A nonparametric approach for daily rainfall simulation

Ashish Sharma[1] and Upmanu Lall[2]

[1]School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia
[2]Utah Water Research Laboratory, Utah State University, Logan, Utah, USA

Abstract: A nonparametric model for daily rainfall simulation is presented. Nearest neighbour methods are used to conditionally simulate rainfall spells and amounts. A "local" subset of the observed record is used to formulate the conditional densities needed for simulation. This provides an effective yet simple way to model local and seasonal features in the observed rainfall time series. The model is applied in two stages. First dry and wet spell lengths are conditionally simulated. Next rainfall amounts for each day of the wet spell are simulated assuming an order one Markov dependence structure. Higher order dependence is modelled by considering the number of days from start of the spell as an additional variable. Rainfall distributional characteristics are observed to have distinctly different characteristics depending on the length of the wet spell. A procedure is developed to reproduce such differences in the simulations. The model is applied to 123 years of daily rainfall from Sydney, Australia.

1. INTRODUCTION

Modelling precipitation is of importance to several disciplines in science and engineering. Synthetic precipitation sequences are used in applications as diverse as agricultural planning, reservoir and watershed management, erosion prediction, design of landfills, and design of facilities for storage and disposal of hazardous wastes. While the time scales modelled range from a year to a few minutes, simulation of daily precipitation has attracted the most attention. Both mathematical and physically based approaches have been proposed for synthesising precipitation, although complexities in the rainfall generation mechanism have limited the scope and applicability of the latter. Mathematical precipitation simulation models consider rainfall as a random process and attempt to simulate daily rainfall sequences that are "representative" of the observed record. This involves reproducing in simulations, certain features that symbolise a daily rainfall time-series. Some of these are:

- Fractions of wet and dry days
- Distribution of wet and dry spell lengths
- Dependence characteristics of the spells
- The first two or three moments of rainfall amounts
- Dependence characteristics of rainfall
- Seasonal and long term dependencies in rainfall. For example, the rainfall generation mechanism may be highly dependent on long-term climatic features such as the El-Nino Southern Oscillation. Periodicities associated with such features would then be desirable in the precipitation simulations.

Most precipitation simulation models reproduce only a few of the above mentioned features. Attempts to develop a more representative approach result in over-parameterisation. There is need for approaches that are capable of simulating sequences that are representative of the observed precipitation record, are computationally robust, and, are intuitively simple to understand, accept, and use in real applications. One such approach is presented in this paper.

Our rainfall simulation strategy is based on nonparametric nearest neighbour methods [*Scott*, 1992]. It involves conditionally resampling values from the observed record based on assumptions about the dependence structure of wet/dry spells and rainfall amounts. Use of nonparametric methods enables accurate representation of the distributional features that characterise rainfall. As a result, moments (mean, variance, skewness) and dependence properties (serial correlations and nonlinear conditional expectations) are both naturally modelled. Resulting simulations are hence more representative of the observed data than those from contemporary precipitation simulation models. We illustrate the use and utility of the approach by application to a 123 year long daily precipitation record from Sydney, Australia.

What follows is a discussion of traditional approaches for modelling daily precipitation. This is followed by a brief introduction on nonparametric methods and how they can be applied to simulation. Next, the proposed simulation approach is outlined. Results from application of the proposed approach to daily precipitation follows. We conclude with a discussion on the advantage of using the nonparametric approach over conventional precipitation models and mention some possible alterations to the present model structure that should increase its applicability to more diverse situations and scenarios.

2. BACKGROUND

2.1. Traditional approaches

Researchers from several disciplines have proposed both physical and stochastic approaches for synthesising daily rainfall. Complexities in the underlying rainfall generation mechanism limits the utility of physics based approaches. Statistical models treat rainfall as composed of two random processes – rainfall occurrence (denoting whether a day is wet or dry, or, more generally, the length of consecutive wet or dry days) and rainfall amount on a wet day. Rainfall occurrence has been modelled as a Markov process, the state (wet or dry) of the current and a few prior days deciding the state of the day that follows [*Stern and Coe*, 1984]. Transformations of the Markov transition probability have been used to rewrite the model as a generalised linear model (GLM) [*Nelder and Wedderburn*, 1972]. This has resulted in simplified procedures for parameter estimation and the ready use of commonly available statistical packages that have in-built GLM modules. The other approach that has been used is to represent the length of the current dry or wet spell as a random variable and model it as an alternating renewal process. Both approaches are appealing, although researchers have noted [*Woolhiser*, 1992] that parameter estimation is often a problem with alternating renewal process approaches.

Once a day has been classified as wet, the rainfall amount can be simulated using a number of plausible approaches. Unconditional simulation assuming a generic distribution, or conditional simulation conditioned on the state of the past few days (wet or dry) is an often used strategy. It has been noted [*Buishand*, 1977] that rainfall tends to have different probability distributions depending on whether the day is a solitary wet day, or whether it is bounded on one or both sides by another wet day. As a result, some models [*Buishand*, 1977; *Chapman*, 1997] try to fit distributions to each of these three classes. Once the distributions have been fitted, values are simulated depending on which of the above three classes the current day falls in.

Several probability distributions have been associated with rainfall spells and amounts. Some of the distributions that have been used to characterise spells lengths are – truncated negative binomial, truncated geometric and mixtures of two geometric distributions. Rainfall amounts form a highly skewed series and have been characterised by several positively skewed distributions such as gamma, shifted gamma, exponential, mixed exponential, kappa, and Weibull. Readers are referred to [*Woolhiser*, 1992] for a good review of daily rainfall simulation models and references to the distributional choices cited above. A "moving window" approach to model seasonality was recently presented by [*Rajagopalan et al.*, 1996] for modelling the seasonal features in rainfall states (the occurrence of a wet or dry day in a year). This approach is conceptually similar to the use of discrete "seasons" except that the discrete season is a period centered about the current day of interest (the period being called a "moving window" since it moves as we go to the next day in the time series). The need to choose an appropriate number of Fourier series terms for different statistics of interest, or to deal with discontinuities at the edge of each fixed discrete season, is obviated. This is the approach used in the work presented here.

A rainfall time series has a marked seasonal behaviour. Seasonality has traditionally been represented in two ways. Some researchers impose a seasonal trend on the parameters that describe the model (e.g., a polynomial or more commonly a Fourier representation for each of the important model parameters). In other cases, the model is applied to discrete segments of the year (seasons or months).

While traditional rainfall simulators work well under certain conditions, they usually require modifications if applied to data from other climatic regimes. In some cases different model forms (distributional or dependence characteristics) are needed for the same location in different seasons of the year. Some typical difficulties in using such models are:

- Representation of distributional characteristics – Distributions of rainfall amounts or spell lengths are crucial to the structure of the simulated time series. The selection of an appropriate distribution is difficult for a univariate and especially for a multivariate variable space.
- Representation of dependence characteristics – Models assuming a Markovian representation for rainfall reproduce the assumed memory well in the simulations. However, most applications assume an order one model for simplicity.
- Representation of seasonality – Modelling seasonality sometimes results in difficulties in parameter estimation, leading researchers to assume stationarity in discrete segments of the annual cycle. Such representations offer a poor approximation of seasonal characteristics in the rainfall data.

The simulation approach presented here offers a feasible alternative that does not have the problems mentioned above. Nonparametric estimates of the joint and conditional density are used, thus avoiding the problems associated with representation of distributional and dependence characteristics. The use of a local neighbourhood (moving window) for each conditional simulation reproduces seasonal characteristics effectively. A short background on nonparametric approaches and their utility in simulation and forecasting follows.

2.2. Nonparametric methods

A parametric regression model is fully defined or indexed by a finite set of parameters. Examples include linear regression, where the parameters are the slope and intercept, and a Gaussian probability density function that is completely specified by a scale and location parameter. *Scott* [1992] has an interesting discussion aimed at defining when an estimator is nonparametric. He suggests that a necessary condition for an estimator to be nonparametric is that it "work" for a large class of functions, e.g., all once and twice differentiable functions. An important point he makes is that for an estimator to be nonparametric, the influence of any data point x_i on the nonparametric regression estimate should vanish asymptotically for any $0<\varepsilon<|x-x_i|$. This is not true for parametric estimators. For example, it is well known that linear regression estimators are globally influenced by the presence of outliers.

A nonparametric density estimator works in a way similar to a nonparametric regression estimator. The density is estimated using observations within a local neighbourhood of the point of estimation. While the familiar histogram is an example of a crude nonparametric density estimator, kernel and nearest neighbour techniques are the more efficient choices. The underlying idea in these is same as the histogram – count the relative frequency of the data lying in a local neighbourhood about the point of estimation. For a histogram, this neighbourhood is a fixed discrete bin. With kernel techniques the neighbourhood depends on the extent of the kernel functions (smooth functions such as a Gaussian PDF, centred at each sample observation). With nearest neighbour techniques, the neighbourhood is composed of the "k" nearest neighbours of the estimation point, k being an appropriately chosen integer. *Scott* [1992] and *Silverman* [1986] are good introductory texts on nonparametric methods. Researchers in hydrology have only recently started applying these methods to hydrology. Some recent nonparametric hydrologic applications relevant to hydrologic simulation are: streamflow simulation and disaggregation using nearest neighbour and kernel methods [*Lall and Sharma*, 1996; *Sharma et al.*, 1997; *Tarboton et al.*, 1997]; simulation of rainfall using a nonhomogenous Markov model [*Rajagopalan et al.*, 1996]; simulation of rainfall spells using a seasonally homogenous resampling approach [*Lall et al.*, 1996]; and simulation of multivariate daily weather sequences using kernel methods [*Rajagopalan et al.*, 1997].

Multivariate probability density estimation is a straightforward extension of the univariate case. A conditional probability distribution is a slice of the joint distribution at a given conditioning plane (a conditioning line in case of an order one model). In the nearest neighbour context *Lall and Sharma* [1996] have developed a conditional probability distribution based on the nearest neighbours of the conditioning plane. This defines a relative probability or weight associated with each neighbour and forms the basis for the resampling approach presented in the next section. These conditional probabilities are estimated as:

$$p(i) = \frac{1/i}{\sum_{j=1}^{k} 1/j} \qquad (1)$$

where p(i) is the probability associated with the i'th nearest neighbour and k is the number of neighbours considered. Since the above equation depends only on the observed data and their relative locations in relation to the conditioning plane, issues related to distribution selection are automatically avoided. Additionally, problems of boundary leakage (of importance to most hydrological variables including precipitation) are automatically resolved. For more details about the conditional probability estimator in (1), its statistical attributes, and illustrations of its use in the context of streamflow simulation, readers are referred to *Lall and Sharma* [1996].

The number of neighbours "k" can be chosen using any appropriate order selection strategy such as Generalised Cross Validation (GCV) [*Craven and Wahba*, 1979]. *Lall and Sharma* [1996] also suggest an ad-hoc rule for choosing k, stated as:

$$k = \sqrt{n} \qquad (2)$$

where n is the length of the observed sample record. This rule has been tested and found to compare well with the GCV chosen optimal. For simplicity, we shall use the above rule in the nonparametric daily rainfall simulation model presented next.

3. PRECIPITATION SIMULATION MODEL

This section describes the nonparametric model for simulation of daily precipitation sequences. The model consists of two phases. The first phase involves conditional simulation of wet and dry spell lengths. The second phase involves simulation of precipitation amounts for each wet spell. All significant dependencies (in spells or amounts) are modelled. Seasonality is modelled by considering a "moving window" centred at the day being simulated. The window forms a local (hence seasonally representative) subset of the data. Values (for all years) falling within this window are used to estimate the conditional probability. Similar windows are used for simulation of both rainfall spells and amounts. Algorithmic details for each of these are presented next.

3.1. Model algorithm

Here we present the algorithm used for simulation of daily rainfall. The algorithm consists of three parts
A – Pre-processing of observations
B - Simulating dry and wet spell lengths, and,
C - Simulating rainfall amounts for all days in a wet spell.

3.1.1. Pre-processing

Both discrete and continuous variables are used in formulating the nonparametric simulation procedure. This necessitates the use of scaling factors to ensure that each variable has the same weighting in selection of nearest neighbours. For example, simulation of rainfall amounts (discussed later) uses the amount for the previous day, the day from start of the spell, and the length of the wet spell as the three variables that comprise the conditioning vector. Hence, standard deviations of each are estimated before hand and variables scaled when estimating euclidean distances for identification of nearest neighbours. Localised estimated of the standard deviation are used to account for possible seasonal variations. The scaling weights are then estimated as:

$$w_{jt} = 1/s_{jt} \qquad (3)$$

where w_{jt} denotes the weight for variable j on day t of the calendar year, and s_{jt} is the estimated local standard deviation (standard deviation for all values falling within moving window centered at day t of the year).

3.1.2. Spell lengths

The conditioning variable used for simulation of dry spell lengths is taken to be the number of days in the preceding wet spell. This choice was based on the strong (and statistically significant) dependency that was observed between these two variables. The conditioning variable for simulation of wet spell lengths is taken as the length of the preceding dry spell. The algorithm for the spell simulation component starting at day t in the calendar year is as follows:

1. Formulate moving window centred at day t of the year. The window records only the calendar dates that fall within the window. Observations corresponding to these dates (for all years of the observed record) then form the local subset used for resampling.
2. Identify the length of each dry spell originating within the moving window. Once this is done, record the following – the length of the wet spell that precedes the dry spell (n_{wi}) and, the length of the dry spell (n_{di}).
3. Identify the k nearest neighbours of the previous occurrence of the conditioning variable (the wet spell length n_{wt}). This is equivalent to estimation of the euclidean distance between the conditioning value (n_{wt}) and each of the wet spell lengths recorded in step 2. This distance is given as:

$$d_i = \left| n_{wt} - n_{wi} \right| \qquad (4)$$

where i ranges from 1 to the number of observations in the window. Observations having the k smallest distances are the k nearest neighbours of the conditioning plane.

4. Estimate the conditional probability in (1) (In reality this needs to be estimated only once and hence is estimated in the pre-processing stage). Randomly select a neighbour based on this estimate. The dry spell length corresponding to the chosen neighbour is the new simulation (this is denoted as n_{dt}).
5. Re-formulate the moving window centred at the new value of t (the day following the end of the dry spell – t+n_{dt} where n_{dt} is the dry spell length simulated in step 4). Identify the length of each wet spell that originates in the window and record the following – the dry spell length preceding the wet spell (n_{di}), and, the length of each wet spell (n_{wi}).
6. Identify the nearest neighbours using dry spell length as the new conditioning variable. Distances are now calculated using the equation

$$d_i = \left| n_{dt} - n_{di} \right| \tag{5}$$

where n_{dt} is the dry spell length simulated in step 4.

7. Randomly select a neighbour using the conditional probability of (1). The wet spell length corresponding to this selection completes the spell simulation part of the model.

As no prior values are available at the start of the simulation, the algorithm is initialised randomly by choosing a spell (wet or dry) in the moving window. If this is a dry spell, the following wet spell length is simulated using step 5 of the above algorithm. If it is a wet spell, rainfall amounts for each of the days are simulated using the procedure described next.

3.1.3. Rainfall amounts

Rainfall amounts are conditioned on three variables – the day from start of the wet spell, the length of the current wet spell, and the amount on the previous day in the spell (not considered for the first day of the spell). The algorithmic details of the rainfall amount simulation phase in the model are as follows:

1. Formulate moving window centred at the first day of the wet spell.
2. Identify the rainfall amounts (non-zero values) that fall within the window for the entire historical record. This forms the local subset that is used to resample the daily rainfall values. Record the following variables – days from start of the spell (l_i), the length of the current wet spell (n_{wi}), the amount for the present day (p_i), and the amount for the day that follows (p_{i+}). Days for which p_{i+} is zero (last day in wet spell) are not considered.
3. Identify the k nearest neighbours of the conditioning plane. These neighbours are chosen using a scaled distance measure. The conditioning variables used are the number of days from start of the wet spell (l_t), the length of the current wet spell (n_{wt}), and the rainfall amount on the previous day in the spell (p_t). The third variable (rainfall amount) is not used when simulating the first day in the spell. The distance is calculated as:

$$d_i = \sqrt{\left(w_{1t}(l_t - l_i)\right)^2 + \left(w_{2t}(n_{wt} - n_{wi})\right)^2 + \left(w_{3t}(p_t - p_i)\right)^2} \tag{6}$$

where w_{1t}, w_{2t} and w_{3t} are the weights associated with the three conditioning variables (these weights being estimated in the pre-processing stage using equation (3)).

4. Randomly identify a neighbour using the conditional probabilities in (1). The rainfall amount for the day that follows the chosen day (p_{i+}, if i is the chosen day) is the new simulated amount.
5. Repeat the procedure till end of spell is reached.

The above procedure is simple and easy to use and provides a natural way to model dependence in the rainfall time series. Use of the day from start of spell as a conditioning variable provides an easy way to approximate any higher order dependence that may be present in individual spells. The dependence of probability distributions on the nature and length of the spell is adequately modelled by use of wet spell length as one of the conditioning variables. Use of the moving window provides an efficient way to model seasonal features. The conditional probability in (1) ensures that sample distributional properties are appropriately represented in the simulations. Although more elaborate strategies for identifying the length of the moving window could be formulated, we chose a fixed value of 60 days in our applications. Applications of the nonparametric simulation model to 123 years of daily rainfall data from Sydney, Australia is given next.

4. APPLICATION

Here we describe the application of the nonparametric daily rainfall simulation model described in the previous section to 123 years of daily rainfall (1859 to 1981) from Sydney, Australia. Seasonal characteristics of the observed record are illustrated in Figure 1. While the first three variables have marked seasonal characteristics, skewness does not show a clear seasonal trend. It is interesting to note that the average precipitation is relatively lower in the "wetter" second half of the year. Sudden short bursts of rainfall in the first half could be a possible reason for this behaviour.

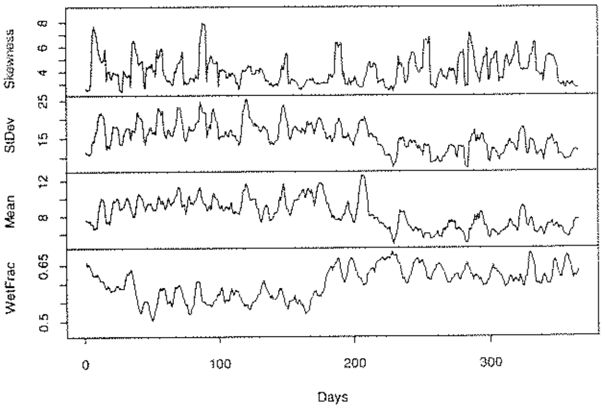

Figure 1

Seasonal variability in historical daily precipitation record. "WetFrac" denotes the fraction of wet days. "StDev" refers to the estimated standard deviation. All statistics are estimated using a fixed window length of 60 days. Only non-zero values are used in estimating the mean, standard deviation and skewness from the data.

The nonparametric simulation model was applied to simulate 30 samples of daily rainfall of the same length as the historical record (123 years). Figure 2 illustrates seasonal features of one of the simulated samples. Note the similarity in Figures 1 and 2.

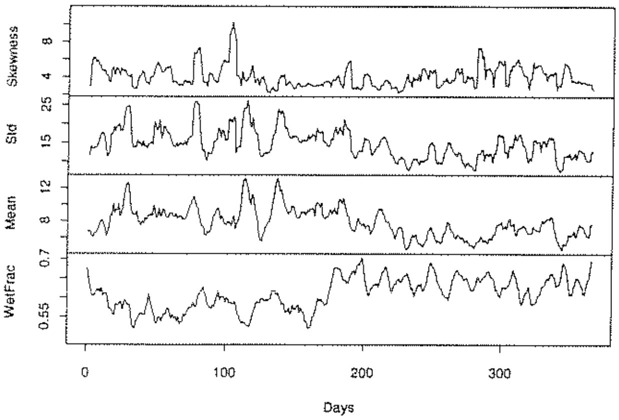

Figure 2
Seasonal features of one of the samples simulated using the nonparametric rainfall model.

For ease of presentation, the results from the simulations are presented at a monthly scale. Figure 3 illustrates average dry and wet spell lengths for each month. Standard deviations of the dry and wet spell lengths are shown in Figure 4. The continuous line in both figures shows the historical statistic (average or standard deviation as the case may be). A boxplot is used to illustrate the variability in each statistic across the 30 simulations. A boxplot, as used here, consists of a line in the middle of the box denoting the 50% quantile (median), the box with edges representing 25% and 75% quantiles, and whiskers that extend to the 5% and 95% quantiles of the statistics shown.

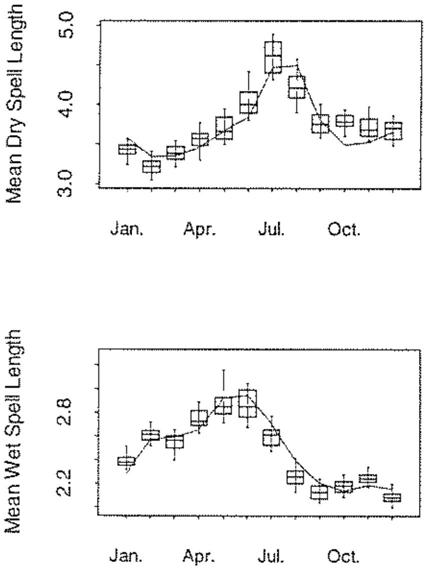

Figure 3
Average observed and simulated dry and wet spell lengths (in days).

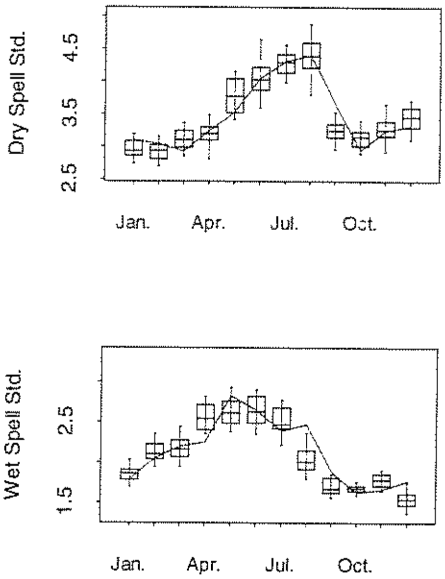

Figure 4
Standard deviations of observed and simulated dry and wet spell lengths. "Std." refers to standard deviation in the figure.

The simulated spell length means and standard deviations (Figures 3 and 4) compare well with the observed spell lengths. It is noticeable that both wet and dry spell lengths have marked seasonal characteristics, both in the averages and the standard deviations. These features are reproduced well in the simulations from the nonparametric model.

Figure 5 illustrates an important feature of the Sydney rainfall time series. Shown are the observed and simulated average rainfalls associated with wet spells that last 1, 2, 3, 4, 5, and 6 or more days. Interestingly, for most months the average rainfall amounts tend to increase as spells get longer. This is contrary to intuition although researchers [*Buishand*, 1977; *Chapman*, 1997] have noted that average rainfall on solitary wet days is usually smaller than for longer spells. Simulation strategies that assume stationarity in rainfall amounts for all spell lengths are unlikely to produce meaningful results. Models that specify different distributions for days that are bounded by 0, 1 or 2 wet days [*Buishand*, 1977], would perform better. However, assumptions of stationarity for long spells (longer than 2 days) would distort the simulated rainfall amounts.

As can be observed from Figure 5, the nonparametric simulation model is able to accurately represent this nonstationarity in the rainfall generation mechanism. Use of l_t, the number of days from start of the spell, and n_{wt}, the wet spell length, as two of the conditioning variables in the simulation algorithm (see previous section) is the likely cause of this accuracy.

Several other characteristics of the simulations were estimated and compared to the observations. Some of these were distributional characteristics (nonparametric estimates of the probability density of rainfall for different months of the year), higher order moment statistics (skewness and kurtosis) and dependence characteristics in the spell and amount processes. These are not presented for lack of space but are available from the authors on request.

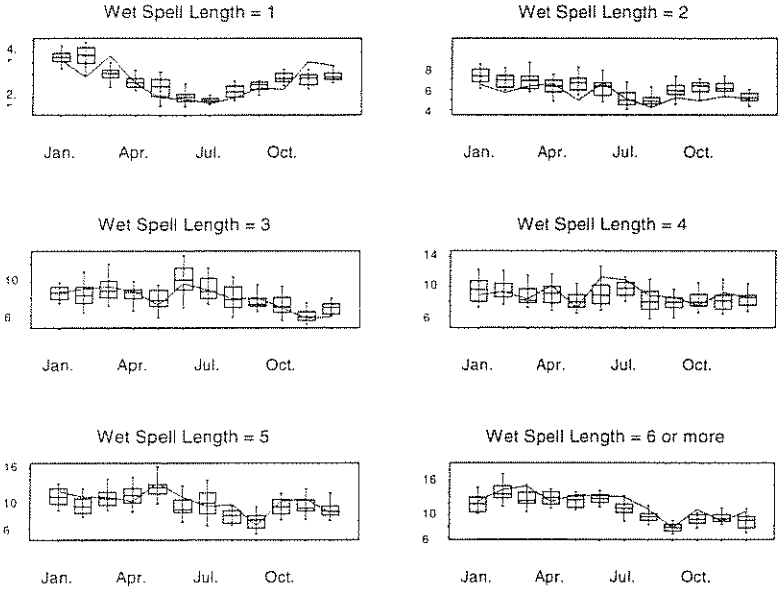

Figure 5
Average rainfall (in mm) as a function of spell length. Spell lengths of 1, 2, 3, 4, 5 and 6 or more days are considered.

5. SUMMARY AND CONCLUSIONS

A nonparametric approach for daily rainfall simulation was developed. Nonparametric nearest neighbour methods were employed to estimate the conditional probabilities of rainfall spell and amounts. These nearest neighbours were selected from a local neighbourhood of the current calendar day. Use of this local neighbourhood resulted in an accurate representation of seasonal characteristics in the observed rainfall record.

A conditional simulation approach was used to simulate dry and wet spell lengths. The simulated dry spells were conditioned on the number of days in the prior wet spell. Wet spell lengths were simulated conditional to the previous dry spell length. The nearest neighbour approach provided a simple framework to formulate the conditional probabilities needed for simulation.

Rainfall amounts were simulated once the wet spell length had been obtained. Simulation of amounts was conditional to two variables – the rainfall amount on the previous day (assuming order one dependence in the rainfall time series), and the number of days from start of the current wet spell. The nearest neighbour methodology was again used.

The model was applied to simulate daily rainfall from 123 years of data from Sydney, Australia. The simulations were tested for their ability to reproduce seasonal, dependence and distributional characteristics. Results from these tests illustrated the ability of the nonparametric model to simulate samples that are "representative" of the historical record.

Although model simulations are not currently capable of representing long range dependence characteristics as represented by long-term climatic fluctuations, work is currently in progress in this direction. Use of additional conditioning variables that have similar periodicities as are observed in atmospheric indices such as the Southern Oscillation Index, is being investigated. Results from this shall be discussed in another paper.

6. REFERENCES

Buishand, T.A., Stochastic modeling of daily rainfall sequences, *Meded. Landbouwhogesch. Wageningen, 77* (3), 211 pp, 1977.

Chapman, T.G., Stochastic models for daily rainfall in the Western Pacific, *Maths and Computers in Simulation (in press)*, 1997.

Craven, P., and G. Wahba, Smoothing noisy data with spline functions, *Numerical Mathematics, 31*, 377-403, 1979.

Lall, U., B. Rajagopalan, and D.G. Tarboton, A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resources Research, 32* (9), 2803-2823, 1996.

Lall, U., and A. Sharma, A nearest neighbor bootstrap for time series resampling, *Water Resources Research, 32* (3), 679-693, 1996.

Nelder, J.A., and R.W.M. Wedderburn, Generalized linear models, *Journal of Royal Statistical Society A, 135*, 370-384, 1972.

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

Rajagopalan, B., U. Lall, D.G. Tarboton, and D.S. Bowles, Multivariate Nonparametric Resampling Scheme For Generation of Daily Weather Variables, *Stochastic Hydrology & Hydraulics, 11* (1), 65-93, 1997.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualisation*, 317 pp., John Wiley & Sons Inc., New York, 1992.

Sharma, A., D.G. Tarboton, and U. Lall, Streamflow Simulation - a Nonparametric Approach, *Water Resources Research, 33* (2), 291-308, 1997.

Silverman, B.W., *Density estimation for statistics and data analysis*, 175 pp., Chapman and Hall, New York, 1986.

Stern, R.D., and R. Coe, A model fitting analysis of daily rainfall data, *Journal of Royal Statistical Society A, 147* (1), 1-34, 1984.

Tarboton, D.G., A. Sharma, and U. Lall, Disaggregation Procedures for Stochastic Hydrology based on Nonparametric Density Estimation, *Water Resources Research (under review)*, 1997.

Woolhiser, D.A., Modeling daily precipitation - progress and problems, in *Statistics in the environmental and earth sciences*, edited by A.T. Walden, and P. Guttorp, pp. 306, Edward Arnold, London, U.K., 1992.