Predictor Selection for a Daily Rainfall Occurrence Model using Partial Informational Correlation

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: Stochastic generation of daily rainfall is an important part of many engineering applications. It is not surprising that several methods for generation of daily rainfall have evolved over time. While many such methods work well at representing the variations in rainfall from one day to the next, they are found lacking at representing features at longer time scales. There is a need to develop methods for stochastic generation of daily rainfall that can incorporate low-frequency features such as drought or long wet periods in the generated record. This paper is part of a study to develop approaches for generation of daily rainfall sequences at a given location that can represent the variability in rainfall at both short (daily) and long (seasonal, annual and inter-annual) time scales. The generation of daily rainfall occurrence (whether a day is "wet" or "dry") is the problem addressed in this two paper series. This first paper of the series presents an approach to select relevant predictors of rainfall occurrence, and the second paper presents a nonparametric stochastic model for generating rainfall occurrence over time. In this paper, the relationships between daily rainfall occurrence and variables formed from previous values in the rainfall occurrence sequence are examined, using a partial measure of association for discrete variables termed the partial informational correlation (PIC). We identify the best predictors of daily rainfall occurrence using a stepwise implementation of PIC. We demonstrate our procedure by applying it to long-term rainfall data from Melbourne and Sydney. The utility of the selected predictors is then evaluated by forecasting the rainfall occurrence for each day in the historical record in a leave-one cross-validation mode.

Keywords: Nonparametric; Modelling; Daily rainfall; Stochastic hydrology; Variability

1. INTRODUCTION

Long series of generated daily rainfall can provide multiple input sequences for catchment water management studies. These sequences supplement the historical record and form an ensemble of inputs to a catchment water management model. They can be used in a Monte Carlo simulation [Salas 1993] to enable risk-based assessment of the water management plans being considered. Consider a simplistic example. Under a particular management plan, a catchment water management model may indicate that no water is available for irrigation uses in the two most severe historical droughts. Monte Carlo simulation using long sequences of good quality synthetic data may help quantify the risk of failure more accurately, enabling meaningful comparison between alternate management plans. Accurate depiction of the risks involved with each management plan requires that the generated rainfall sequences are representative of the features observed in the historical rainfall record. Generation of synthetic rainfall sequences that reproduce these features can be a difficult task, especially if low frequency variability (representing droughts and sustained periods of high rainfall) is present in the rainfall record. Most approaches for daily rainfall generation are limited in their ability to reproduce such low-frequency attributes in the generated rainfall sequences. The variability of seasonal and annual totals generated by these approaches is known to be lower than the respective observed values [Buishand, 1978; Wilks and Wilby, 1999]. Such reduced variability affects the representation of sustained droughts or periods of continuously high rainfall in the generated sequences, features that are of great interest in catchment planning and management. There is a need to develop methods for stochastic generation of daily rainfall that can reproduce low-frequency features in the generated rainfall sequences, with an accurate representation of the variability of longer-term

totals. Water resource managers need to be convinced that the generated sequences are representative of historical features before they will use such sequences in risk-based assessment of water management plans, especially when their catchment models are sensitive to climatic variability.

This two paper series is part of a study to develop stochastic approaches for generation of single-site daily rainfall that represents the historical variability in rainfall at both short and long time scales. The rainfall generation problem is approached in two stages: generation of rainfall occurrence, and subsequent generation of rainfall amounts on the simulated wet days. An approach for generation of daily rainfall occurrence (whether a day will be "wet" or "dry") is presented here. This first paper concentrates on identifying predictors to be used in a model for generating the daily rainfall occurrence state. The second paper presents an approach for generating rainfall occurrence time series that represent the day-to-day variations in the historical record, and can represent variability at longer (seasonal, annual and inter-annual) time scales through the use of appropriately specified predictors. The generation of the rainfall amounts is not discussed here and will be presented in future work.

Many traditional methods for daily rainfall generation assume that the daily occurrence or amount depends exclusively on the rainfall that occurred on the previous day, an assumption that ensures under-representation of variability at longer time scales. The approach adopted here for stochastic generation of daily rainfall occurrence is to incorporate both short-term (day to day) and longer-term (seasonal, annual and inter-annual) features into the generated record through the use of appropriately selected predictor variables that describe these features. Use is made of "aggregate" variables that describe how many wet days have been observed over a period of time. We formulate a number of aggregate variables that represent the rainfall state over varying aggregation periods. These aggregate variables are used in our set of candidate predictors for daily rainfall occurrence. The candidate predictors are formed solely from previous values in the rainfall occurrence sequence, and are formulated to represent the short-term and longer-term variability that exists in the historical rainfall occurrence record. This approach can capture the dependence structure of the data in a way that involves fewer parameters than a traditional multi-order Markov chain model.

The presence of a number of possible predictor variables necessitates the use of a predictor identification criterion. Such a criterion should be capable of working with discrete variables, as our predictand (rainfall occurrence) as well as the candidate predictors, all assume discrete values. Additionally, the criterion should be such that no explicit or implicit assumptions about the nature of variability or dependence are made in selecting the predictors. Unfortunately, traditional methods used to select the model order, such as the Akaike Information Criterion [Akaike, 1974; Wilks and Wilby, 1999], make assumptions about the probability distributions of the variables involved. We present here a nonparametric procedure for measuring the dependence between discrete variables that avoids specification of probability distributions (such as Binomial, Multinomial or Poisson). The procedure also avoids assuming linear or a specified non-linear dependence. Our proposed procedure is termed partial informational correlation (PIC). This is a "partial" measure of dependence, which allows it to be used to identify predictors in a stepwise approach. Our proposed approach is related to the Partial Mutual Information (PMI) criterion for a system of continuous random variables [Sharma, 2000].

2. PARTIAL INFORMATION

The mutual information (MI) criterion [Sharma, 2000; Linfoot, 1957] is a measure of dependence that can detect and quantify both linear and non-linear relationships. Sharma [2000] shows that MI performs better than correlation in detecting and quantifying a range of non-linear dependence structures, and that it also performs well in quantifying linear dependence. We believe that the mutual information criterion can quantify a broader range of underlying dependence structures than any other available method.

Partial Information (PI) is a measure of partial information that has been developed by the authors, as an extension of the theory of mutual information. Full details of the theory behind partial information can be found in Harrold et al. [2001b]. For any given multivariate sample where the variables are discrete, and X denotes the response and P_1, P_2, ..., P_K denote the predictor variables, the PI score can be estimated as:

$$\hat{\mathrm{PI}}(X, P_K \mid \mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\hat{p}(x_i, p_{ki} \mid \mathbf{z}_i)}{\hat{p}(x_i \mid \mathbf{z}_i)\hat{p}(p_{ki} \mid \mathbf{z}_i)}\right] \quad (1)$$

where $(x_i, p_{1i}, p_{2i}, \ldots, p_{ki})$ is the i^{th} multivariate sample data point in a sample of size n, $\mathbf{z}_i = (p_{1i}, p_{2i}, \ldots, p_{(k-1)i})$, and $\hat{p}(x_i \mid \mathbf{z}_i)$, $\hat{p}(p_{ki} \mid \mathbf{z}_i)$, and $\hat{p}(x_i, p_{ki} \mid \mathbf{z}_i)$ are the

conditional probability mass functions estimated at the sample data points.

$\hat{\mathrm{PI}}(X, P_K \mid \mathbf{z})$ estimates the partial dependence between X and P_K, after accounting for the effect of the existing predictor set $\mathbf{z}$. The rationale behind PI is the definition of dependence (or independence). The conditional joint probability density function is equal to the product of the two conditional probability densities if there exists no dependence between X and P_2, after accounting for the effect of $\mathbf{z}$. The PI score in (1) would, in that case, equal a value of 0 (the ratio of the joint and marginal densities being one, and the log of this equals zero). A high value of the PI score would indicate dependence between X and P_2, after accounting for the effect of $\mathbf{z}$. PI is scale invariant, and remains unchanged if either variable undergoes any linear or non-linear transformation.

When the first predictor is being investigated, the pre-existing predictor set $\mathbf{z}$ is empty and (1) collapses to the equation for mutual information. This also occurs if $\mathbf{z}$ has only one possible value, which leads to a useful method of calculating PI for discrete data. If the sum in (1) is expressed as separate equations for each possible $\mathbf{z}_i$, (1) becomes a weighted sum of as many MI estimates, as there are possible states of $\mathbf{z}_i$ (a finite number as $\mathbf{z}_i$ is discrete).

Mutual information can be easily transformed to give a statistic that lies in the range 0 to 1, where 0 represents no dependence and 1 represents perfect dependence. The rescaled statistic is called informational correlation [Linfoot, 1957]. This can be applied to partial information to give us partial informational correlation (PIC):

$$\hat{\mathrm{PIC}} = \sqrt{1 - \exp(-2\hat{\mathrm{PI}})} \qquad (2)$$

As PIC assumes a range of 0 to 1, and can be thought of as a generic measure of correlation independent of distributional specifications, all results presented in the sections that follow use PIC as the measure of dependence.

When a measure of dependence is estimated from a small sample, there is some uncertainty as to whether the calculated value represents dependence that is significant, i.e. that the underlying (population) dependence is greater than zero. We use a randomisation test [Maritz, 1981] to test significance, where the X variable is repeatedly resampled without replacement to form a number of randomised samples that have no dependence between X and the other variables. If the calculated PIC value is greater than the upper 95th percentile PIC from the randomised samples, the calculated PIC value represents significant dependence, and we expect that there is a less than 5% chance that the variables are independent. The 95th percentile PIC as described here will be denoted as PIC_{95} in the discussion that follows. PIC_{95} is used to indicate whether the estimated value of PIC is statistically significant.

3. SELECTION OF PREDICTORS FOR RAINFALL OCCURRENCE

We apply the partial information criterion to the problem of predictor selection for a daily rainfall occurrence model, using long-term daily rainfall from 13 locations in Australia. Only Melbourne and Sydney results are presented here due to space limitations. A threshold of 0.3 mm was used to define a "wet day" [after Buishand, 1978], and a September-August water year was used.

Our set of candidate predictors was chosen to represent a range of short, medium and longer-term features in the daily rainfall occurrence sequence. The candidate predictors for rainfall occurrence that we adopt are the length (L) of the previous dry/wet spell leading up to the current day, the number of wet days in the last M days where M = 1, 2, 3, a "wetness index" for the last D days where D = 7, 14, 30, 60, and 183 days, and a "wetness index" for the last Y years where Y = 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 years. Thus 19 candidate predictors are formulated. These candidate predictors are denoted as L, 1d, 2d, 3d, 7d, 14d, 30d, 60d, 183d, 1y, 2y, 3y, 4y, 5y, 6y, 7y, 8y, 9y, and 10y, in the discussion that follows. As discussed in the introduction, these candidate predictors are "aggregate" variables that describe how wet it has been over a period of time. Values of these predictors were calculated for every day in the historical rainfall occurrence record, except in the first ten years of the record.

Because we are working with rainfall occurrence, which is a two-state discrete variable, it is important that we formulate discrete predictors with a limited number of states (this allows more accurate estimation of the probability mass functions in (1)). Each of the "wetness indexes" (i.e. candidate predictors 7d to 10y) is a state variable that describes how wet it has been over the previous period. Each wetness index can take integer values between 1 and 5, where 1 = very dry, 2 = dry, 3 = average, 4 = wet, and 5 = very wet. Values are assigned based on comparison with the ranked historical values of the number of wet days in each period of length D days (or Y years) ending in the sample being investigated. Candidate predictor L (the length of the previous dry/wet spell) is also a state variable, taking values between -3 and 3, where -3 = a very long

wet spell, -2 = a long wet spell, -1 = a short wet spell, 1 = a short dry spell, 2 = a long dry spell, and 3 = a very long dry spell. Values are assigned based on the ranked historical lengths of the dry/wet spells that end in the sample being investigated. Dry and wet spells are ranked separately.

The PIC scores for each candidate predictor are calculated for each of 24 seasonal windows, which span the year. The seasonal windows provide a sufficiently large sample for calculation purposes, while keeping seasonal effects negligible in the sample formed by the window. Each seasonal window includes values from all except the first ten years of the historical record. The six seasonal windows forming Spring (March-April-May), Summer (June-July-August), Autumn (September-October-November) and Winter (December-January-February) have their PIC results averaged to give a single parsimonious predictor set for each of the four quarters. A stepwise predictor selection algorithm [see Harrold et al., 2001b] standardised the results using randomisation tests. This method produced consistent results and allowed the candidate predictors to be compared against each other.

4. RESULTS: SELECTED PREDICTORS FOR RAINFALL OCCURRENCE

Table 1 shows the three most significant predictors for daily rainfall occurrence that were found for Melbourne and Sydney, for each of four quarters. The PIC/PIC_{95} ratio is also shown. The predictors that were significant at the 95% level (i.e. with PIC/$PIC_{95} \geq 1.0$) are underlined.

Rainfall occurrence on the previous day (i.e. candidate predictor 1d) is identified as a significant predictor in every quarter for both Melbourne and Sydney. This was also true at the other 11 locations we tested. Our results showed that when the first predictor was being chosen, the 1d predictor easily outperformed all other candidate predictors at all the locations being tested. However, when a second predictor was chosen, the difference in performance between the candidate predictors was not as clear-cut. Additionally, in no case were more than two predictors identified as significant at the 95% level. The location that came closest to having a third significant predictor was Sydney in winter (predictor 3d, S=0.97).

The results in Table 1 show that rainfall occurrence on the previous day is the best single predictor for rainfall occurrence on the current day. This is an implicit assumption in rainfall occurrence models that generate occurrence one day at a time, such as a Markov chain. Note that a combination of candidate predictors 1d and 2d is equivalent to the predictors used in a traditional second-order Markov chain model. Our results indicate that a first-order or second-order Markov chain model should provide a reasonable fit to the short-term historical features of most of the 13 rainfall occurrence records that we tested. However, predictor combinations that were identified for Melbourne and Sydney in winter are quite different to the predictors used in a traditional Markov chain model. In all the cases where more than one predictor is identified as significant, a model that uses these identified multiple predictors should have the potential to better reproduce the short-term historical features of the rainfall record. We test this in the next section of this paper.

Table 1. Selected predictors for rainfall occurrence.

Location	Sep-Nov			Dec-Feb			Mar-May			Jun-Aug		
	1	2	3	1	2	3	1	2	3	1	2	3
Melbourne	***1d***	5y	2d	***1d***	6y	183d	***1d***	10y	60d	***1d***	***4y***	30d
1855-1998	5.68	0.87	0.9	6.01	0.86	0.95	6.30	0.99	0.89	4.20	1.21	0.93
Sydney	***1d***	183d	60d	***1d***	60d	4y	***1d***	***L***	2y	***1d***	***9y***	3d
1859-1998	6.12	0.93	0.9	5.93	0.88	0.89	8.05	1.08	0.91	7.85	1.15	0.97

Table 2. MSE for forecasts of Melbourne rainfall occurrence.

Predictor combination	spring	summer	autumn	winter	annual
Unconditional resampling	0.490	0.404	0.463	0.500	0.464
1-predictor	0.459	0.372	0.427	0.477	0.434
2-predictors	0.459	0.371	0.426	0.474	0.432
3-predictors	0.458	0.371	0.428	0.475	0.433

Table 3. MSE for forecasts of Sydney rainfall occurrence.

Predictor combination	spring	summer	autumn	winter	annual
Unconditional resampling	0.467	0.476	0.491	0.465	0.475
1-predictor	0.433	0.439	0.427	0.397	0.424
2-predictors	0.433	0.438	0.427	0.395	0.423
3-predictors	0.434	0.439	0.427	0.393	0.423

5. FORECASTS USING THE IDENTIFIED PREDICTORS

In order to test the utility of the selected predictors in forecasting the rainfall occurrence state, a leave-one cross validation analysis of rainfall occurrence forecasts was conducted. This involved predicting the rainfall one day at a time for the full historical record using a simple forecast approach, and comparing the predicted rainfall state with what was observed in reality. The rainfall prediction approach adopted here is related to the rainfall generation model described in the second paper of this series [Harrold et al., 2001a]. The approach adopted consisted of using the current state of the predictors to conditionally generate one hundred forecasts of the rainfall occurrence state on the current day, the prediction model being formulated based on all observations in a seasonal subset of the historic record except the observations from the year corresponding to the day being predicted. Such an approach, known as leave-one-cross-validation, allows the model to be tested on data points not used in model development.

With the dry state denoted as "0" and the wet state as "1", the leave-one-cross-validation predictions were compared to the true rainfall state, as inferred from the historical record. If the one-day ahead predictions are working well, one expects that the difference between the actual state and the average predicted state would be small. The measure of error used for the i^{th} day of the year is:

$$\mathrm{MSE}_i = \frac{1}{n}\sum_{j=1}^{n}\left(x_{j,i} - E_{(-j)}(\hat{x}_{j,i} \mid \mathbf{z}_{j,i})\right)^2 \qquad (3)$$

where MSE_i is used to denote the mean square error, n is the number of years of historical data used in the calculation, $x_{j,i}$ is the observed rainfall occurrence state ("0" representing dry and "1" representing wet) on the i^{th} day of year j, and $E_{(-j)}(\ )$ represents the expectation operator, estimated here as the average value of the predicted rainfall state, conditional to $\mathbf{z}_{j,i}$, the predictor variables associated with $x_{j,i}$. The prediction model is formulated so as to use all observations in a seasonal subset of the historic record except those from the year for which the prediction is made (i.e year j).

The values of MSE obtained using various combinations of predictors for the forecasts for Melbourne and Sydney are presented in Table 2 and Table 3. Values from all days in a season are averaged to give the results shown in the table. The predictor combinations are:

a. No predictors (unconditional resampling)
b. 1-predictor (rainfall occurrence yesterday, 1d)
c. 2-predictors for each quarter, from Table 1
d. 3-predictors for each quarter, from Table 1.

The results show that the relative reduction in MSE due to using 1-predictor (rainfall occurrence yesterday) instead of unconditional resampling is large. This reinforces our conclusion from the previous section that rainfall occurrence on the previous day is the single best predictor for rainfall occurrence on the current day. Incorporating this predictor into our forecast model significantly reduces the error of the forecast. The results for Melbourne show that 2-predictors produce a lower prediction error than 1-predictor in winter (June-August). For this season, our predictor selection methods using PIC identify two predictors as significant (cf. Table 1). 2-predictors are also slightly better than 1-predictor in summer and autumn. For autumn, our predictor selection algorithm was bordering on selection of two predictors. The use of more than two predictors is not justified (except, possibly, in spring), as the prediction errors are not lower in the 3-predictors case than in the 2-predictors case. With the

exception of spring, this agrees with our result from the predictor selection algorithm.

The results for Sydney show that the use of more than two predictors is not justified in spring, summer and autumn, as the MSE are not lower in the 3-predictors case than in the 2-predictors case. In winter, the MSE results suggest that 3-predictors should be used. This agrees with our result from the predictor selection algorithm. For winter, Table 1 shows that we were on the borderline of selecting three predictors (as noted in the discussion after the table).

We also tested alternate predictor sets, chosen to represent specific time scales (daily, seasonal, annual and interannual), and found that the predictor combinations identified using PIC have equal or lower prediction errors than the alternate predictor combinations. This confirmed that our PIC predictor identification method gives an optimal or near-optimal predictor set for prediction of rainfall occurrence one day ahead of the present.

6. CONCLUSIONS

This paper has presented a new measure of dependence that we call partial informational correlation (PIC), and applied PIC to discrete time series data. PIC is a partial measure of dependence derived from mutual information theory, which is sensitive to both linear and non-linear dependence. We have used PIC in an approach for identifying predictors for use in a rainfall occurrence simulation model. We formulated a new set of candidate predictors for daily rainfall occurrence formed solely from previous values in the sequence. We tested our approach at Melbourne and Sydney. The selected predictors are validated using a leave-one cross-validation analysis of rainfall occurrence forecasts.

The predictor identification methods in this paper use PIC as the criterion to pick the best possible choice for rainfall occurrence on the current day. We are, in effect, "using the best information we have (drawn only from the historical time series of rainfall occurrence) to pick whether today is going to be wet or dry". It is therefore appropriate to validate the performance of the selected predictors using leave-one cross validation analysis of rainfall occurrence forecasts, as was done here. Our conclusion from this validation is that our PIC predictor identification method gives us an optimal or near-optimal predictor set for the short-term prediction of daily rainfall occurrence. The method is a nonparametric alternative to the use of traditionally used order selection techniques such as the Akaike Information Criterion.

We stated in the introduction that we are particularly interested in representing the longer-term dependence that is associated with historical features such as droughts and sustained periods of high rainfall. A different method for measuring the performance of a generation model based on the assessment of a large number of long generated sequences is proposed in Harrold et al. [2001a]. A model for stochastically generating rainfall occurrences is also described in the second paper.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Australian Research Council and the NSW Department of Land and Water Conservation for funding this research.

8. REFERENCES

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

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, A nonparametric model for daily rainfall occurrence that reproduces long term variability, In: MODSIM 2001 Congress, Canberra, 10-13 December, 2001a.

Harrold, T.I., A Sharma and S Sheather, Stochastic generation of daily rainfall occurrence: 1. Selection of relevant predictors, *Water Resources Research* (submitted), 2001b.

Linfoot, E.H., An informational measure of correlation, *Information and Control* 1, 85-89, 1957.

Maritz, J.S., *Distribution-free statistical methods*, Chapman and Hall, London, 1981.

Salas, J.D., Analysis and modelling of hydrologic time series, In: Maidment, D.R. (ed.). *Handbook of Hydrology*, McGraw-Hill, New York, 1993.

Sharma, A., Seasonal to interannual rainfall ensemble forecasts for improved water supply management: 1. A strategy for system predictor identification, *Journal of Hydrology*, 239, 232-239, 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.