0.36	0.56	0.46
Barcaldine	-0.17	-0.17	0.11	0.91	-0.60	-0.02	-0.03	-0.03	0.26	0.30
Cape Capricorn	-0.18	-0.18	0.37	0.12	-0.15	0.64	0.05	-0.25	0.55	-0.51
Rockhampton	-0.12	-0.12	0.45	0.07	0.29	1.71	1.34	0.61	-0.10	-0.53
Cape Moreton	-0.22	-0.22	0.62	0.06	0.51	0.55	-0.72	-0.63	0.56	0.09
Brisbane	-0.15	-0.15	0.45	0.10	0.28	0.74	0.49	0.96	0.70	-0.34
Miles	-0.15	-0.15	0.19	-0.62	-0.12	0.98	-0.44	0.04	-0.37	0.04
Cunnamulla	-0.09	-0.09	0.21	0.23	0.15	1.26	0.58	-1.03	-1.43	-0.24
Wentworth	-0.17	-0.17	0.25	0.52	0.21	0.85	-0.60	-0.97	-1.24	-1.16
Balranald	-0.21	-0.21	0.24	0.73	0.34	0.51	-0.66	-0.84	-0.28	-0.45
Bingara	0.35	0.35	0.20	0.10	0.03	1.18	0.79	0.58	-0.13	-1.38
Mudgee	-0.17	-0.17	0.60	0.30	0.54	1.26	0.65	0.96	0.39	-0.72
Sydney	-0.12	-0.12	0.25	0.41	-0.07	1.35	0.47	0.52	-0.51	0.08
Moruya	-0.13	-0.13	0.30	0.31	-0.12	0.68	0.53	-0.13	-0.08	-0.11
Hay	-0.22	-0.22	0.40	0.93	0.32	0.83	0.13	-0.32	-0.35	-0.79
Tongala	-0.16	-0.16	0.53	1.05	0.36	0.25	-0.98	-1.50	-1.27	-1.22
Caniambo	-0.16	-0.16	0.46	0.15	0.33	0.70	0.12	-0.32	0.46	0.59
Orbost	-0.18	-0.18	0.61	-0.96	0.19	0.91	0.61	0.92	0.87	-0.02
Frankford	-0.20	-0.20	0.31	-0.07	-0.60	0.84	0.97	1.09	0.56	0.71
Fingal	-0.13	-0.13	0.85	0.41	0.97	1.22	0.31	0.04	0.91	-0.41
Sandford	-0.16	-0.16	0.02	-0.11	0.25	-0.53	-0.18	0.59	0.14	0.79

6. REFERENCES

Srikanthan, R. and T. A. McMahon, *Stochastic generation climate data: A review*, CRCCH Report 00/16, Monash University, Clayton, 34pp, 2000.

Srikanthan, R. and T.A McMahon, *Stochastic generation of rainfall and evaporation data*, AWRC Technical Paper No. 84, 301pp, 1985.

Srikanthan, R., M.A. Thyer, G. A. Kuczera, and T. A. McMahon, *Application of hidden state Markov model to Australian annual rainfall data*, CRCCH Working Document (in press), 2002a.

Srikanthan, R., G. A. Kuczera, M.A. Thyer, and T. A. McMahon, *Stochastic generation of*

*annual rainfall data*, CRCCH Technical report (in press), 2002b.

Thyer, M. A., *Modelling long-term persistence in hydrological time series*, Ph D Thesis, University of Newcastle, 2001.

Thyer, M.A. and G. A., Kuczera, *Modelling long-term persistence in rainfall time series: Sydney rainfall case study*, Hydrology and Water Resources Symposium, Institution of Engineer, Australia: 550-555, 1999.

Thyer, M. A. and G. A. Kuczera, *Modelling long-term persistence in hydro-climatic time series using a hidden state Markov model*, *Water Resources Research*, 36(11), 3301-3310, 2000.