Generation of synthetic daily rainfall for thirteen locations in Australia using a nonparametric approach

T.I. Harrold[a], **A. Sharma**[b], **and S. Sheather**[c]

[a]Research Institute for Humanity and Nature, Kyoto, Japan (harrold@chikyu.ac.jp)

[b] School of Civil and Environmental Engineering, The University of New South Wales, Australia (a.sharma@unsw.edu.au)

[c] Australian Graduate School of Management

Abstract: Many existing methods of daily rainfall generation assume that daily rainfall depends exclusively on the rainfall that occurred in the past one, two, or three days, an assumption that results in an under-representation of variability at longer time-scales. Such reduced variability affects the representation of sustained wet spells and droughts, features that are of great interest in catchment planning and management. The approach of Harrold et al. (2002) is designed to give a better representation of rainfall variability. This paper applies the approach of Harrold et al. (2002) to daily rainfall from 13 locations in Australia. These locations provide a broad range of rainfall regimes, ranging from temperate to semi-arid to tropical. Conclusions are drawn regarding the flexibility and ease of application of the approach, and the length of record required to calibrate the multi-predictor rainfall occurrence and rainfall amount models.

Keywords: *Daily rainfall, stochastic model, nonparametric methods, longer-term variability, Australia*

1. INTRODUCTION

Australian rainfall records contain complex low-frequency features that are associated with climate variability. Harrold et al. (2002) presented a nonparametric model for generation of daily rainfall that considers these low-frequency features in its formulation, and showed that their nonparametric model, which incorporates longer-term predictors that are internal to the rainfall sequence, produces a more appropriate representation of both the short-term correlation structure of rainfall amounts and of rainfall variability at long time scales compared to models which incorporate fewer predictors. The sequences generated by the nonparametric model may be of use in catchment studies, especially when climate-related variability is being investigated; such sequences are used as a tool for exploring the potential variability in the catchment response.

A focus of this paper, which was not included in Harrold et al (2002) due to space limitations, is on the methodology for predictor selection. Traditionally used criteria for selecting the predictors for a model, including the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), are based on one-day-ahead forecasts made using the model. However, the quality of the one-day-ahead forecasts does not indicate whether the sequences generated by the model will reproduce historical longer-term variability. Jimoh and Webster [1996] propose that the use of frequency-duration curves of dry and wet spell lengths can provide an alternative method of model identification. Others who have used alternative methods for assessing model performance include Gregory et al. (1993), who state that reproduction of seasonal variance provides a crucial test of stochastic weather generators; and Wilks (1999), who tested seasonal variance, extreme daily precipitation, and runs of consecutive dry and wet days. A procedure for selecting predictors based on the quality of generated sequences is presented here.

In this paper, the approach of Harrold et al. (2002) is applied to rainfall from 13 locations in Australia. The selected locations provide a range of climates (from temperate to semi-arid to tropical) on which to test the approach.

2. THE DAILY RAINFALL MODEL

Generation of daily rainfall can be treated as a two-stage process, with the entire sequence of wet and dry days being generated before the amounts on wet days are calculated. This is the approach of Harrold et al. (2002). A brief outline of the approach is given here.

2.1 Rainfall occurrence

The occurrence model (termed ROG(j) to denote "Rainfall Occurrence Generator" with j predictor

variables) uses a moving window approach (Rajagopalan et al. 1996) to give a smooth representation of seasonal features. A window length of 15 days, centred on the current calendar day, is used to form a local subset of data for use in the model for that day. Simulation proceeds using nearest-neighbour methods (Sharma and Lall, 1999) to resample historical values and insert them into the generated sequence. The resampling is conditional to the values of the predictors that are being used in the model, which for Sydney rainfall were chosen as:

1. Rainfall occurrence on the previous day.
2. The wetness state (very wet, wet, average, dry, or very dry) for the previous 90 days.
3. The wetness state for the previous year, leading up to the current day.
4. The wetness state for the previous five years, leading up to the current day.

In this approach, a value is resampled from the successors to the k historical "neighbours" to the current pattern formed by the predictors in the generated sequence. k is a key parameter in this methodology; a poor choice of k can lead to bias in the generated sequences.

2.2 Rainfall amount

Chapman (1998) showed that stochastic models that treat rainfall amounts as separate classes based on the number of adjoining wet days (0, 1, or 2), result in a better fit than stochastic models that treat the data together, because the distributional characteristics of each class are different. Harrold et al. (2002) give separate treatment to four classes of amount, subdividing Class 1 into Class 1a (days at the start of wet spells) and Class 1b (days at the end of wet spells).

The Harrold et al. (2002) model for rainfall amounts is called RAG to denote "rainfall amount generator". The one-predictor RAG(1) model uses rainfall amount on the previous day as a short-term predictor. This model generates amounts from a conditional probability density function formed from the Class 0 amounts (or Class 1a, Class 1b, or Class 2 amounts, as appropriate, along with the rainfall amounts on the previous day) that fall within a 31-day moving window centred on the day of interest. RAG is implemented using kernel estimation of the probability densities (Sharma and O'Neill 2002). Kernel density estimation methods form a smoothed empirical probability distribution from the historical record, and generate values from this empirical distribution. Details on this methodology can be found in Srikanthan et al. (2003) and Harrold et al. (2002).

Because of complex low-frequency features in the historical record of amounts, Harrold et al. (2002) introduced a second predictor into their model for both Sydney and Melbourne rainfall amounts, conditioning the model on the wetness state for the previous year (very dry, dry, average, wet, or very wet), based on the number of wet days over this period. The resulting two-predictor model, which is denoted as RAG(2), is formulated in the same way as RAG(1), except the observations are separated into five datasets according to the historical values of the wetness state. The values of the wetness state in the generated sequence determine which dataset to use in the simulation for a particular day.

3. PREDICTOR SELECTION

Harrold et al. (2002) propose that variability in rainfall can be reproduced by linking an occurrence model that reproduces observed longer-term variability in the pattern of wet and dry days with a simpler model for rainfall amounts. The selection of predictors for both the occurrence and amounts models in this approach is based on the quality of the generated sequences. One predictor at a time is added to the existing predictor set, and the resulting model is evaluated by generating 100 sequences from the model, of the same length as the historical record, and then comparing the characteristics of the generated sequences with the characteristics of the historical record. The best performing predictor is chosen at each step of this procedure. We use this assessment method as the basis for selecting the predictors and the value of smoothing parameters for both the occurrence and amounts models.

A graphical illustration of this predictor selection method is given in Figures 1 and 2. The Figures show a "panel of plots" for the ROG(1) and ROG(4) models for Sydney rainfall occurrence, where ROG(4) incorporates all the predictors listed in section 2.1, and ROG(1) only incorporates the short-term predictor. In each plot, the historical values of each statistic are shown as either dots, or as connected line segments, and the distribution of the generated values of each statistic (from each of the 100 generated sequences) are shown as either 5%, median, and 95% lines, or as box plots with the whiskers defining the 5% and 95% values. Detailed descriptions of each plot type in the "panel of plots" are given in Harrold (2002), and are not repeated here for space reasons; instead, both "panels of plots" are shown here to give an overall impression of the difference in performance between the two models.

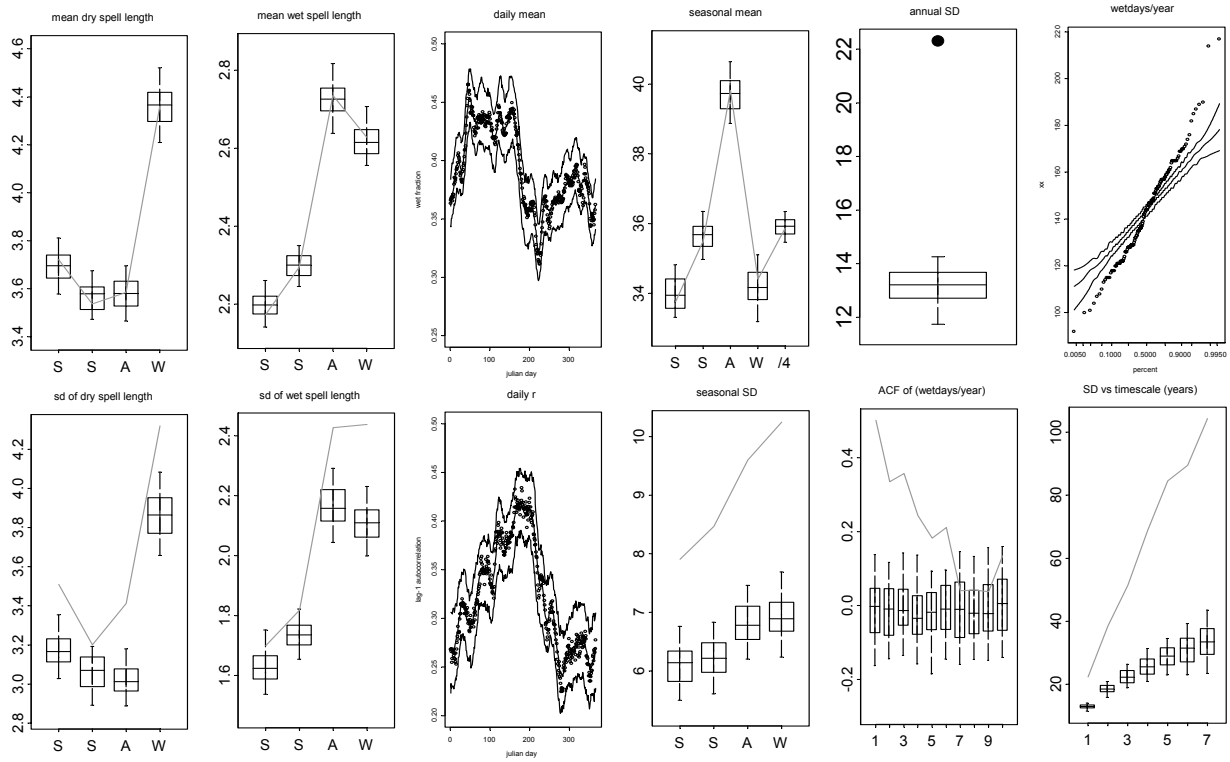

Figure 1. ROG(1): Panel of plots for Sydney rainfall occurrence.

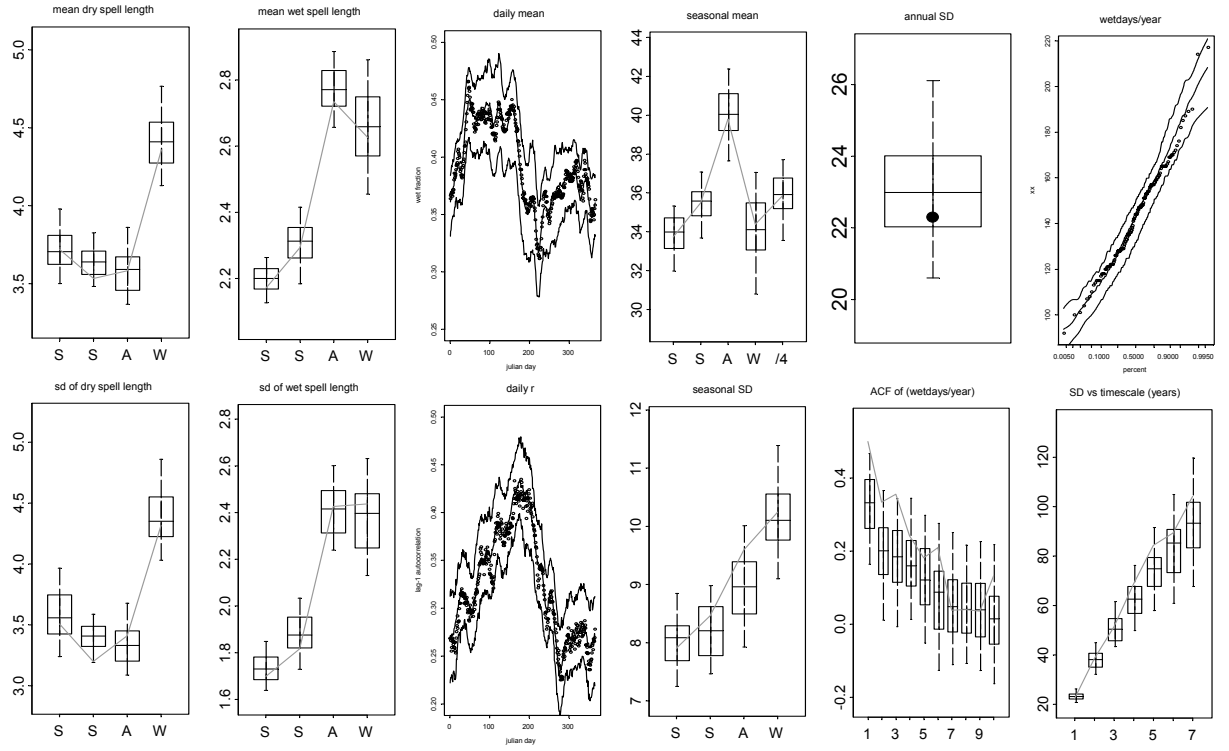

Figure 2. ROG(4): Panel of plots for Sydney rainfall occurrence.

A key statistic that is shown in both Figure 1 and Figure 2 is the distribution of wet days per year, shown in the upper right-hand plot. If the median of the generated sequences provides a good fit to

the observed values of this distribution, then other statistics (especially annual means and annual standard deviations, but also seasonal means, seasonal standard deviations, and the distribution of wet and dry spell lengths) were also well matched by the ROG model. The sum of squared residuals (SSR), based on the differences between the observed values and the median generated values in Figure 1, can be used to assess the quality of the model that produced the sequences.

Table 1. ROG models for eight of the 13 locations.

Location	Record length (years)	Model	α [a]	Annual mean (days)[a]	SSR of wetdays/year	Lag-1 corr. of wetdays/year
Adelaide	115			122.8		0.04
		ROG(1)	1	*-0.2*	371	*-0.05*
		ROG(2)	8	*-0.2*	294	*-0.01*
Alice Springs	81			38.2		0.35
		ROG(1)	1	*0.0*	674	*-0.35*
		ROG(2)	8	*-3.7*	1603	*-0.32*
Brisbane	96			117.0		0.00
		ROG(1)	1	*0.1*	3037	*-0.01*
		ROG(2)	8	*-0.2*	421	*0.03*
Broome	48			46.3		-0.03
		ROG(1)	1	*-0.2*	1015	*0.01*
		ROG(2)	8	*-2.3*	792	*0.01*
Cowra	37			96.5		0.17
		ROG(1)	1	*-0.4*	1612	*-0.15*
		ROG(2)	8	*-2.7*	1478	*-0.16*
Darwin	59			98.9		0.10
		ROG(1)	1	*-0.4*	939	*-0.11*
		ROG(2)	8	*-0.2*	336	*-0.09*
Melbourne	144			149.2		0.53
		ROG(1)	1	*0.1*	5949	*-0.53*
		ROG(2)	8	*0.0*	1227	*-0.47*
		ROG(3)	2	*0.2*	374	*-0.25*
		ROG(4)[c]	0.5	*-0.1*	306	*-0.15*
Sydney	140			143.5		0.50
		ROG(1)	1	*0.2*	11710	*-0.51*
		ROG(2)	6	*-0.2*	1657	*-0.44*
		ROG(3)	2	*-0.2*	954	*-0.27*
		ROG(4)	1	*0.0*	950	*-0.18*

[a] α is a smoothing parameter used to specify the number of nearest neighbours *k* used in each model; $k = \alpha\sqrt{n}$ where *n* is the sample size.

[b] Values shown in italics in this and the last column are biases. "Bias" is the absolute difference between the observed value and the mean of values from 100 sequences generated by the given model.

[c] For Melbourne, the multi-year predictor was chosen as the 4-year wetness state.

4. RESULTS FOR 13 LOCATIONS

In applying the above predictor selection methodology, we found that the specification of the short-term predictor was not problematic; Harrold (2002) shows that the value on the previous day is the best short-term predictor for either occurrence or amount. More effort was required, however, to select the smoothing parameter *k* for the multi-predictor models, and to determine whether seasonal-level, annual-level, and multi-year predictors were required at a particular location. We obtained *k* by trialling a range of possible values, given by $k = \alpha\sqrt{n}$ where *n* is the sample size. Even though the range of statistics shown in Figures 1 and 2 were examined, in this paper we summarise the performance of a model incorporating a given predictor set and *k* value, using the following statistics:

1. Bias in the annual mean wet days per year;

2. SSR from the distribution of wet days per year;

3. Bias in the lag-1 correlation of wet days per year.

Bias in the annual mean is related to inappropriate selection of *k*. The distribution of wet days per year can be optimised by appropriate selection of the seasonal and annual-level predictors, and the correlation of wet days per year can be optimised through the use of the multi-year predictor.

The results for selection of the ROG model for eight of the 13 locations are shown in Table 1. Initial results for testing one-predictor and two-predictor ROG models are shown for six locations (using rainfall occurrence on the previous day as the first predictor, and the 90-day wetness state as the second predictor), followed by the models selected for Melbourne and Sydney. The results for the latter two models are as reported in Harrold et al. (2002). It can be seen that the initial choice of α for the two-predictor models works for some locations (Adelaide, Brisbane and Darwin), but not for others (Alice Springs, Broome, and Cowra, where the generated annual means are biased). ROG(2) with $\alpha = 0.8$ provides a good fit to the Adelaide and Brisbane observed data; it appears that no more than two predictors for occurrence are required at these locations (we also trialed ROG(3) models here, but the addition of the annual-level predictor did not improve the results).

Alice Springs, Cowra and Darwin have correlations of wet days per year that are greater than zero. We trialed ROG(3) models at these locations. For Alice Springs and Cowra, the ROG(3) model gave a better representation of the lag-1 correlations than the ROG(2) model. The addition of a multi-year predictor to form ROG(4) may also be worthwhile.

We also trialed ROG models at Kalgoorlie, Mackay, Monto, Perth, and Tenterfield. Initial results suggest that full 4-predictor models may be required at Kalgoorlie, Mackay and Perth, and three-predictor models should provide a good fit to Monto and Tenterfield rainfall occurrence.

5. CONCLUSIONS

This paper has applied the model of Harrold et al. (2002) to the historical rainfall record from 13 locations in Australia. The models and method of predictor selection described in this paper are flexible and easy to apply, although the procedure involves generation of many years of rainfall data from models that use data-intensive methods, and significant post-processing is required to evaluate the generated sequences.

The length of record required to calibrate the multi-predictor rainfall occurrence and rainfall amount models depends on location, but in general, the longer the record the better. For Adelaide and Brisbane, only two predictors were required for the rainfall occurrence model, but for other locations (such as Sydney and Melbourne), four predictors were required, and the longer-term features of the historical record were still not perfectly reproduced. It is difficult to give a physical interpretation for the differences in results from location to location. However, limiting the predictor set to daily-level, seasonal-level, annual-level, and multi-year predictors seems appropriate for all the locations that were tested.

6. ACKNOWLEDGEMENTS

The support of the Australian Research Council and the Japan Society for the Promotion of Science is gratefully acknowledged. Rainfall data was obtained from the Australian Bureau of Meteorology.

7. REFERENCES

Akaike, H., A new look at the statistical model identification, *IEEE Transactions on Automation and Control*, AC-19, 716-723, 1974.

Gregory, J.M., T.M.L. Wigley, and P.D. Jones, Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals, *Climate Dynamics*, 8, p299-310, 1993.

Harrold, T.I., Stochastic generation of daily rainfall for catchment water management studies, PhD Thesis, School of Civil Engineering, University of New South Wales, 2002. http://adt.caul.edu.au.

Harrold, T. I., A. Sharma and S. J. Sheather, Representation of long-term variability in daily rainfall generation. Hydrology and Water Resources Symposium, Institution of Engineers, Australia, 2002.

Jimoh, O., and P. Webster, The optimum order of a Markov chain for daily rainfall in Nigeria, *Journal of Hydrology*, 185, 45-69, 1996.

Rajagopalan, B., U. Lall, and D. G. Tarboton, Nonhomogeneous Markov model for daily precipitation, *Journal of Hydrologic Engineering*, 1(1), 33-40, 1996.

Schwarz, G., Estimating the dimension of a model, Annals of Statistics, 6, 461-464, 1978.

Sharma, A. and U. Lall, A nonparametric approach to daily rainfall simulation, *Mathematics and Computers in Simulation*, 48, 367-371, 1999.

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

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

R. Srikanthan, T.I. Harrold, A. Sharma and T.A. McMahon, Comparison of two approaches for generation of daily rainfall data, *Modsim 2003 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, Townsville, 2003.

Wilks, D.S., Interannual variability and extreme-value characteristics of several stochastic daily precipitation models, Agricultural and Forest Meteorology, 93(3), 153-169, 1999.