Stochastic Generation of Monthly Rainfall Data Using a Nonparametric Approach

R Srikanthan[ad], **A Sharma**[b] **and T A McMahon**[cd]

[a]Hydrology Unit, Bureau of Meteorology, Melbourne, Australia. Email: r.Srikanthan@bom.gov.au

[b]Department of Civil Engineering, University of New South Wales, Sydney, Australia

[c]Department of Civil and Environmental Engineering, University of Melbourne, Melbourne, Australia

[d]Cooperative Research Centre for Catchment Hydrology, Monash University, Clayton, Australia

Abstract: Models to generate stochastic monthly streamflow data can be applied to generate monthly rainfall data, but this presents problems for most regions, which have significant months of no rainfall. Srikanthan and McMahon (1985) recommended the method of fragments to disaggregate the annual rainfall data generated by a first order autoregressive model. The main drawbacks of this approach are the inability to preserve the monthly correlation between the first month of a year and the last month of the previous year and the occurrence of similar patterns from a short length of historic data. Maheepala and Perera (1996) proposed a modification to the selection of fragments that preserves the year-end monthly correlation to improve on the first drawback. Porter and Pink (1991) used synthetic fragments from a Thomas-Fiering monthly model to overcome the second drawback. For sites with considerable number of zero rainfall months, there will be problems with the application of the Thomas-Fiering monthly model to generate the synthetic fragments. Sharma and O'Neil (2002) developed a nonparametric approach to model the inter-annual dependence in monthly streamflows. This method is applied to generate rainfall data from 10 rainfall stations located in various parts of Australia, and results compared to the modified method of fragments. Several annual and monthly statistics were calculated to evaluate the performance of the model and to compare with the modified method of fragments. It was found that both the models preserved the annual and monthly characteristics adequately although it was concluded that the nonparametric approach offered some advantages.

Keywords: *Stochastic generation; Monthly rainfall, Method of fragments; NPL1*

1. INTRODUCTION

Monthly rainfall data are used in the simulation of water resources systems, and in the estimation of water yield from large catchments. In order to assess the system response to climatic variability, long sequences of stochastically generated monthly data are used. Even though the stochastic models to generate monthly streamflow data can be applied to generate monthly rainfall data, this presents problems for most regions, where there are large number of months of no rainfall. In an earlier study, Srikanthan and McMahon (1985) recommended the method of fragments to disaggregate the annual rainfall data generated by a first order autoregressive model. Porter and Pink (1991) reported that the use of the method of fragments resulted in the conspicuous repetition of monthly patterns when generating data much longer than the historical data. They proposed to obtain the monthly fragments from a generated monthly flow sequence from a Thomas-Fiering monthly model. For sites with a considerable number of months of no rainfall, there will be problems to generate the synthetic fragments .using the Thomas-Fiering monthly model. Another problem with the method of fragments is that its inability to preserve the monthly correlation between the first month of a year and the last month of the previous year. Maheepala and Perera (1996) proposed a modification to the selection of fragments that preserves the year-end monthly correlation and applied this modification successfully to generate monthly streamflows using synthetic fragments.

Recently, Sharma and O'Neil (2002) developed a nonparametric approach to model the short-term (month to month) as well as the intern-annual (month to year) dependence in the generated monthly streamflows. In this paper, this method (denoted NPL1, or nonarametric with long-term dependence order 1) is applied to generate rainfall data from 10 rainfall stations located in various parts of Australia. The results from this model are compared with those from a modified method of fragments (MFM).

Table 1. Details of the rainfall stations selected

Number	Name	Latitude (degree)	Longitude (degree)	Length of Record (years)	Annual Mean (mm)	% of months with no rainfall
006036	Meedo	-25.66	114.62	95	216	0 - 68
009034	Perth	-31.95	115.84	116	868	0 - 11
014902	Katherine Council	-14.46	132.26	112	974	0 - 90
015540	Alice Springs PO	-23.71	133.87	113	280	12 - 45
023000	Adelaide	-34.93	138.58	140	530	0 - 5
028004	Palmerville	-16.00	144.08	110	1034	0 - 57
040214	Brisbane	-27.48	153.03	134	1154	0 - 4
066062	Sydney	-33.86	151.20	141	1226	0 - 1
086071	Melbourne	-37.81	144.97	140	657	0 - 1
094061	Sandford	-42.93	147.52	112	578	0 - 1

2. RAINFALL DATA

Ten rainfall stations were selected to cover the Australian continent. The locations of the selected rainfall stations are shown in Figure 1 while the details are shown in Table 1. The number of months of no rainfall, calculated separately for each of 12 calendar months, varies from 0 to 90 %. The large percentage of no rainfall months renders the application of the Thomas-Fiering model very difficult.

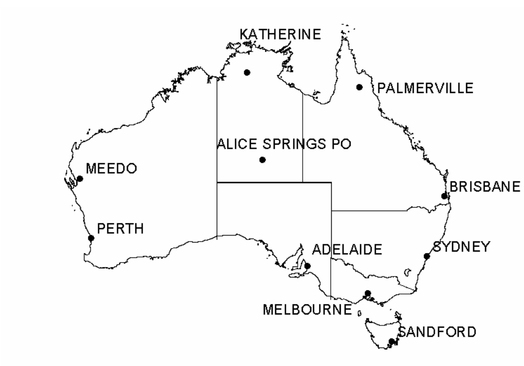

Figure 1. Location of the rainfall stations selected.

3. NONPARAMETRIC MODEL

The nonparametric NPL1 model is designed to preserve both the short term (month to month) as well as the inter-annual (month to year, year to year) dependences in simulated rainfall data. The model used in this study uses only the dependence on the previous month and the previous 12 months rainfall total. The generation of monthly rainfall data proceeds from the following conditional probability density:

$$f(X_t \mid X_{t-1}, Z_t) = \frac{f(X_t, X_{t-1}, Z_t)}{f_m(X_{t-1}, Z_t)} \qquad (1)$$

where X_t rainfall for month t

Z_t previous 12 months rainfall ($= X_{t-1} + X_{t-2} + \ldots + X_{t-12}$)

f_m marginal probability density of X_{t-1} and Z_t

Using a Gaussian kernel function, the above conditional probability density is estimated as Sharma and O'Neill, 2002):

$$\hat{f}(X_t \mid X_{t-1}, Z_t) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda^2 S'}} w_i \exp\left(-\frac{(x_i - b_i)^2}{2\lambda^2 S'} \right) \qquad (2)$$

where λ is a smoothing parameter, known as the "bandwidth" of the known density estimate

S' is a measure of spread of the conditional probability density, expressed as:

$$S' = S_{11} - \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix} \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}$$

where the covariance matrix of the variable set (X_t, X_{t-1}, Z_t) is written as

$$\mathrm{Cov}(X_t, X_{t-1}, Z_t) = \begin{bmatrix} S_{11} & S_{12} & S_{1z} \\ S_{12} & S_{22} & S_{2z} \\ S_{1z} & S_{2z} & S_{zz} \end{bmatrix}$$

w_i is the weight associated with each kernel that constitutes the conditional probability density:

$$w_i = \frac{\exp\left\{ -\frac{1}{2\lambda^2} \begin{bmatrix} X_{t-1} - x_{i-1} \\ Z_t - z_i \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} X_{t-1} - x_{i-1} \\ Z_t - z_i \end{bmatrix} \right\}}{\sum_{j=1}^{n} \exp\left\{ -\frac{1}{2\lambda^2} \begin{bmatrix} X_{t-1} - x_{j-1} \\ Z_t - z_j \end{bmatrix}^T \begin{bmatrix} S_{22} S_{2z} \\ S_{2z} S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} X_{t-1} - x_{j-1} \\ Z_t - z_j \end{bmatrix} \right\}}$$

b_i is the conditional mean associated with each kernel:

$$b_i = x_i + \begin{bmatrix} S_{12} \\ S_{1z} \end{bmatrix}^T \begin{bmatrix} S_{22} & S_{2z} \\ S_{2z} & S_{zz} \end{bmatrix}^{-1} \begin{bmatrix} X_{t-1} - x_{i-1} \\ Z_t - z_i \end{bmatrix}$$

x_t and z_t represent the rainfall for month t and the sum of the prior 12 monthly rainfalls respectively.

The conditional probability density estimate in Equation (2) can be viewed as consisting of n kernels having relative areas equal to weight w_i, centred at b_i, and having a spread proportional to S'. Each of these is a slice of the trivariate kernels that constitute the joint probability density of (X_t, X_{t-1}, Z_t), along the conditioning plane specified by (X_{t-1}, Z_t). The weight w_i depends directly on how far the kernel is from the conditioning plane. A small weight implies that the kernel is far from the conditioning plane and does not make up a significant proportion of the conditional density estimate. On the other hand, a large w_i implies the kernel is close to the conditioning plane and constitutes a significant portion of the conditional density estimate. Consequently, data generation will proceed with more emphasis given to the observed data points lying closer to the conditioning plane and less emphasis given to the data points that lie farther away. Readers are referred to Sharma et al. (1997) and Sharma and O'Neal (2002) for details on the above multivariate kernel density estimator.

Monthly rainfall data is generated by following the steps below:

Step 1. Estimate the bandwidth and the covariances S_{11}, S_{12}, S_{1z}, S_{22}, S_{2z}, S_{zz}.

Step2. Start the data generation by arbitrarily assigning values to X_{t-1} and Z_t.

Step 3. Given X_{t-1} and Z_t, estimate the weight w_i associated with each kernel.

Step 4. Pick a data point with probability w_i.

Step 5. The new value of X_t can be obtained as $X_t = b_i + \lambda(S')^{1/2}W_i$, where W_i is a Gaussian random variable with zero mean and unit standard deviation.

Step 6. Repeat steps 3 to 5 until the required length of data is generated.

The first few generated values are discarded to reduce the effect of the arbitrary initialisation used. In this study, the first 16 years of generated data are discarded.

To overcome the problem of generating negative rainfall, a variable kernel (Scott, 1992) has been used for data points close to the zero rainfall boundary. The bandwidth of the conditional kernel used for generating a new rainfall in step 5 is reduced depending on the distance of its centre (b_i) from the zero rainfall boundary. The modified step 5 is as follows:

Step 5a. Estimate a transformed band width λ' such that

$$\begin{aligned} \lambda &= \lambda && \text{if } F_{N(bi,\lambda^2 S')}(X_t \le 0) \le \alpha \\ &= \lambda' && \text{if } F_{N(bi,\lambda^2 S')}(X_t \le 0) > \alpha \end{aligned}$$

where $F_{N(bi,\lambda'^2 S')}(X_t \le 0) = \alpha$ and $F_{N(\mu,\sigma^2)}$ is the cumulative probability of a Normal distribution with mean μ and variance σ^2, with the bandwidth being transformed to λ' if for the selected Normal kernel, the probability of the rainfall X_t being less than or equal to zero, is estimated to be greater than a specified threshold α.

Step 5b. Generate a new value of X_t using $X_t = b_i + \lambda'(S')^{1/2}W_i$, where W_i is a Gaussian random variate with zero mean and unit variance.

Step 5c. Repeat step 5b if the generated value of X_t is less than zero, until a positive value results.

4. MODIFIED METHOD OF FRAGMENTS

The observed monthly rainfalls are standardised year by year so that the sum of the monthly rainfall in any year equals unity. This is carried out by dividing the monthly rainfall in a year by the corresponding annual rainfall. By doing so, from a record of n years, one will have n sets of fragments of monthly rainfalls. The appropriate monthly fragments for a given year, k, is selected by considering the closeness of the generated annual rainfall data and the monthly rainfall for the last month of the previous year of the already disaggregated data to the corresponding historical values (Maheepala and Perera, 1996). This is achieved by selecting the monthly fragments of a year, i, in the generated monthly series that produces a minimum value for α_i, which is defined below:

$$\alpha_i = \left(\frac{x'_k - x_i}{s_x} \right)^2 + \left(\frac{y'_{k-1} - y_{i-1}}{s_y} \right)^2 \tag{3}$$

where x'_k = generated annual rainfall for year k

x_i = historical annual rainfall for year i

s_x = standard deviation of the annual rainfall

y'_{k-1} = disaggregated monthly rainfall for the last month of year k-1

y_{i-1} = historical monthly rainfall for the last month of year i-1

s_y = standard deviation of the monthly rainfall for the last month of a year

The generated annual rainfalls are disaggregated by multiplying the generated rainfall by each of the 12 fragments to give 12 generated monthly rainfalls. In this study, the generated annual rainfall is obtained from a first order autoregressive model with parameter uncertainty (Srikanthan et al., 2002a).

5. MODEL EVALUATION

The model evaluation is carried out at annual and monthly time periods. The parameters used to evaluate the annual level are the annual mean, standard deviation, coefficient of skewness, lag one autocorrelation coefficient, extreme events, adjusted range and low rainfall sums. The maximum and minimum rainfall depth occurring in the historic record and in each of the generated sequences are taken as the extreme events. Rank one 2-, 3-, 5-, 7- and 10-year low rainfall sums are used. The adjusted range (R) is obtained from

$$R = \max \{D_k\} - \min \{D_k\} \quad k = 1, 2, \ldots , n \qquad (4)$$

where $D_k = \sum_{t=1}^{k} (x_t - \overline{x})$ and $\overline{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$

The extreme events, low rainfall sums and the adjusted range are standardised by dividing by the historical mean annual rainfall.

At the monthly level, the monthly means, standard deviations, coefficients of skewness, serial correlation coefficients between successive months, maximum and minimum monthly rainfall and relative frequency of no rainfall months are used. The serial correlation coefficient corresponding to month t is the correlation between the t and (t-1) monthly pairs. The extreme events are divided by the historical overall monthly mean (i.e. annual mean divided by 12).

6. DISCUSSION OF RESULTS

The annual and monthly parameters mentioned in section 5 were estimated from each of the 100 replicates for both the models. For each parameter, the mean, median, 2.5-, 25-, 75- and 97.5- percentile values are obtained for comparison Due to lack of space, only part of the results are presented in this paper. For a full set of results, the reader is referred to Srikanthan et al. (2002b).

6.1. Annual Parameters

The average values of annual mean, standard deviation, coefficient of skewness and lag one autocorrelation coefficient from 100 replicates are presented in Table 2.

Both models reproduced the annual mean, standard deviation and coefficient of skewness for all the stations. The lag one autocorrelation was small for all the stations and both the models preserved it except for Meedo and Alice Springs. For Alice Springs, the nonparametric model gave a small average value while the method of fragments gave a satisfactory value. However, the observed correlation is within the 95% confidence limits. Table 3 shows that the extreme events are reproduced. The adjusted range and the low rainfall sums are also reproduced (Table 4).

6.2. Monthly Parameters

The mean of each of the seven monthly parameters (Section 5) estimated from the 100 replicates were compared with the corresponding historical values. The comparison for Meedo is given in Table 5. Both models reproduced the monthly mean, standard deviation, coefficient of skewness and correlation well. The modified method of fragments using historical fragments preserved the monthly correlation for the first month in the year. Both models preserved the extreme events and the relative frequency of no rainfall months.

The percentile values of the parameters from the generated data and the historical parameters were plotted for both the models (Srikanthan et al. 2002b). For the first five parameters, the historical values are all within the 95 % confidence limits. The method of fragments produced more extreme rainfall totals than the historical minimum rainfalls for the low rainfall months. The nonparametric model performs better than the method of fragments with regard to minimum rainfall. (Figure 2). Both models satisfactorily reproduced the relative frequency of no rainfall months.

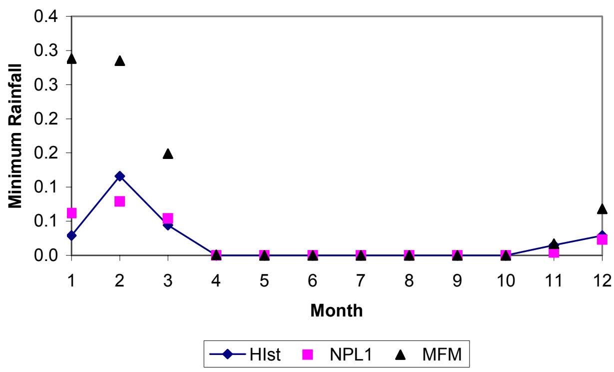

Figure 2. Comparison of historical and generated monthly minimum rainfall for Palmerville.

Table 2. Comparison of historical and generated annual mean, standard deviation, skewness and lag one autocorrelation coefficient.

	Mean			Standard deviation			Skewness			Lag 1 autocorrelation		
Station	Hist	MFM	NPL1	Hist	MFM	NPL1	Hist	MFM	NPL1	Hist	MFM	NPL1
Meedo	214	220	216	102	96	102	1.105	0.878	0.873	-0.032	0.258	0.059
Perth	868	874	876	162	164	168	0.103	0.639	0.210	-0.051	-0.019	0.030
Katherine	975	973	983	251	261	275	0.046	0.153	0.219	0.069	0.084	0.066
Alice	282	285	279	145	146	144	1.498	1.401	1.182	0.289	0.230	0.068
Adelaide	530	531	532	108	107	115	0.057	0.610	0.232	-0.010	0.076	0.045
Palmerville	1035	1043	1041	302	327	321	0.661	0.681	0.474	0.070	0.062	0.057
Brisbane	1154	1161	1162	358	384	356	0.592	0.953	0.659	0.016	0.054	0.106
Sydney	1225	1223	1230	331	342	347	0.607	0.825	0.524	0.101	0.129	0.070
Melbourne	659	657	660	127	125	132	0.001	0.571	0.257	0.012	0.104	-0.064
Sandford	576	578	583	131	130	130	0.432	0.748	0.404	0.011	0.105	0.037

Table 3. Comparison of historical and generated annual maximum, minimum and adjusted range.

	Maximum			Minimum			Range		
Station	Hist	MFM	NPL1	Hist	MFM	NPL1	Hist	MFM	NPL1
Meedo	2.619	2.513	2.560	0.305	0.263	0.220	5.817	6.233	5.740
Perth	1.542	1.620	1.533	0.586	0.630	0.561	3.571	2.386	2.478
Katherine	1.615	1.717	1.771	0.451	0.373	0.355	4.344	3.549	3.715
Alice	3.207	3.131	2.838	0.191	0.257	0.170	8.205	7.723	6.754
Adelaide	1.485	1.673	1.628	0.487	0.592	0.506	2.275	3.000	3.148
Palmerville	2.027	2.031	1.923	0.433	0.401	0.369	3.595	3.998	3.855
Brisbane	1.944	2.247	1.969	0.357	0.414	0.389	6.490	4.786	4.599
Sydney	1.790	1.992	1.874	0.476	0.473	0.421	5.785	4.332	4.164
Melbourne	1.468	1.632	1.545	0.504	0.599	0.544	2.239	2.840	2.628
Sandford	1.603	1.759	1.657	0.566	0.576	0.537	2.846	3.061	2.986

Table 4. Comparison of historical and generated 2-, 5- and 10-year low rainfall sums.

	2-year			5-year			10-year		
Station	Hist	MFM	NPL1	Hist	MFM	NPL1	Hist	MFM	NPL1
Meedo	0.772	0.751	0.739	2.797	2.813	2.817	7.515	7.061	7.061
Perth	1.439	1.449	1.371	4.012	4.124	4.043	8.803	8.849	8.748
Katherine	1.173	1.053	0.995	3.605	3.544	3.459	8.232	8.038	8.041
Alice	0.716	0.705	0.628	2.446	2.584	2.742	6.320	6.535	6.868
Adelaide	1.389	1.367	1.296	4.020	3.958	3.878	8.369	8.559	8.483
Palmerville	1.259	1.075	1.033	3.389	3.500	3.438	7.590	7.987	7.914
Brisbane	0.976	1.066	1.072	3.607	3.422	3.443	7.732	7.829	7.866
Sydney	1.262	1.142	1.139	3.650	3.534	3.585	7.945	7.909	8.086
Melbourne	1.313	1.376	1.320	4.107	3.962	4.008	8.899	8.603	8.688
Sandford	1.323	1.323	1.361	3.912	3.893	3.947	8.730	8.470	8.591

Table 5. Comparison of historical and generated monthly parameters for Meedo.

Parameter	Model	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Mean (mm)	Hist	15.6	20.3	18.8	13.6	34.8	42.1	34.0	18.5	5.1	5.8	3.2	3.6
	MFM	19.8	20.6	21.6	12.0	32.7	44.2	34.6	18.7	5.4	4.4	3.1	2.6
	NPL1	16.4	21.8	18.4	12.1	34.8	42.2	33.0	18.9	5.2	6.2	3.1	3.8
Std Dev (mm)	Hist	32.4	29.6	27.8	23.1	36.6	31.2	35.0	19.1	6.3	13.1	8.8	10.2
	MFM	38.7	26.6	31.0	16.8	36.3	30.2	32.8	17.5	7.4	10.0	6.6	6.7
	NPL1	33.0	30.5	27.8	18.1	37.5	32.2	34.9	19.1	6.5	12.5	7.9	10.1
Skew	Hist	4.01	2.36	2.31	3.53	1.62	0.96	2.10	1.41	1.72	3.70	4.79	5.51
	MFM	3.28	1.77	1.96	2.38	1.63	0.66	1.98	1.31	1.97	4.11	3.70	4.00
	NPL1	3.49	2.11	2.20	2.75	1.60	0.91	1.96	1.25	1.64	3.31	4.17	4.19
Correl.	Hist	-0.11	0.13	0.02	0.02	0.06	0.07	-0.03	-0.05	0.03	0.03	0.30	0.04
	MFM	-0.12	0.17	-0.02	0.08	-0.04	0.09	0.08	-0.11	0.06	0.09	-0.02	-0.04
	NPL1	-0.13	0.15	0.02	0.03	0.06	0.04	-0.03	-0.06	0.05	0.02	0.13	0.04
Max	Hist	12.1	8.3	8.5	7.5	9.3	7.3	10.2	5.0	1.8	4.2	3.4	4.6
	MFM	11.6	7.0	8.1	5.2	9.2	7.1	9.3	4.5	1.9	3.4	2.2	2.4
	NPL1	10.7	7.9	7.8	6.0	9.1	7.7	9.8	4.6	1.7	3.8	2.8	3.5
Min	Hist	0	0	0	0	0	0.072	0	0	0	0	0	0
	MFM	0	0	0	0	0	0.09	0.001	0	0	0	0	0
	NPL1	0	0	0	0	0	0.03	0	0	0	0	0	0
No rainfall (%)	Hist	35.1	28.7	36.2	33.0	10.6	0	5.3	10.6	35.1	41.5	63.8	68.1
	MFM	35.8	28.6	35.5	32.9	13.2	0	5.5	7.1	37.7	40.6	58.8	69.0
	NPL1	37.2	29.4	38.6	32.6	11.3	0.2	5.7	10.8	33.9	42.1	65.6	68.6

7. CONCLUSIONS

The method of fragments and the nonparametric model, developed by Sharma and O'Neill (2001), were evaluated using monthly rainfall data from 10 rainfall stations located in different parts of Australia. The evaluation was carried out at the annual and monthly level. Both models were found to preserve the annual and monthly characteristics adequately. However, the nonparametric model has the following advantages over the method of fragments.

- It does not repeat the same yearly patterns as in the method of fragments.
- Minimum rainfalls were generated better than method of fragments.
- It eliminates the need for having a separate model that simulates the annual rainfall values.
- It also eliminates the need for choosing a starting month for forming annual totals as it uses only the monthly data.

8. REFERENCES

Maheepala, S. and C. J. C., Perera, Monthly hydrologic data generation by disaggregation. *J. Hydrol*., 178, 277-291, 1996.

Porter, J. W. and B. J., Pink, *A method of synthetic fragments for disaggregation in stochastic data generation*. Hydrology and Water Resources Symposium, Institution of Engineers, Australia, 187-191 1991.

Scott, D. W., Multivariate density estimation: Theory, Practice and Visualisation. John Wiley & Sons Inc., New York, 1992.

Sharma, A., D. G. Tarboton, and U. Lall, Streamflow simulation: A nonparametric approach. *Water Res. Res. 33* (2), 291-308, 1997.

Sharma, A. and R.O'Neill, A nonparametric approach for representing interannual dependence in monthly streamflow. *Water Res. Res.* 138(7), 5-1:5-10. 2002.

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

Srikanthan, R., T. A McMahon, M. A. Thyer, and G.Kuczera,. Stochastic generation of annual rainfall data. Technical Report 02/6, CRC for Catchment Hydrology, Monash University, Clayton, 53pp, 2002a

Srikanthan, R., T. A McMahon, and A. Sharma, Stochastic generation of monthly rainfall data. Technical report 02/8, CRC for Catchment Hydrology, Monash University, Clayton, 40p, 2002b.