A Nonparametric Model for Daily Rainfall Occurrence that Reproduces Long-term Variability

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

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

[b] *Australian Graduate School of Management, The University of New South Wales.*

Abstract: A goal of stochastic hydrology is to generate synthetic daily rainfall that is representative of the statistical characteristics of the historical record. Such sequences are used in catchment studies to assess the uncertainty in the catchment response that is due to climatic variability. A major deficiency of existing stochastic generation methods is their inability to represent variability at longer (seasonal, annual and inter-annual) time scales. This paper presents a nonparametric model for generating single-site daily rainfall occurrence, which is formulated to reproduce longer-term variability and low-frequency features such as drought and sustained wet periods, while still reproducing characteristics at daily time scales. The model assumes a Markovian dependence structure (assuming that a finite number of previous values in the sequence are sufficient to characterise the rainfall state on the next day). Parsimony is achieved within the Markovian framework by using "aggregate" variables that describe how wet it has been over a period of time. Actual simulation proceeds by resampling from the historical record of rainfall occurrence, conditional to the current values of the associated predictors. The use of a seasonally representative sample at any given time of year ensures an accurate representation of the seasonal variations present in the rainfall time series. The model is applied to historical daily rainfall from Sydney. We find that the use of multiple predictors produces sequences that more closely reproduce the longer-term variability present in the historic records.

Keywords: Nonparametric; Modelling; Daily rainfall; Stochastic hydrology

1. INTRODUCTION

Synthetic sequences of rainfall that are statistically consistent with the observed characteristics of the historical record can provide useful input data for catchment water management studies. These synthetic sequences are each assumed to be equally likely to occur in the future, with the same likelihood as the historical sequence, and can be used to quantify the uncertainty in the catchment response that results from climatic variability.

Methods for generating synthetic sequences of daily rainfall are reviewed by Woolhiser [1992]. Wilks and Wilby [1999] and Srikanthan and McMahon [2000] note that these methods typically do not reproduce the longer-term variability found in historical daily rainfall records. This paper presents a nonparametric model for the stochastic generation of single-site daily rainfall occurrence that is designed to address this issue. The model uses multiple predictors and is formulated to reproduce low-frequency features such as drought and sustained wet periods.

The use of multiple predictors necessitates the use of a predictor identification criterion. Jimoh and Webster [1996] question the use of traditionally used criteria, such as the Akaike Information Criterion (AIC), for selecting the number of parameters used in a rainfall occurrence model. They suggest that such criteria be used with caution. These criteria use maximum likelihood (or a similar measure) to give an indication of the goodness of fit of the model parameters to the data, and they depend on an assumed probability mass function of the residuals from a one-day-ahead forecast. The predictor selection methods described in the first paper of this series [Harrold et al., 2001a] provide a nonparametric alternative to the use of criteria such as the AIC. However, the quality of the one-day-ahead forecasts made using the selected predictors does not indicate whether the sequences generated by the model will reproduce historical longer-term variability. A different method is required to assess the longer-term behaviour of generated sequences. Such a method is described in section 3 of this paper, after a description of the rainfall occurrence model is given (section 2). Section 4 assesses the generated sequences from the models formed to reproduce longer-term variability and from the models formed from the predictors selected in Harrold et al. [2001a]. Conclusions are presented in section 5.

2. THE RESAMPLING MODEL FOR RAINFALL OCCURRENCE

Our resampling model, termed ROG to denote "rainfall occurrence generator", is based on Sharma and Lall [1999]. The differences in our approach from the Sharma and Lall [1999] model are:

- Rainfall occurrence is resampled one day at a time, rather than resampling entire wet spells or entire dry spells. An advantage of resampling one day at a time is that unprecedentedly long wet and dry spells may occur in the generated sequences.

- Multiple predictors for rainfall occurrence are used here. Sharma and Lall used a single predictor. The use of multiple predictors will enable us to capture both short-term and longer-term dependence structure, and result in generated sequences that more closely reproduce the variability found in the historical record.

We simulate short-term dependence using predictors such as rainfall occurrence at a short time lag from the present, and longer-term dependence using predictors that describe how wet it has been over a longer length of time. Simulation proceeds by resampling from the historical record of rainfall occurrence, conditional to the current values of the predictors. We model the seasonal variations present in the rainfall time series using a 15-day moving window, centred on the current day. This forms a seasonally representative sample at any given time of year, which includes data from all years. We use a threshold of 0.3mm to decide whether a day is wet or dry [after Buishand 1978].

Details of the resampling model are presented in Harrold et al. [2001b]. (Note that one of the key steps in the procedure is the selection of the number of nearest neighbours used for resampling. This is not discussed here due reasons of space and simplicity of presentation). The model conditionally resamples from a seasonal subset of the historic record, using nearest neighbour techniques. The performance of the model is evaluated by comparing 100 sequences generated by the model, each of the same length as the historical record, with the historical record.

3. PREDICTOR SELECTION

The rainfall occurrence model formed from the selected predictors should generate sequences that are statistically similar to the historical sequence, with both short-term and longer-term statistics reproduced. A method of predictor selection, based upon this principle, is to first select a short-term predictor, then a medium-term predictor, then a long-term predictor, and then a very-long-term predictor if required. This approach captures the features of the historical record in a parsimonious way. One predictor at a time is added to the existing predictor set, and the resulting model is evaluated by comparing the short-term, medium-term, and longer-term characteristics of the 100 sequences generated by the model with the characteristics of the historical record. The best performing predictor (as shown by these comparisons) is chosen at each stage.

4. RESULTS

When we applied our ROG model and our approach for stepwise selection of predictors to Sydney rainfall occurrence (1859-1998), we added the following predictors to the model:

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.

We formulated our rainfall occurrence model using these predictors. ROG(1) denotes a model which uses rainfall occurrence on the previous day (Y_{t-1}) as the single predictor. This model resamples from the days in our seasonally representative sample that are preceded by wet/dry days if the current value of Y_{t-1} is wet/dry. ROG(2), ROG(3), and ROG(4) denote models which use the two, three, or four selected predictors, respectively. These more complicated models conditionally resample from the seasonally representative sample based on the current values of the predictors.

Results are presented here for ROG(1) and ROG(4). The results for ROG(1) are presented in Figures 1 to 6. Figures 1 to 2 demonstrate the ability of a model that incorporates a single short-term predictor to reproduce the short-term statistical characteristics of the historical record. Figures 3 to 6 demonstrate the inability of such a model to reproduce the medium-term and longer-term characteristics of the historical record. The results for ROG(4) are presented in Figures 7 to 11. Figures 7 to 11 show how the use of a combination of short-term and longer-term predictors gives a model that better reproduces the medium-term and longer-term characteristics of the historical record. We also found that the short-term characteristics were reproduced by ROG(4).

Figure 1 shows the fraction of wet days for the ROG(1) model for Sydney, as it varies with time of year. The distribution of the statistic from the 100 generated sequences is shown by the 5th percentile, median, and 95th percentile lines. Superimposed on this graph are the historical values (circles). It can be seen that ROG(1) adequately reproduces the historical values. ROG(1) also adequately reproduces the historical daily lag-one correlations, as shown in Figure 2.

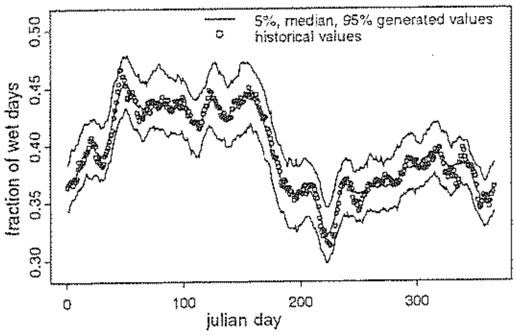

Figure 1 ROG(1): Fraction of wet days for Sydney as a function of julian day.

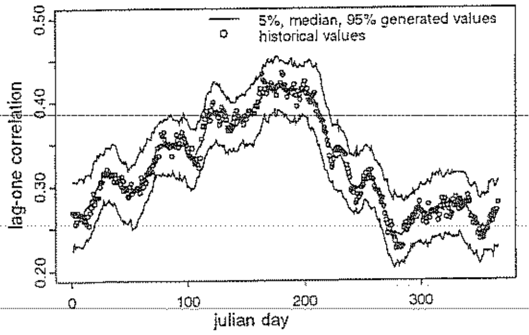

Figure 2. ROG(1): Lag-one correlations of Sydney rainfall occurrence as a function of julian day.

Figure 3 shows the mean length of the wet spells (sequences of consecutive wet days) and the standard deviation of the wet spells ending in each of four seasons (Spring, Summer, Autumn and Winter) for ROG(1) for Sydney. Values of each statistic were calculated for each of the 100 generated sequences, and the distribution of these values is shown as a boxplot. The historical values are superimposed on the plot, connected by a line. The generated sequences adequately reproduce the historical mean wet spell lengths, but not the historical standard deviations. Similar results were obtained for dry spells. Boxplots formed from the longest wet spell and the longest dry spell ending in each season, for each of the 100 generated sequences from ROG(1), are shown in Figure 4, with the historical values superimposed on the plot.

The mean number of wet days per season, and the standard deviation of the number of wet days per season, are shown in Figure 5. The generated sequences from ROG(1) reproduce the historical means, but not the historical standard deviations.

A probability plot of the distribution of wet days per year is given in Figure 6. The figure shows that the driest year on record for Sydney has only 90 wet days, but the wettest year has approximately 220 wet days. The generated sequences from ROG(1) do not reproduce this distribution. Note that the standard deviation of the number of wet days per year is directly related to this distribution, and is not reproduced either.

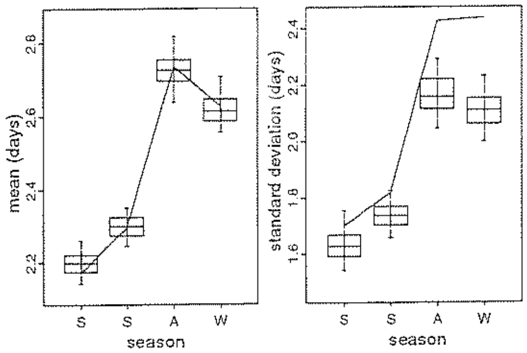

Figure 3. ROG(1): Mean (left panel) and standard deviation of wet spell lengths in each season for Sydney.

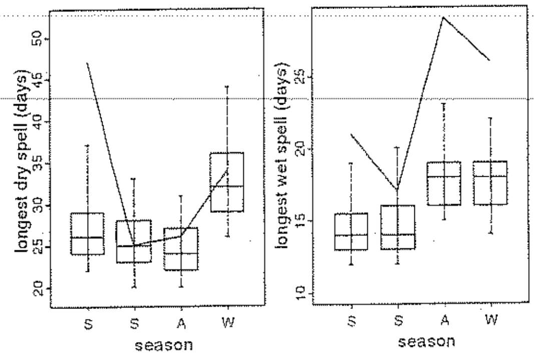

Figure 4. ROG(1): Longest dry spell (left panel) and longest wet spell in each season for Sydney.

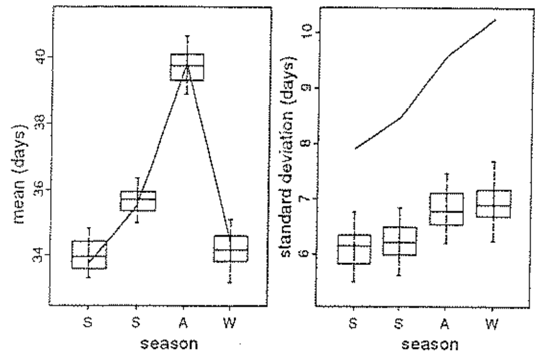

Figure 5. ROG(1): Mean (left panel) and standard deviation of wet days per season for Sydney.

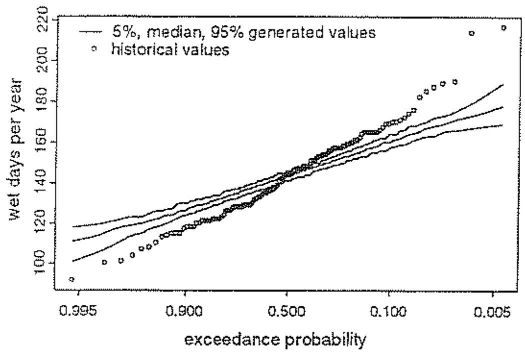

Figure 6. ROG(1): Distribution of wet days per year for Sydney.

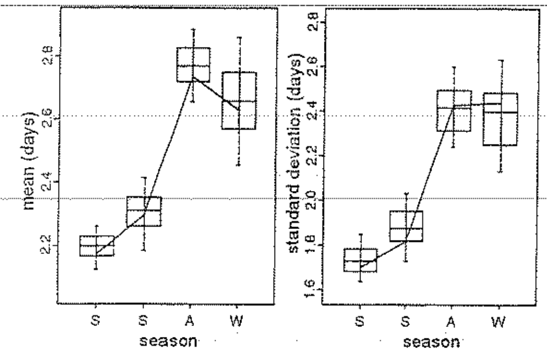

Figure 7. ROG(4): Mean (left panel) and standard deviation of wet spell lengths in each season for Sydney.

We found that generated sequences from ROG(4) adequately reproduce the historical fraction of wet days and the historical daily lag-one correlations, with similar results to those shown in Figures 1 and 2. Generated sequences from ROG(4) also adequately reproduce both the mean wet spell lengths, and the standard deviations of the spell lengths, as shown in Figure 7. The result for the standard deviations is a marked improvement over ROG(1) (cf. Figure 3). This improvement was also found in the standard deviations of dry spell lengths. The representation of extreme dry and wet spell lengths (Figure 8) has also improved.

Figure 9 shows the results for ROG(4) for the mean number of wet days per season, and the standard deviation of the number of wet days per season. There is a substantial improvement in the representation of the seasonal standard deviations, compared to the results for ROG(1) (cf. Figure 5).

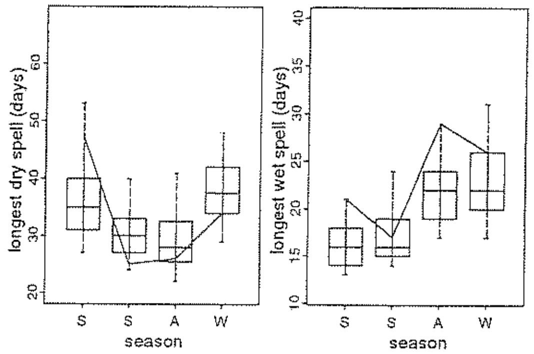

Figure 8. ROG(4): Longest dry spell (left panel) and longest wet spell in each season for Sydney.

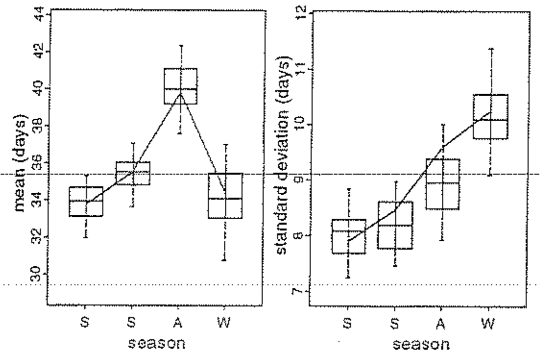

Figure 9. ROG(4): Mean (left panel) and standard deviation of wet days per season for Sydney.

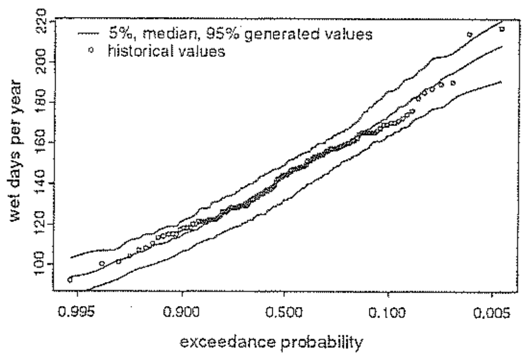

Figure 10. ROG(4): Distribution of wet days per year for Sydney.

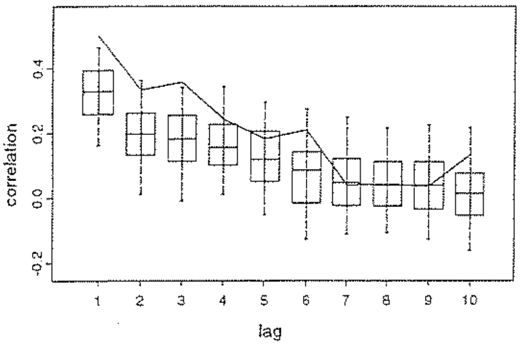

Figure 11. ROG(4): Autocorrelation function of wet days per year for Sydney.

The probability plot of the distribution of wet days per year for ROG(4) is given in Figure 10. The generated sequences from ROG(4) adequately reproduce this distribution. We consider this to be a major achievement of our model. The standard deviation of the number of wet days per year is also reproduced by ROG(4).

Figure 11 shows the autocorrelation function of the number of wet days per year, from lag one up to lag ten. A line joins the ten historical values, and box plots of the values from the generated sequences from ROG(4) are shown. This very-long-term statistic was not reproduced until the fourth predictor (the five-year wetness state) was added to our model. While the historical values are not perfectly reproduced by ROG(4), the degree of long-term variability in the sequences produced by the model is good, considering the complexity of the natural processes that contribute to the historical variability.

We found that the probability plots of the distribution of the number of wet days per year (Figures 6 and 10) gave a very good indication of the overall quality of the generated sequences, and that if the generated sequences could reproduce this distribution, then other statistics were also well reproduced. We therefore used the probability plots to compare all the models. The sum of squared residuals (SSR) from the probability plots were calculated based on the differences between the historical values and the mean of the generated values. The results of these calculations are shown in Table 1. A smaller SSR indicates a better-fitting model. In the Table, ROG(2A) and ROG(3A) denote the models formed from the two-predictors and the three-predictors (respectively) selected for each season using the predictor identification methods described in the first paper of this series [Harrold et al., 2001a].

Table 1 shows that models formed from predictors found using the method described in section 3 of this paper (i.e. ROG(2), ROG(3), and ROG(4)) reproduce the distribution of wet days per year better than models formed from predictors identified using the methods outlined in Harrold et al. [2001a] (i.e. ROG(2A) and ROG(3A). This difference was also found to apply to other seasonal, annual and very-long-term statistics. In formulating ROG(4), the first predictor was found to be good at matching daily level dependencies, the second predictor good at matching seasonal level dependencies, the third predictor good at matching annual level dependencies, and the fourth predictor good at matching interannual level dependencies. A combination of these four predictors forms a model that can capture much of the natural variability of the historical record. In contrast, the predictor identification methods used to form ROG(2A) and ROG(3A) pick the best possible choice for rainfall occurrence on the current day. This is a one-step-ahead prediction, and, apart from matching short-term characteristics, this method does not produce generated sequences that are statistically similar to the historical sequence.

Table 1. Sum of squared residuals (SSR) from the distribution of wet days per year, for various models for Sydney rainfall occurrence.

Model	SSR
ROG(1)	11710
ROG(2)	1657
ROG(3)	954
ROG(4)	950
ROG(2A)	7035
ROG(3A)	2786

We applied our ROG model to Melbourne rainfall occurrence and obtained similar results to those reported here, arriving at a recommended four-predictor model using the same predictors as the Sydney model (except the fourth predictor selected was the wetness state for the previous four years). We also applied a model formed from a combination of three predictors to rainfall occurrence from eleven additional Australian locations, including Adelaide, Alice Springs, Brisbane, Broome, Cowra, Darwin, Kalgoorlie, Mackay, Monto, Perth, and Tenterfield. The results that were obtained indicated that some locations, such as Adelaide, may only require two predictors to adequately model the short-term and longer-term variability of the historical rainfall occurrence pattern, while other locations may require three or four predictors. The result for Adelaide may indicate that its rainfall record is less affected by long-term climatic influences (such as the El Nino Southern Oscillation) than other locations.

5. CONCLUSIONS

We have presented a new model for generating long synthetic sequences of daily rainfall occurrence that reproduce both the short-term and longer-term variability of the historical record, and we have presented a method for selecting the predictors used in the model. These generated sequences provide a better representation of droughts and sustained wet periods than was previously possible. Such features are of great interest in catchment management studies, and the generated sequences can be used in such studies to enable better quantification of the uncertainty in the catchment response that is due to climatic variability. The model resamples from a seasonal subset of the historical record of rainfall occurrence, conditional to the values of a set of multiple predictors. The predictors are formed solely from previous values in the sequence, and represent short-term, seasonal, annual, and inter-annual features of the rainfall sequence. Predictors are selected in a stepwise procedure, with the shorter-term predictors selected first. We find that the use of these multiple predictors in our resampling model produces generated sequences that more closely reproduce the historical longer-term variability. These improvements to the long-term behaviour of the generated sequences do not compromise the short-term performance of the model (i.e. features such as daily means and lag-one correlations were similar regardless of the number of predictors used). The final models recommended for both Sydney and Melbourne each use four predictors. Initial results for 11 other Australian locations suggest that, in general, a two-predictor, three-predictor, or four-predictor model can be found which can produce synthetic sequences which closely match the historical variability at that location.

The predictor selection methods described in this paper are designed to give a model that adequately reproduces both short-term and longer-term statistical properties of the historical series. In contrast, traditionally used parametric model order-selection criteria (as discussed in the introduction to this paper, and as discussed in Harrold et. al. [2001a]) are designed to give a model that produces accurate forecasts of rainfall occurrence one day ahead of the present. We show in this paper that these criteria are not good indicators of whether synthetic sequences produced by the chosen model will reproduce the longer-term statistical properties of the historical series.

6. FURTHER WORK

Further research will focus on the problem of stochastically generating rainfall amount values for all the wet days simulated by the rainfall occurrence model. Details on the method for identifying predictors for such a model, and on the procedure used to generate the rainfall amounts will be published at a later stage.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Australian Research Council and the NSW Department of Land and Water Conservation for funding this research, and the constructive comments of the anonymous MODSIM reviewer.

8. REFERENCES

Buishand, T.A., Some remarks on the use of daily rainfall models, *Journal of Hydrology*, 36, 295-308, 1978.

Harrold, T.I., A Sharma and S. Sheather, Predictor selection for a daily rainfall occurrence model using partial informational correlation, In: MODSIM 2001 Congress, Canberra, 10-13 December, 2001a.

Harrold, T.I., A Sharma and S Sheather, Stochastic generation of daily rainfall occurrence: 2. A nonparametric simulation model, *Water Resources Research* (submitted), 2001b.

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

Sharma, A., and U. Lall, A nonparametric approach for daily rainfall simulation, *Mathematics and Computers in Simulation*, 48, 361-371, 1999.

Srikanthan, R., and T.A. McMahon, Stochastic generation of annual, monthly and daily climate data: a review, Report 00/16, Cooperative Research Centre for Catchment Hydrology, Monash University, 2000.

Wilks, D.S., and R.L. Wilby, The weather generation game: a review of stochastic weather models, *Progress in Physical Geography*, 23(3) 329-357, 1999.

Woolhiser, D.A., Modelling daily precipitation - progress and problems, In: Walton, A., and P. Gutton (eds), *Statistics in the Environmental and Earth Sciences*, Edward Arnold, London, 1992.