A Semi-Parametric Stochastic Model For Simultaneous Stochastic Simulation Of Daily Precipitation Amounts At Multiple Sites

Mehrotra, R. and A. Sharma

The University of New South Wales, Sydney, NSW, Australia, E-Mail: raj@civeng.unsw.edu.au

Keywords: *Markov-chain; Rainfall; Weather generator; Spatial correlation; KDE approach.*

EXTENDED ABSTRACT

A semi-parametric stochastic model for generation of daily precipitation amounts, simultaneously at a collection of stations in a way that preserves realistic spatial correlations, accommodates seasonality, and reproduces a number of key aspects of the distributional and dependence properties of observed rainfall is described and illustrated. Following a conventional weather generator formulation, rainfall occurrences are modeled at the first stage and the rainfall amounts on simulated wet days are modeled subsequently. The rainfall occurrences at each individual site are simulated using a two-state, second-order Markov model. This model is found to produce better results for the statistics of long wet and dry spells. The rainfall amounts on the simulated wet days are generated using a non-parametric kernel density based approach. The amount model is conditioned on the rainfall amount on the previous day. Multisite spatial correlations of rainfall occurrences and amounts are reproduced by driving the single-site models with spatially correlated random numbers following a procedure described in Wilks (1998). The seasonal transition in the generation process is maintained by estimating the correlations on a day-to-day basis using a moving window formulation. The procedure of simulating rainfall at individual station and introducing the spatial dependence by means of spatially correlated random numbers, allows more flexibility to model temporal rainfall attributes of importance at individual station without introducing unnecessary complexity. The model is applied on a network of 30 raingauge stations around Sydney in Australia and the results evaluated. The study region exhibits substantial topographic and spatio- temporal rainfall variations, and thus provides a challenging setting to evaluate the simulation model. The analyses of the results show that the model is able to reproduce successfully the spatial correlations of rainfall occurrence and amounts (as shown by the scatter plots of Figure 1) and temporal rainfall characteristics (for example, number of wet days and average rainfall amount as shown in Figure 2) of general interest to the hydrologists. In addition, rainfall characteristics at higher time scale are also found to be captured well by the model.

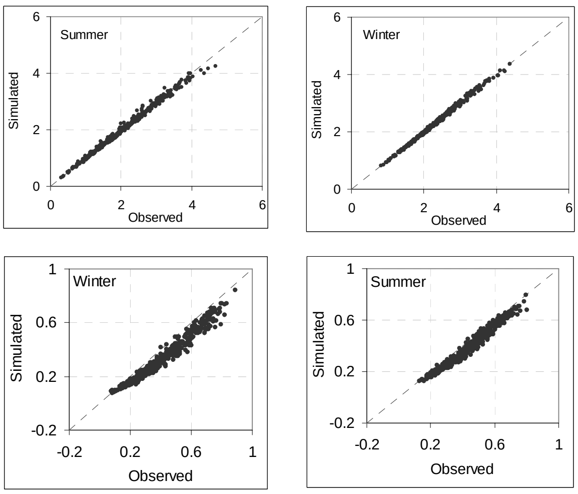

Figure 1. Observed and modeled log-odds ratios and cross correlations at all stations and seasons.

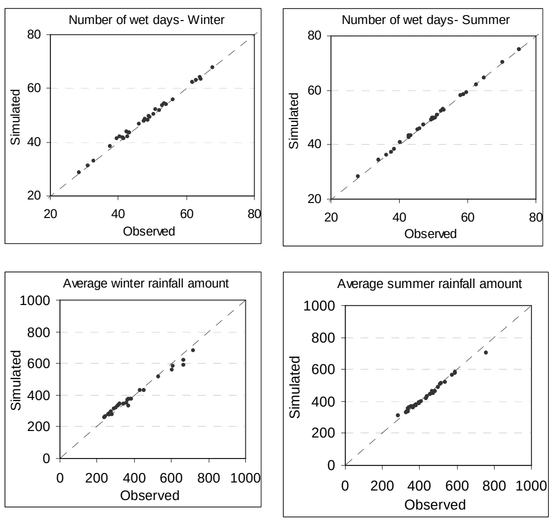

Figure 2. The scatter plots of observed and modelled number of wet days and average rainfall amount at all stations and seasons.

1. INTRODUCTION

Simultaneous simulation of weather variables at multiple point locations is often desired in many hydrological and agricultural ecosystem models. An important limiting factor in the applications of these variables in modeling activities can be their availability with sufficient temporal and spatial coverage. To overcome this, stochastic models sometimes also known as 'weather generators' are commonly used to generate synthetic sequences of weather variables that are statistically consistent with the observed characteristics of the historical record in time. These synthetic sequences provide information in enhancing our understanding of hydrological system response, and in the design and operation of water resource systems. The single site weather generators based on Markovian dependence are based on a relatively simple stochastic process and are easy to formulate and implement. However, the single-site weather generators cannot satisfactorily reproduce the strong spatial dependence among weather variables, often necessary to evaluate the hydrological or agricultural behavior of a region. The spatial correlation of weather variables, specially the precipitation, may have essential effects on the discharge of a river and the formation of floods. It is nowadays widely accepted that multi-station precipitation simulation can only be achieved with the weather models that preserve spatial correlation among stations.

This vital issue of spatial dependence among series of weather variables at multiple locations has been addressed in a number of space-time applications (e.g. Bras and Rodriguez-Iturbe, 1976; Hay *et al*., 1991; Hughes and Guttorp, 1994; Bardossy and Plate, 1992; Hughes *et al*, 1999; Stehlík and Bárdossy, 2002). These approaches are comparatively complex in both calibration and implementation, and therefore have been only modestly successful in multi-site simulation of rainfall and other weather variables.

In an effort to alleviate the problem of the excessive number of parameters of multi-site weather simulators, Wilks (1998) proposed a simple extension of commonly used single-site weather simulators primarily for rainfall, to multi-sites by driving each of the individual single-site rainfall occurrence and amounts models with temporally independent but spatially correlated random numbers. This logic has since than been applied to varieties of multi-site simulation studies including, simulation of daily precipitation, daily maximum temperature, daily minimum temperature, and daily solar radiation (Wilks, 1999a); downscaling of precipitation (Wilks, 1999b); downscaling of daily precipitation, and daily maximum and minimum temperatures conditional on daily circulation patterns (Qian *et al*., 2002); generation of rainfall occurrences (Mehrotra *et al.*, 2005) and downscaling of rainfall occurrence and amounts (Mehrotra and Sharma, 2005).

The modification proposed by Wilks (1998) simulates amounts on wet days assuming temporal independence. However, Buishand (1978) found significant correlation between precipitation amounts on successive wet days. Gregory *et al.* (1993) suggest that reproduction of the structure of daily autocorrelation provides a crucial test for a stochastic rainfall model.

The kernel density approach belongs to a class of models that are nonparametric, or, models where assumption about the form of the probability distribution or dependence between variables is minimal. This method has been used extensively for a range of hydro-climatological applications (Sharma *et al.*, 1997; Sharma, 2000; Sharma and O'Neill, 2002; Harrold *et al.,* 2003; Mehrotra and Sharma, 2005).

This paper presents a mix of parametric and nonparametric approaches for the generation of rainfall at multiple locations. The paper is organised as follows. The methodology and the models used are discussed in Section 2. Details on the application of the various models considered, the data and the study region used, and a comparison of the various results obtained, are presented in Section 3. We conclude the paper by presenting the summary and conclusions drawn from the results in section 4.

2. METHODOLOGY

With the conventional weather generator formulation, it is quite common to model first precipitation occurrence (whether a day is wet or dry at a location) and subsequently model the precipitation amounts on simulated wet days. As rainfall occurrence and amounts are simulated independently at each location, observed spatial dependence across the stations is not directly reproduced and is introduced by using spatially correlated random numbers. This procedure of simulating rainfall at individual station and introducing the spatial dependence by means of spatially correlated random numbers, allows more flexibility to model temporal rainfall attributes of importance at individual station without introducing unnecessary complexity.

2.1. Rainfall occurrence

Daily precipitation occurrence is usually modeled using a two-state, first-order Markov chain, according to which the probability of precipitation depends only on whether the previous day was wet or dry. Here, a slightly more complex Markov model for precipitation occurrence was used (Stern and Coe, 1984; Wilks, 1999a,b), involving four precipitation probabilities: (1) the probability of precipitation following two consecutive dry days (p001), (2) the probability of precipitation following two consecutive wet days (p111), (3) the probability of precipitation if the previous day was dry but the day before that was wet (p101), and (4) the probability of precipitation if the previous day was wet but the day before that was dry (p011). The two state second-order Markov model is used here because it produces better results for the statistics of long wet and dry spells, and also helps in identifying whether the wet day belongs to the start or middle of a wet spell. Sets of these four transition probabilities are estimated separately for each location and for each day using a moving window formulation.

2.2. Rainfall Amounts

A nonzero precipitation amount must be generated for each day and location that the model described in Section 2.1 simulates to be wet. Conventionally, parametric methods generate nonzero precipitation amounts by fitting a single probability distribution to all locations and months. The model for rainfall amounts that is presented in this paper is nonparametric, and is based on the kernel density procedure as described in Sharma *et al.* (1997); Sharma (2000) and Harrold *et al.* (2003). It simulates the rainfall amounts at individual stations conditional on the previous days' rainfall. The use of rainfall amounts on the previous day as a conditioning variable imparts a Markovian order dependence to the simulated series.

2.3. Modeling Spatial Correlations of Rainfall Occurrence and Amounts

Rainfall occurrence and amount models outlined in sections 2.1 and 2.2 generate series of occurrence and amounts at individual stations, hence resulting in values that are theoretically spatially uncorrelated. We induce spatial dependence in the generated rainfall fields by making use of spatially correlated and serially independent random numbers in generation of rainfall occurrence and amounts at individual stations separately, adopting a procedure outlined in Wilks (1998). The general procedure of estimating the correlation matrices of random numbers for rainfall occurrence and amounts models is explained next. It may be noted that these matrices are estimated independently for occurrence and amounts models.

Denote $\mathbf{u}_t$ as a vector of uniform [0,1] variates of length n_s at time step t, with n_s being the number of stations. The vector $\mathbf{u}_t$ ($\equiv u_t(1), u_t(2), ...u_t(n_s)$) is defined such that for locations k and l, $corr[u_t(k), u_{t+1}(l)] = 0$ (or, random numbers are independent across time), but $corr[u_t(k), u_t(l)] \neq 0$ (or, random numbers are correlated across space). As a result, there is spatial dependence between individual elements of the vector $\mathbf{u}_t$, this dependence being introduced to induce observed spatial dependence in the response variables they are used to simulate.

We denote the response series at a station pair k and l as $Y_t(k)$ and $Y_t(l)$, respectively. For rainfall occurrence these series comprise binary variables (0 if rainfall is less than 0.3 mm and 1 if it is equal to or greater than 0.3mm, after Buishand, (1978) and Harrold *et al.* (2003)) while for rainfall amounts these include only those observations when rainfall is equal to or greater than 0.3 mm. The sample coefficient of correlation between the two variables $Y_t(k)$ and $Y_t(l)$ is written as:

$$\xi(k,l) = Corr[Y_t(k), Y_t(l)] \qquad (1)$$

The spatial dependence between the uniform random variates $u_t(k)$, $u_t(l)$ is specified such that the resulting quantiles from the conditional CDF exhibit a correlation equivalent to $\xi(k,l)$. This spatial dependence is specified by first transforming the uniform random variates to those of a standard Normal distribution:

$$u_t(k) = \Phi[v_t(k)] \qquad (2),$$

where $\Phi[.]$ indicates the standard normal CDF. Let the correlation between the standard Normal variates be denoted as:

$$\omega(k,l) = Corr[v_t(k), v_t(l)] \qquad (3)$$

Our aim then becomes to find a value for $\omega(k,l)$ such that the Uniform random variates $u_t(k)$, $u_t(l)$ (or equivalently $v_t(k)$, $v_t(l)$) generated based on it, lead to rainfall series $Y_t(k)$, $Y_t(l)$ (for rainfall occurrences these are series of 0 and 1 while for amounts these are non zero values i.e. values greater than or equal to 0.3 mm) that exhibit a correlation of $\xi^o(k,l)$, which denotes the observed value of $\xi(k,l)$ (estimated as the sample coefficient of correlation of the observed rainfall series (of occurrence or amounts) $Y_t^o(k)$ and $Y_t^o(l)$ at stations k and l). Estimating this value involves an empirical procedure that is described below.

Direct computation of $\omega(k,l)$ from $\xi^o(k,l)$ is complex as $v_t(k)$ or $v_t(l)$ corresponding to observed rainfall series $Y_t^o(k)$ or $Y_t^o(l)$ is difficult to specify. However, as noted by *Wilks* (1998), there exists a monotonic relationship between $\omega(k,l)$ and the simulated $\xi(k,l)$ for a given station pair k and l. Therefore, in practice, one obtains $\omega(k,l)$ using a trial and error by assigning a reasonable value to $\omega(k,l)$, simulating the series $Y_t(k)$ and $Y_t(l)$ at stations k and l, evaluating $\xi(k,l)$ and comparing it with $\xi^o(k,l)$. The process is repeated with different values of $\omega(k,l)$ till an acceptable value of $\xi(k,l)$, reasonably close to $\xi^o(k,l)$, is obtained. It is also possible to invert the relationship between $\omega(k,l)$ and $\xi(k,l)$ using a nonlinear root finding algorithm. The procedure is repeated with all possible combinations of station pairs and a final correlation matrix $\boldsymbol{\omega}$ (n_s x n_s) of normal deviates separately for rainfall occurrence and amounts is obtained. Further details on the methodology are available in Wilks (1998) and Mehrotra and Sharma (2005).

Note that the estimation of the correlation coefficient of the normally distributed random numbers as described here is assumed independent of the process used to drive the rainfall generation mechanism. That is, although, rainfall occurrences or amounts are generated by using the second order Markov model or conditional KDE model, the correlations are computed independent of the assumed conditional distribution. This assumption while allows considerable simplifications in the calculations, has negligible affects on the final results as presented and discussed in the subsequent sections.

3. APPLICATION OF MODELS AND RESULT DISCUSSION

3.1. Data and Study Area

The study region is located around Sydney, eastern Australia spanning between 147°E - 153°E longitude and 31°S - 36°S latitude (Figure 3). For this study, a 43-year continuous record (from 1960 to 2002) of daily rainfall at 30 stations around Sydney, eastern Australia was used. The inter-station distances between station pairs vary approximately from 20 to 340 km. Missing values at some stations (<0.5%), were estimated using inverse distance averaging and the records of nearby stations.

3.2. Model Application

Relationship between correlations of series of normally distributed random numbers and simulated rainfall occurrence and amount series at a station pair is estimated on a daily basis considering the observations falling within a moving window of length 31 days centred on the current day. Rainfall occurrences at individual sites are generated using a two state, second order Markov model as described in sub section 2.1, while rainfall amounts on wet days are generated conditional on the previous day rainfall amount using a nonparametric kernel density approach as described in sub section 2.2.

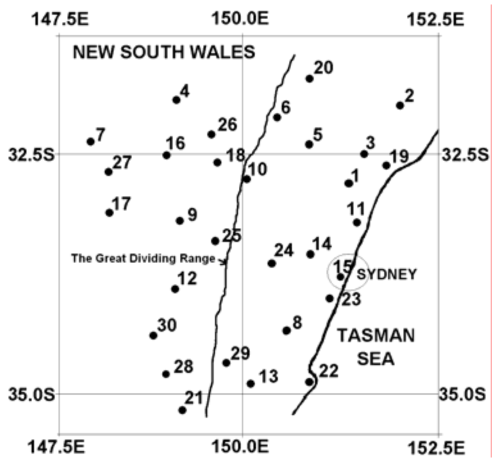

Figure 3. Map of the study area showing the locations of rain-gauge stations.

It was observed that the procedure used for generation of rainfall amounts considers the correlation matrix of random numbers derived using the rainfall series of wet days only and therefore the correlation of simulated rainfall amounts on occasions when only one station (of a station pair) is wet, did not match well with the observed ones. This situation occurs when a wet station lies close to the boundary of a wet region. We address this issue by incorporating two additional variables in the predictor set representing respectively, the ratio of number of wet stations to the total number of stations in the neighbourhood (within a given radius) and, the average distance of neighbouring wet stations from the station under consideration. Following Mehrotra and Sharma (2005), a region represented by a radius of 150 kilometres was found to be the most effective in representing the influence of local neighbourhood. The conditioning vector considered for simulation of rainfall amount at a station thus includes previous day rainfall and two variables defining the local wetness density, totaling three variables.

3.3. Model results

In all the results that follow, the statistics reported are ascertained by generating 100 realisations of the rainfall from the model. The performance of the occurrence and amounts models is evaluated on a seasonal basis for their ability to simulate the observed spatial and temporal characteristics of rainfall including those of importance in water resource management. These include number of wet days, average daily rainfall amount, wet and dry spells, maximum daily rainfall amount and mean and standard deviation of monthly rainfall totals. Due to space limitation, however, results of only a few important statistics are reproduced here.

Spatial dependence of rainfall occurrence and amounts

The log-odds ratio, reflecting the spatial correlation between rainfall occurrences at each pair of stations provides a measure of accurate reproduction of the spatial dependence. The top row of Figure 1 presents observed and modeled log-odds ratios at all stations for both seasons. Each point on the graph indicates the ratio evaluated for a pair of raingauge stations. The dependence between the stations is accurately preserved by the model.

Bottom row of Figure 1 presents the scatter plots of cross correlations in the observed and simulated daily rainfall amounts for all station pairs and seasons. The amount model provides a good fit to the rainfall amounts except some bias during winter.

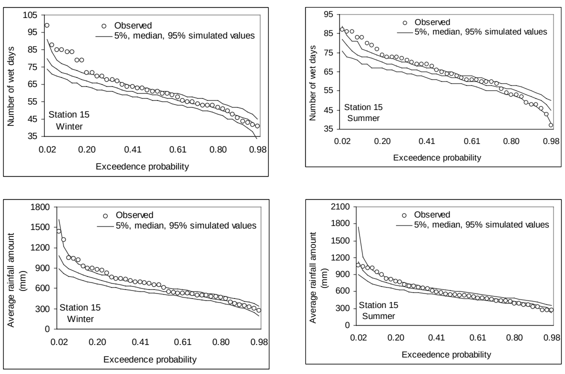

Figure 4. Distribution plots of wet days and seasonal rainfall amount in the observed and simulated rainfall series at the representative station for both seasons.

Number of wet days and average rainfall amount

It is vital that the average number of wet days and rainfall amount be reproduced accurately before using the generated rainfall series as an input to any water balance modeling exercise. Top row of Figure 2 presents the scatter plots of observed and modelled number of wet days at all stations for both seasons. Each point on the graph represents a station. As can be seen from the graph, the model provides a good fit to the number of wet days at all stations. Bottom row of Figure 2 shows the scatter plots of the observed and model simulated seasonal rainfall amount for all stations and both seasons. The model provides an accurate simulation of seasonal rainfall amounts.

For efficient design and management of water resource projects, not only the number of wet days but their accurate distribution in the simulated series is also important. Top row of Figure 4 compares the distribution plots of wet days in the observed and simulated rainfall series for a representative station only (no 15 of Figure 3, picked up to represent the coastal wet region). The 5th percentile, median, and 95th percentile values are shown as continuous lines while the historical values are superimposed as circles. As shown in the plot, the model adequately reproduces the distribution of wet days for both seasons. Similarly, bottom row of Figure 4 compares the distribution of observed and simulated seasonal rainfall amounts at the representative station. In general, performance of the model in reproducing the distribution of seasonal rainfall amount and number of wet days is satisfactory, except for the upper ends of the graphs, where it somewhat under estimates the extreme values.

Wet and dry spell characteristics

The generated series intended for use in catchment modeling studies should be capable of reproducing sustained wet and dry days. Figure 5 presents the scatter plots of lowest, median and highest yearly maximum wet and dry spells at each station. As shown in the plots, these characteristics are not well reproduced by the model; specifically the highest dry spells are consistently under estimated at majority of stations. It appears that the second order Markov model used in the study is not capable enough to capture these longer time scales characteristics. Perhaps, conditioning on additional variables incorporating longer time memory of

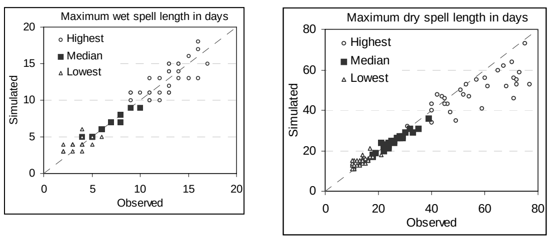

Figure 5: Scatter plots of observed and modeled yearly highest, median and lowest maximum wet and dry spells at all stations.

rainfall series as suggested by Harrold e*t al.* (2003) might help in improving the results further.

Daily maximum rainfall

Accurate estimation and distribution of the observed rainfall peaks at a station is of significance in catchment studies dealing with flood estimation and analysis. It has essential effect on the maximum discharge that can be expected from the catchment. Figure 6 provides the distribution plots of observed and modeled daily maximum rainfall at the representative station for winter and summer. On an average, the model is quite successful in simulating this important rainfall characteristic.

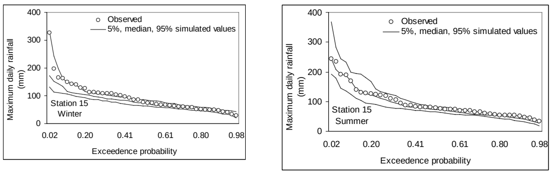

Figure 6: Distribution plots of observed and modeled daily maximum rainfall at the representative station for both seasons.

Implied monthly rainfall characteristics

As the model is formed considering the daily rainfall occurrence and amount series, it would of general interest to investigate further the capability of the model in modeling the longer time scale rainfall attributes. These longer time-scale characteristics can be of essential interest for some applications, and constitute a challenging assessment of the performance of the daily model (Buishand, 1978; Wilks, 1998). Figure 7 presents the scatter plots of a few statistics of the aggregated observed and generated rainfall series at monthly level including: standard deviations of monthly rainfall totals (panel a), standard deviations of number of wet days in a month (panel b) and, cross correlations of monthly rainfall totals and number of wet days in a month (panels c and d). The points on the graph are shown for all months and stations for panels a and b and for all station pairs and months for panels c and d. Since the monthly means are determined by the mean number of wet days per month and mean rainfall per day (Wilks, 1992; Wilks, 1998) and these observed statistics are well reproduced by the model (Figure 2), the plots of the averages of monthly rainfall totals and number of wet days are not presented here. As can be seen from the Figure 7, there is consistent under estimation of standard deviation (indicative of the interannual variation of the series used) of monthly rainfall totals and monthly number of wet days. This deviation is more pronounced for higher standard deviations of monthly rainfall rainfalls. This kind of behaviour of aggregated series of rainfall has also been noted by others in the past (Buishand, 1978; Katz and Parlange, 1993; Wilks, 1998). Similarly, inter-site cross correlations of monthly rainfall totals and monthly number of wet days indicate a wide scatter across the stations (panels c and d). In these graphs, some of the deviation can be attributed to sampling uncertainty in the limited observational record (only 43 years of record is used in the analysis). In general, these statistics at monthly time scale are reasonably well represented by the model, even though the models were fitted to the daily rainfall series. We feel that inclusion of additional conditioning variables representing longer time memory of the rainfall series as suggested by Harrold e*t al.* (2003) might help improving these results. For rainfall amount, as individual models are fitted to each site, inclusion of additional variables in the nonparametric KDE model is easy and straightforward.

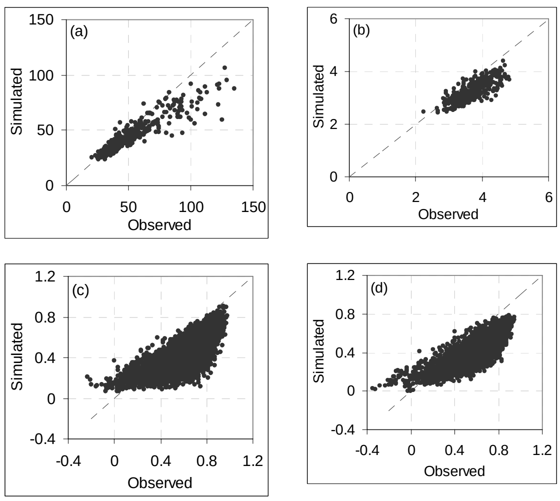

Figure 7. Statistics of observed and simulated monthly aggregated rainfall series: (a) and (b)- standard deviations of monthly rainfall amount and number of wet days in a month, (c) and (d)- cross correlations of monthly rainfall amount and number of wet days in a month. Points on the panels a and b are shown for each month and station, while for c and d are shown for each month and station pair.

4. SUMMARY AND CONCLUSIONS

This paper has demonstrated the applicability of a relatively simple rainfall generation framework for multi-site rainfall amounts. The approach simulates rainfall occurrences at all stations using a parametric Markov model, while rainfall

amounts on the wet days as specified by the occurrence model, are simulated using a conditional kernel density estimation procedure. Spatial correlations in the rainfall occurrence and amount series are induced by making use of spatially dependent random numbers.
Rainfall simulation models having the capability to simulate rainfall at networks of stations and maintaining spatial correlation structure are best suited for use in catchment management practice, where the nature of spatial variations in rainfall has important influences on streamflow and flooding.

The study region exhibits substantial topographic and spatio- temporal rainfall variations, and thus provides a challenging setting to evaluate the simulation model. Results of the study indicate the proposed simulation framework is effective in reproducing the various important spatial and temporal attributes of the observed rainfall occurrence and amount process over the region.

5. REFERENCES

Bardossy, A., and E.J. Plate (1992), Space-time models for daily rainfall using atmospheric circulation patterns, *Water. Resour. Res. 28*, 1247-1259.

Bras, R., Rodriguez-Iturbe, I. (1976), Rainfall generation: a nonstationary time varying multidimensional model, Water Resources Research, 12, 450–456.

Buishand, T.A. (1978), Some remarks on the use of daily rainfall models, J. Hydrol., 36, 295–308.

Gregory, J.M., T.M.L. Wigley and P.D. Jones (1993), Application of Markov models to area average daily precipitation series and interannual variability of seasonal totals, *Clim. Dynm*, *8*, 299-310.

Harrold, T.I., A. Sharma, and S.J. Sheather (2003), A nonparametric model for stochastic generation of daily rainfall amounts, Water Resour. Res., 39(12), 1300, doi:10.1029/2003WR002570.

Hay, L, J. McCabe, D.M. Wolock, and M.A. Ayers (1991), Simulation of precipitation by weather type analysis, Water Resour. Res. 27, 493-501.

Hughes, J.P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res. 30*, 1535-1546.

Hughes, J. P., P. Guttorp, and S.P. Charles (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *Appl. Stat. 48*(1), 15-30.

Katz, R.W., Parlange, M.B. (1993), Effects of an index of atmospheric circulation on stochastic properties of precipitation, *Water. Resour. Res., 29*, 2335-2344.

Mehrotra, R., R. Srikanthan and A. Sharma (2005), A comparison of three stochastic multi-site precipitation occurrence generators, submitted to *J. Hydrol*. For possible publication.

Mehrotra, R. and A. Sharma (2005), A nonparametric stochastic downscaling framework for daily rainfall at multiple locations, submitted to *J. Geophys. Res.* for possible publication.

Qian, B., J. Corte-Real, and H. Xu (2002), Multisite stochastic weather models for impact studies, *Int. J. Climatol, 22*, 1377–1397.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 — A nonparametric probabilistic forecast model, *J. Hydrol.*, 239, 249–258.

Sharma, A. and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, *38*(7), 5.1-5.10.

Sharma, A., D.G. Tarboton, and U. Lall (1997), Streamflow simulation: a nonparametric approach, *Water Resour. Res.*, *33*(2), 291-308.

Stehlík, J., and A. Bárdossy (2002), Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation, *J. Hydrol., 256*, 120-14.

Stern R.D. and R. Coe (1984), A model fitting analysis of daily rainfall data. Journal Royal Statistics Society. A 147:1-37.

Wilks, D.S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol., 210*, 178–191.

Wilks D.S. (1999a), Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain, Agri. and Forest Meteorol. 96:85–101.

Wilks, D.S. (1999b), Multisite downscaling of daily precipitation with a stochastic weather generator, Clim. Res., 11, 125-136.