30th Hydrology and Water Resources Symposium
4 - 7 December 2006
Launceston, TAS

Stochastic Modelling of Annual Rainfall Data

R. Srikanthan[1], M.C. Peel[2], G.G.S. Pegram[3], and T.A. McMahon[2]
1 Hydrology Unit, Bureau of Meteorology, Melbourne, Victoria, 3000,
r.srikanthan@bom.gov.au
2 Civil & Environmental Engineering, University of Melbourne, Victoria, 3010
3 Civil Engineering, University of KwaZulu-Natal, DURBAN, South Africa.

Abstract: Rainfall data are generally required in computer simulations of rainfall-runoff processes, crop growth and water supply systems. The length of historical climate data is usually not long enough to describe the complete range of variability that might be experienced during the life of a water resources or agricultural project. Using the statistical characteristics of historical data, it is possible to generate many sequences of data that better represent the climatic variability. In developing the stochastic models, the data are generally assumed stationary in the broad sense and any long-term fluctuations in the data are ignored. Typically, only in monthly, daily and sub-daily models, is the seasonal variation within a year considered explicitly in stochastic models. However, there is a growing interest and concern about the role of interdecadal variability in climate and its influence on rainfall. One approach is to identify any long-term fluctuations in the observed rainfall and model them explicitly. Empirical Mode Decomposition (EMD) was used to identify any low frequency fluctuations in annual rainfall data from 44 sites in Australia. The results did not allow easy identification of low frequency fluctuations in the data. As a means of aiding interpretation of the EMD results, the following ploy was adopted. The AR1 model, the most widely used model for the generation of annual rainfall data, was used to generate stochastic data based on the statistics of the observed sequences and the EMD analysis was performed on the stochastic data sets. The results of the analysis comparing both the historical and generated data showed that, in general, both the data sets have similar low frequency characteristics except for Perth.

Keywords: Rainfall, stochastic modelling, empirical mode decomposition, interdecadal variability.

1. INTRODUCTION

Climate data, particularly rainfall data, are a major input to water resources and agricultural modelling systems. As the historical record only provides a single realisation of the underlying climate variability, stochastically generated data can be used to assess the impact of climate variability on water resources and agricultural systems. Even though generated annual rainfall data have little direct application, 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 their 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 become the input to various disaggregation schemes.

Srikanthan and McMahon (2001) carried out a review of stochastic generation of climate data and recommended a lag one autoregressive [AR(1)] time series model or the Hidden State Markov (HSM) model to generate annual rainfall data. These two models were applied to annual rainfall data from 44 sites across Australia and the results showed that the AR(1) model is generally adequate for all the stations (Srikanthan et al. 2002). However, there is a growing interest and concern about the role of interdecadal variability in climate and its influence on rainfall. The HSM model implicitly takes this long term variability into account by having two climate states, namely, dry and wet states. The residence time in each of the states is determined by the transition probabilities. Another approach is to identify any long-term fluctuations in the observed rainfall and model them explicitly. In this paper, Empirical Mode Decomposition (EMD) was used to identify any low frequency fluctuations in annual rainfall data from 44 sites in Australia. The results did not allow easy identification of low frequency fluctuations in the data. As a means of aiding interpretation of the EMD results, the following

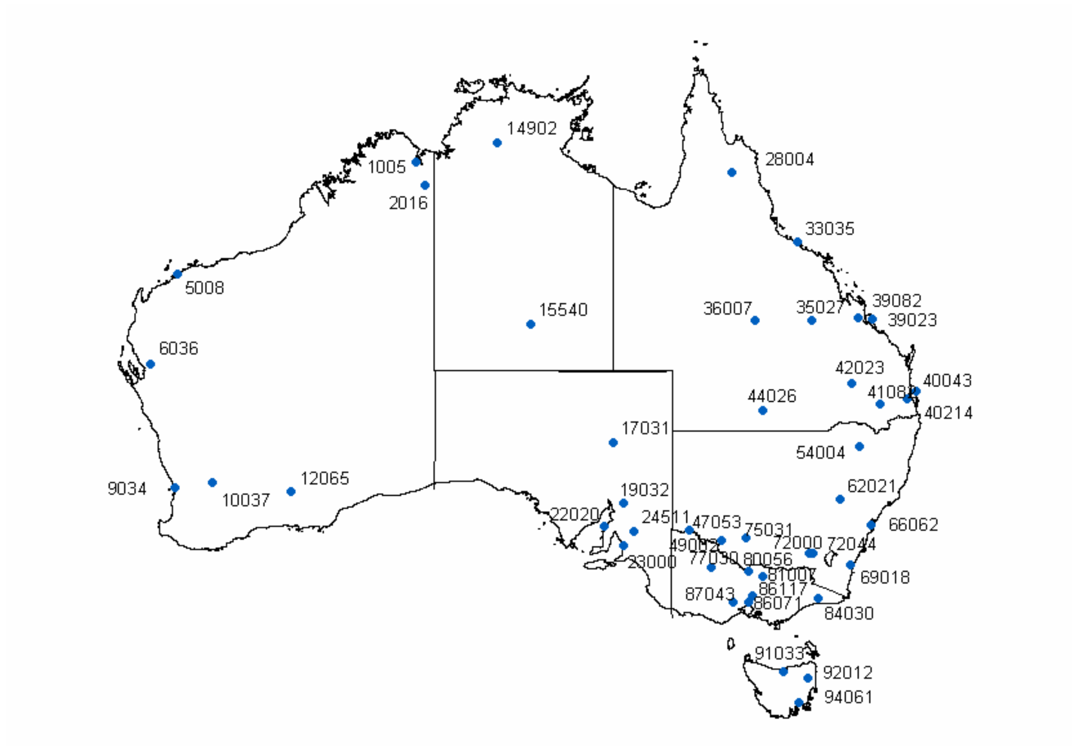

Figure 1. The locations of the 44 selected rainfall stations.

ploy was adopted. The AR1 model, the most widely used model for the generation of annual rainfall data, was used to generate stochastic data and the EMD analysis was performed on the stochastic data. The results of the analysis comparing both the historical and generated data are presented in the paper.

2. RAINFALL DATA

Forty-four rainfall stations with long records were selected. The locations of the stations are shown in Figure 1. The data length varies from 69 to 143 years. In order to have at least one rainfall station in each of the focus catchments of the CRC for Catchment Hydrology (CRCCH), the station Tongala (80056) in Victoria was included, although it has only 69 years of data. The mean annual rainfall of the stations varies from 164 to 1550 mm.

3. EMPIRICAL MODE DECOMPOSITION

Empirical Mode Decomposition (EMD), developed by Huang *et al* (1998), is a form of adaptive time series decomposition. Traditional forms of spectral analysis, like Fourier, assume that a time series (either linear or nonlinear) can be decomposed into a set of linear components. However, as the degree of non-periodic behaviour and non-stationarity in a time series increases, the set of linear components describing that time series increases substantially when using Fourier techniques. In the physical sciences, time series are often non-periodic, more stochastic and even non-stationary, so Fourier based spectral analysis techniques often produce large sets of physically meaningless harmonics when applied to these problems (Huang *et al* 1999). In contrast the EMD method does not assume a time series is linear or stationary prior to analysis, it lets the data speak for themselves. EMD adaptively decomposes a time series into a set of independent intrinsic mode functions (IMFs) and a residual component. When the IMFs and residual are summed together they form the original time series. The IMFs and residual component may display linear and or non-linear behaviour (amplitude and frequency modulation) depending on the nature of the time series being studied. In comparisons of EMD analysis with Fourier Series fitting of hydrological time series, in order to accommodate the non-linearities evident in the IMFs (frequency and amplitude modulation) we found that up to 10 times as many Fourier terms were needed (compared to EMD) for a satisfactory decomposition. The reason is that Fourier analysis is an averaging methodology assuming stationarity while EMD makes no such implicit assumption.

There are four steps (Peel, et al 2005) involved in estimating the IMFs. First, the local extrema, both maxima and minima, are identified. Second, cubic spline curves are fitted to the sequences of maxima and minima. Third, the mean of the two cubic splines is taken and, fourth, the mean of the cubic splines is subtracted from the original time series and the remainder forms a residual. The residual is the first estimate of the first IMF. An IMF must satisfy two conditions (Huang *et al* 1998):

(i) the number of extrema (sum of maxima and minima) and the number of zero crossings must be equal or differ by one, and
(ii) the mean of the cubic splines must be equal to zero at all points.

3.1 Sifting to obtain the first IMF

The mean of the upper and lower envelopes is designated as m_1, and the difference between the data and m_1 is the first 'Proto-Intrinsic Mode Function', h_1. Ideally, h_1 should be an IMF, for the construction of h_1 described above should have made it satisfy all the requirements of an IMF. Yet, even if the fitting is perfect, a gentle hump on a slope can be amplified to become a local extremum by changing the local zero from a rectangular to a curvilinear co-ordinate system. After the first round of sifting, the hump may become a local maximum. Therefore, the sifting process should be applied repeatedly. Once the sifting converges (see Peel et al. 2005 for stopping conditions) h_1 is subtracted from the data to obtain the first residual, r_1.

3.2 Obtaining subsequent IMFs

The four steps and recursive sifting process are then applied to the residual r_1 of the observed time series to obtain the second IMF. As each subsequent IMF is identified it is subtracted from the last available residual, i.e. the difference between the observed time series and the sum of the previous IMFs. The process of IMF identification continues until all IMFs are extracted and a final residual remains. The final residual will be a constant, a monotonic trend or a fluctuation with a cycle longer than the period of record. The average period of each IMF can be calculated by dividing twice the sample size (2 x N) by the number of zero crossings.

4. APPLICATION OF EMD TO ANNUAL RAINFALL DATA

4.1 Historical data

A summary average period and variance of each IMF are shown in Table 1 for the 44 sites. By examining the IMFs of individual rainfall stations, it was not possible to identify any well defined cycles. The IMFs for Melbourne are shown in Figure 2 which are typical of most

Table 1. Summary of the periods and variances of the IMFs obtained for the 44 rainfall stations

(a) Period (years)

		IMF1	IMF2	IMF3	IMF4	IMF5
All	Mean	3.03	6.51	15.03	30.76	51.56
	Std Dev	0.17	0.56	2.65	6.00	8.24
WA	Mean	3.04	6.88	14.31	33.30	47.85
	Std Dev	0.18	0.80	2.95	9.02	6.58
NT	Mean	3.01	7.13	14.93	28.00	
	Std Dev	0.14	0.48			
SA	Mean	2.98	6.36	13.40	30.05	36.67
	Std Dev	0.04	0.57	1.06	5.53	
QLD	Mean	2.99	6.27	16.18	31.40	53.93
	Std Dev	0.16	0.41	3.15	4.19	2.77
NSW	Mean	3.09	6.70	15.43	32.94	52.80
	Std Dev	0.12	0.52	1.76	6.12	
VIC	Mean	3.04	6.50	15.87	31.96	51.88
	Std Dev	0.18	0.48	2.74	4.91	4.57
TAS	Mean	2.91	6.24	12.85	22.73	62.83
	Std Dev	0.18	0.45	2.74	5.41	7.25

(b) Variance (%)

		IMF1	IMF2	IMF3	IMF4	IMF5	Residual
All	Mean	54.57	18.41	13.98	6.71	3.28	5.60
	Std Dev	6.63	5.96	5.66	3.85	2.40	3.54
WA	Mean	55.74	21.49	8.38	6.58	4.25	6.59
	Std Dev	7.15	6.05	3.83	5.73	4.99	2.37
NT	Mean	51.03	25.43	10.32	10.29		2.94
	Std Dev	15.75	11.01	4.06	2.17		1.49
SA	Mean	54.89	18.67	13.96	7.03	5.55	3.22
	Std Dev	5.50	2.24	4.34	4.04	3.58	2.43
QLD	Mean	55.70	16.20	15.31	5.75	3.00	6.74
	Std Dev	5.24	5.79	4.93	3.61	2.45	4.52
NSW	Mean	52.79	18.19	14.71	6.93	2.54	7.09
	Std Dev	7.95	5.87	3.34	2.27		3.27
VIC	Mean	52.95	16.82	18.21	6.80	2.60	4.71
	Std Dev	6.86	6.95	8.67	3.88	1.53	3.08
TAS	Mean	58.62	18.53	12.60	6.43	2.14	2.39
	Std Dev	1.91	3.59	2.47	5.99	0.73	2.68

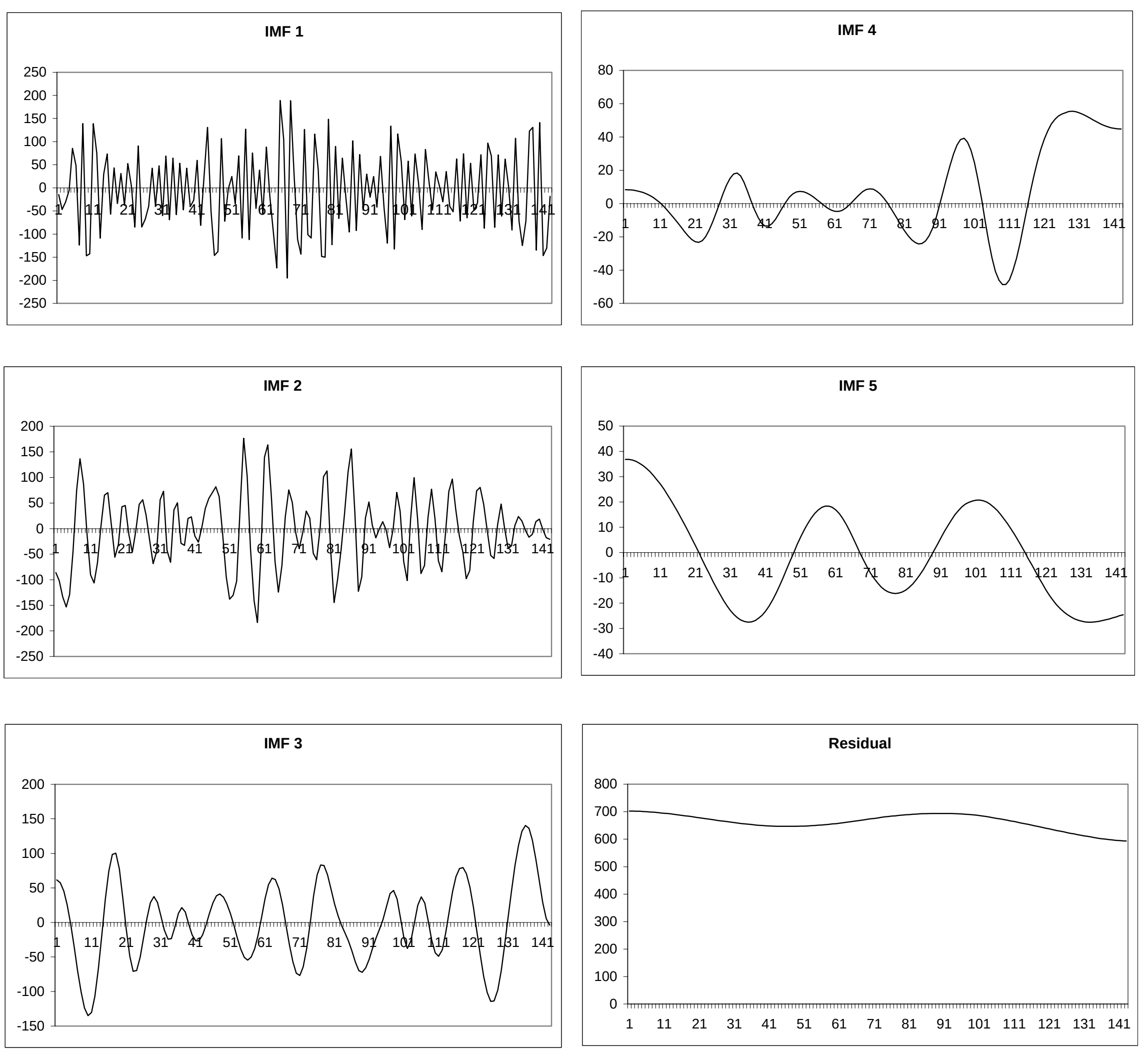

Figure 2. The IMFs and the residual for Melbourne.

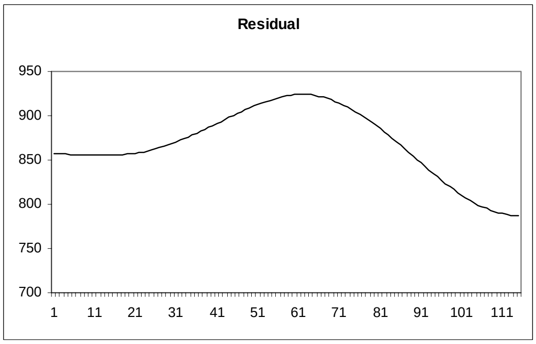

Figure 3. The residual series for Perth.

stations. The residual for Perth has a strange pattern and is shown in Figure 3. Since it was difficult to identify any consistent cycles in the annual rainfall data, we decided to examine the generated data to see whether or not similar cycles are present.

4.2 Generated data

One hundred replicates, each of length equal to the historical record, were generated using an AR(1) model using the statistics estimated from the historical data and the EMD analysis was carried out on each. The periods and the variances were averaged and compared with the corresponding historical values. Due to lack of space, a comparison of the periods of IMFs for only eight stations is presented in Table 2. It can be seen from the table that, except for Perth, the stations have a good resemblance between the historical and generated values. In addition, the variances were plotted against the periods of the IMFs for the historical and generated replicates and superimposed. If the generated data are similar to the historical data, then the historical point should lie within the cloud formed by the generated data.

It can be seen from Figure 4 that, except for Perth (and to a lesser extent Sydney), the historical points lie within the cloud of the points obtained from the generated data and, in addition, the number of IMFs is the same in historical and generated sets. For Perth, the historical data appear to have longer cycles than the generated data, especially for the last two IMFs. This means that the AR(1) model is not adequate for Perth and a higher order model might be necessary. It is worth noting that an ARMA(2,2) model was identified for Perth using Akaike Information Criterion in an earlier study (Srikanthan et al., 2002). But the statistics used to evaluate the models could not discriminate between the AR(1) and ARMA(2,2) model results.

Table 2. Comparison of period (years) of IMFs.

		Period (years)					
		IMF1	IMF2	IMF3	IMF4	IMF5	IMF6
Sydney	Hist	3.1	6.2	15.7	26.6	53.2	
	Gen	3.0	6.5	13.9	31.1	56.8	70.0
Melbourne	Hist	2.8	6.0	13.6	28.6	57.2	
	Gen	3.0	6.5	13.7	30.4	56.5	95.3
Adelaide	Hist	3.0	6.8	13.9	34.8		
	Gen	2.9	6.3	13.4	29.6	55.8	69.5
Alice Springs	Hist	2.9	6.8	14.9	28.0		
	Gen	3.0	6.6	14.3	32.3	47.8	56.0
Katherine	Hist	3.1	7.5	14.9	28.0		
	Gen	3.0	6.4	14.0	30.8	51.5	
Perth	Hist	2.9	7.2	19.2	46.0		
	Gen	2.9	6.3	13.5	30.4	48.8	57.5
Brisbane	Hist	3.1	6.2	15.7	26.6	53.2	
	Gen	3.0	6.4	13.7	30.6	55.5	
Sandford	Hist	3.0	6.0	12.3	20.2		
	Gen	3.0	6.4	13.8	30.6	48.0	

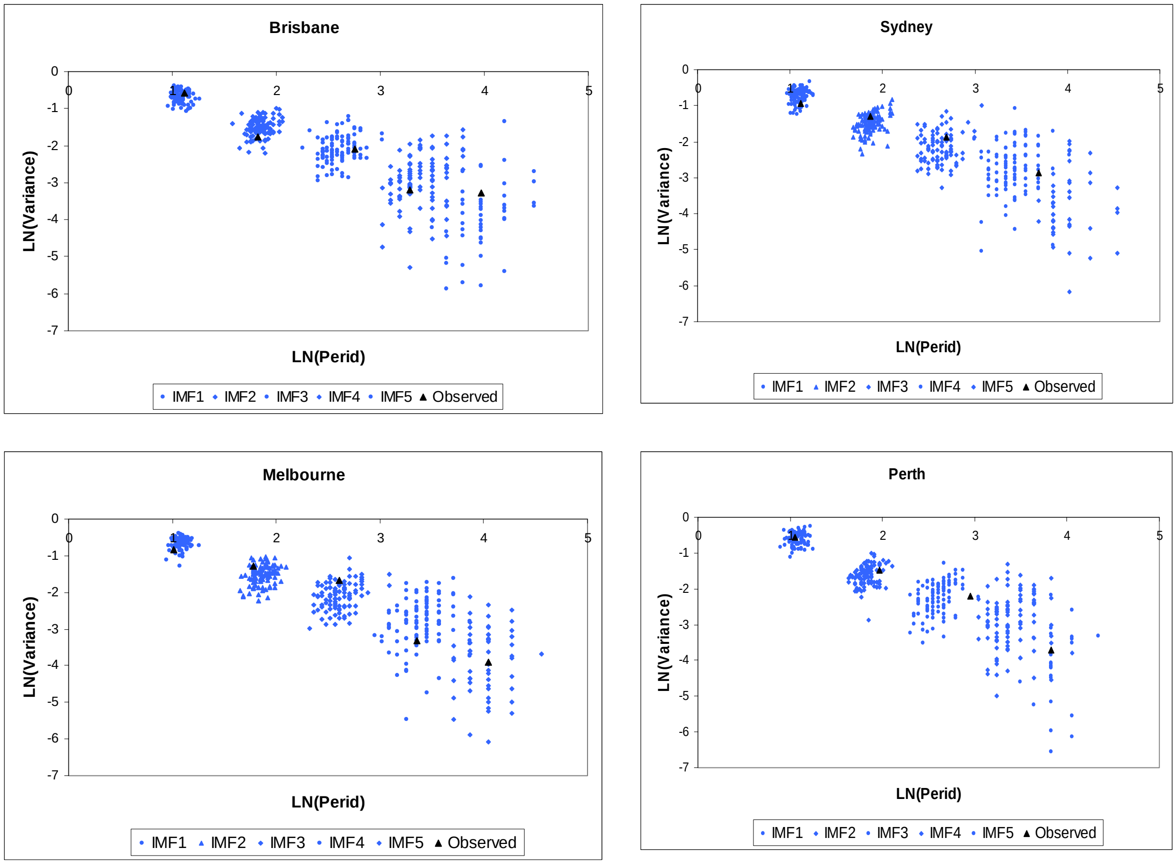

Figure 4. Plots of variance against period for four stations.

5. CONCLUSIONS

EMD analysis was performed on annual rainfall data from 44 stations located in different parts of Australia to determine the presence of long pseudo-periodic fluctuations in the IMFs so that these IMFs might be explicitly incorporated into stochastic data generation models. From the results, it was not possible to obtain any consistent fluctuations for the annual rainfall data. As an alternative, it was decided to apply the EMD analysis to stochastically generated data to determine the IMFs and these were compared with the historical ones. The results show that the IMFs were similar for all the stations except Perth. Hence, excluding Perth, an AR(1) model is generally adequate for Australian annual rainfall data.

6. REFERENCES

Huang, N.E., Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung and H.H. Liu (1998), The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proceedings of the Royal Society London A* 454(1971), 903-995.

Huang, N.E., Z. Shen and S.R. Long (1999), A new view of nonlinear water waves: the Hilbert spectrum, *Annual Review of Fluid Mechanics* 31, 417-457.

Peel, M.C., G. E. Amirthanathan, G. G. S. Pegram, T. A. McMahon, and F. H. S. Chiew (2005), Issues with the Application of Empirical Mode Decomposition Analysis. MODSIM2005, 12-15 Dec 2005, Melbourne, 1681- 1687.

Srikanthan, R. and T. A. McMahon (2001) Stochastic generation of annual, monthly and daily climate data: A review. Hydrology and Earth Systems Sciences 5(4): 653-656.

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