HYDROLOGICAL PROCESSES
Hydrol. Process. **26**, 3182–3194 (2012)
Published online 10 January 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.8245

Probabilistic assessment of one-step-ahead rainfall variation by Split Markov Process

Rajib Maity*

Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India

Abstract:

In this paper, Split Markov Process (SMP) is developed to assess one-step-ahead variation of daily rainfall at a rain gauge station. SMP is an advancement of general Markov Process and specially developed for probabilistic assessment of change in daily rainfall magnitude. The approach is based on a first-order Markov chain to simulate daily rainfall variation at a point through state/sub-state transitional probability matrix (TPM). The state/sub-state TPM is based on the historical transitions from a particular state to a particular sub-state, which is the basic difference between SMP and general Markov Process. The cumulative state/sub-state TPM is represented in a contour plot at different probability levels. The developed cumulative state/sub-state TPM is used to assess the possible range of rainfall in next time step, in a probabilistic sense. Application of SMP is investigated for daily rainfall at four rain gauge stations – Khandwa, Jabalpur, Sambalpur, and Puri, located at various parts in India. There are 99 years of record available out of which approximately 80% of data are used for calibration, and 20% of data are used to assess the performance. Thus, 80 years of daily monsoon rainfall is used to develop the state/sub-state TPM, and 19 years data are used to investigate its performance. Model performance is assessed in terms of hit rate (*HR*), false alarm rate (*FAR*), and percentage captured. It is found that percentage captured is maximum for Khandwa (70%) and minimum for Sambalpur (44%) whereas hit rate is maximum for Sambalpur and minimum for Khandwa (73%). *FAR* is around 30% or below for Jabalpur, Sambalpur, and Puri. *FAR* is maximum for Khandwa (37%). Overall, the assessed range, particularly the upper limit, provides a quantification possible extreme value in the next time step, which is a very useful information to tackle the extreme events, such as flooding, water logging and so on. Copyright © 2011 John Wiley & Sons, Ltd.

KEY WORDS Split Markov Process (SMP); probabilistic assessment; rainfall variation; transitional probability matrix (TPM)

Received 28 March 2011; Accepted 28 July 2011

INTRODUCTION

Rainfall is one of the most complex and difficult component of the hydrologic cycle to model because of the complexity of the atmospheric processes and the wide range of variation in both space and time. However, prior information of rainfall is essential (both at large and small spatio-temporal scale) for proper planning and management of water resources. This is a high priority objective for developmental activities of a country, where the agricultural sector plays a key role for their economic growth. Large spatio-temporal variation of rainfall arises many water-related problems, such as flood and drought, which seriously affect the crop production. Reasonably, accurate rainfall prediction is required, which can help in alleviating such problems by planning for appropriate cropping patterns corresponding to water availability.

At smaller spatio-temporal scale, variation of rainfall has an effect on day-to-day life, such as water logging, heavy traffic jams, shutdown of airports, blackout problem and so on. Heavy rain may paralyse most of daily activities. High intensity of rainfall in Mumbai on 26 July 2005 caused a complete halt for the city, a large number of deaths (almost 1100) and an enormous loss of housing, trade and commerce, agriculture and cattle (as per the status report published by the government). An early information (at least a day before) could have helped in better management of the disaster. According to scientists at National Centre for Medium Range Weather Forecasting, which is a premier institute to provide medium range weather forecast in India, the predictions of severe weather events have enormous limitations (Bohra *et al*., 2006). Even though such events have a very short life but still cause extensive damage. Thus, even though the prediction of rainfall (spatio-temporal) is possible to achieve from numerical weather model, probabilistic information of rainfall could be an added advantage for the concerned community. The main purpose is to provide as much advance notice as possible to the people to save the human and animal lives and properties from an impending disaster. The focus of this paper is the variation of point rainfall at a particular station.

Use of probabilistic rainfall prediction has a long history to predict the near-future occurrence of extreme events (Box *et al*., 1976; Weeks and Boughton, 1987; Wójcik *et al*, 2003). A framework for probabilistic rainfall forecast using nonparametric kernel density estimator is presented in a series of three papers (Sharma, 2000a; Sharma *et al*., 2000; Sharma, 2000b). The approach is developed for station rainfall data. However,

*Correspondence to: Rajib Maity, Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur, 721302 West Bengal, India.
E-mail: rajib@civil.iitkgp.ernet.in, rajibmaity@gmail.com

Copyright © 2011 John Wiley & Sons, Ltd.

the temporal resolution is seasonal to interannual rainfall. Application of Markov Process (MP) for short-term rainfall forecast through a probabilistic way is well accepted for a long time (Gabriel and Neumann, 1962; Chin, 1977; Fraedrich and Müller, 1983; Stern and Coe, 1984; Rajagopalan *et al*., 1996; Jimoh and Webster, 1996; Kaseke and Thompson, 1997; Wilks, 1999; Hayhoe, 2000; Kottegoda *et al*., 2004; Baik *et al*., 2006; Deni *et al*., 2009). For instance, Gabriel and Neumann (1962) found that the first-order Markov chain model could be fitted to daily rainfall data at Tel Aviv in Israel. However, it was argued later that a second-order model would fit the data more suitably (Gates and Tong, 1976). Fraedrich and Müller (1983) predicted the probability of weather state by first order of Markov chains by using data of single station and forecasted daily sunshine measurements and rainfall combined with three hourly past weather observations. Stern and Coe (1984) used a nonstationary Markov chain to model the occurrence of daily rainfall along with Gamma distribution to model the amount of rainfall. Fraedrich and Leslie (1987) used a linear combination of probabilistic approach (Markov chain) and numerical weather prediction for short-term rainfall prediction. A first-order Markov process is a continuous-time process for which the future behaviour, given the past and the present, only depends on the present and not on the past and characterized by set of states and the transition probabilities P_{ij} between the states. Here, P_{ij} is the probability that the state in the next time step is j, given that the same is i at the present time step. Haan *et al*. (1976) developed the stochastic model, which was based on a first-order Markov process and used rainfall data to estimate the Markov transitional probabilities and simulated daily rainfall record of any length, which was based on the estimated transitional probabilities and frequency distributions of rainfall amounts and concluded that simulated data had statistical properties similar to those of historical data. Kaseke and Thompson (1997) developed the partially observed Markov process algorithms for rainfall runoff process model and considered the special case of the martingale estimating function approach on the runoff model in the presence of rainfall. Rajagopalan *et al*. (1996) estimated the daily transition probability matrices nonparametrically and estimated the transition probabilities through a weighted average of transition by kernel estimator. Based on the assumption that the daily rainfall occurrence depends only on the previous day's rainfall, first-order Markov chain model was reported by Kottegoda *et al*. (2004) to fit the observed daily rainfall in Italy. However, application of higher order Markov chain model was established to be suitable for stochastic weather generator for daily rainfall characteristics (Wilks, 1999; Hayhoe, 2000). Optimum order of Markov chain model for a particular data was also addressed (Chin, 1977; Jimoh and Webster, 1996). Deni *et al*. (2009) applied an optimum order model for daily rainfall in Peninsular Malaysia using the Akaike's information criteria and Bayesian information criteria.

Almost all these approaches follow a general path of creating a single set of different states depending on historical record, and the probabilities of transition from one state to another is obtained. However, for rainfall variation study in hydrology, the change in rainfall magnitude, particularly in higher side, is more crucial information as indicated before. Quantifying these changes through a single set of states demands large number of defined states. The word 'large' is subjective and implies more number of required states for more the inherent variability. Generally, in the tropical countries, the variation of daily rainfall is very high, and application of MP may not perform well. Moreover, probabilistic prediction is more useful than that of simple point prediction. Defining another set of sub-states, classifying the changes in magnitude of daily rainfall will be helpful for such probabilistic assessment. This is the theme of this paper. The objective of this study is to develop an approach for change prediction daily rainfall through state to sub-state transition, which is achieved through Split Markov Process (SMP). However, the approach considers daily rainfall in which sequential phases within a particular event of rainfall (e.g. initiation, growth, peak, decay and vanish) is not of interest. Rather, the total rainfall in a day is considered, which is important from water resources point of view. Thus, the transitions through states to sub-states are computed through state/sub-state transitional probability matrix (TPM) for a daily temporal resolution, which is used for probabilistic assessment of one-step-ahead rainfall variation. In MP, the transition from a particular state to another state is investigated. However, in SMP, the daily rainfall magnitude is categorized into different states, and change in magnitude from one temporal step to another is categorized into different sub-states for the probabilistic assessment of rainfall variation. It will be shown later that better performance can be achieved while both states and sub-states are being used, which is the basic characteristic of SMP. The methodology of SMP is explained in next section. The proposed methodology is applied to a station rainfall data at four rain gauge stations – Khandwa, Jabalpur, Sambalpur and Puri, located at various parts in India. Results and discussions are presented afterwards.

METHODOLOGY

General Markov Process

The Markov Process (MP) at discrete time points is characterized by a set of states and the transition probabilities P_{ij} from state i at time step t to state j at time step $t+1$ (Haan *et al*, 1976; Haan, 2002). The matrix representation of all possible P_{ij} forms the transition probability matrix (TPM) of the Markov chain, denoted as P. The definition of the P_{ij} implies that the sums of all elements in any row equal to 1 as the transitions from a particular state to all possible states are 'mutually exhaustive'.

Copyright © 2011 John Wiley & Sons, Ltd.

The property of successive dependence in a time series is modelled through MP. The order of an MP is equal to the number of previous observation(s) on which the present value depends. For example, the conditional probability for m^{th} order MP is expressed as $P[X_t = a_j | X_{t-1} = a_i, X_{t-2} = a_k, \cdots, X_{t-m} = a_l]$. Similarly, a first-order MP is a stochastic process in which the state of the value X_t of the process at time t depends only on the state of X_{t-1} at time $t-1$ and no other previous values. Thus, the transition probability for the first-order MP, P_{ij} , is expressed as

$$P_{ij} = P\left[X_t = a_j \middle| X_{t-1} = a_i\right] \tag{1}$$

The collection of all these probabilities with m different states forms the TPM, which provides information of transition from one state to another state, and thus can be synonymously termed as state-to-state TPM or state/state TPM as against state/sub-state TPM in case of SMP. Detailed methodology of SMP is presented in the following section.

Split Markov Process

Major steps of SMP are shown in a flowchart in Figure 1. It is a data-driven process as in the case of an MP. Basic assumption is the first-order stationarity of the data. However, homogeneity of the data across different stations is not a necessary requirement if SMP is being applied to a specific station. In order to investigate the daily rainfall variation in a probabilistic way, another sub-state is introduced in addition to the existing states. Thus, the states categorize the daily rainfall amount, and the sub-states categorize the daily rainfall variation. The observed rainfall data are classified in different categories depending on its variability, and these categories are denoted as different states, say, S_1, S_2, $\cdots$, S_n , n being the total number of states. The amount of variation in daily rainfall magnitude is obtained by first-order differencing of original data. These variations in daily rainfall magnitude are classified into different categories depending on the range of their variability. These categories are denoted as sub-states, say, $\bar{s}_1$, $\bar{s}_2$, $\cdots$, $\bar{s}_m$, m being the total number of states. The probability of transitions

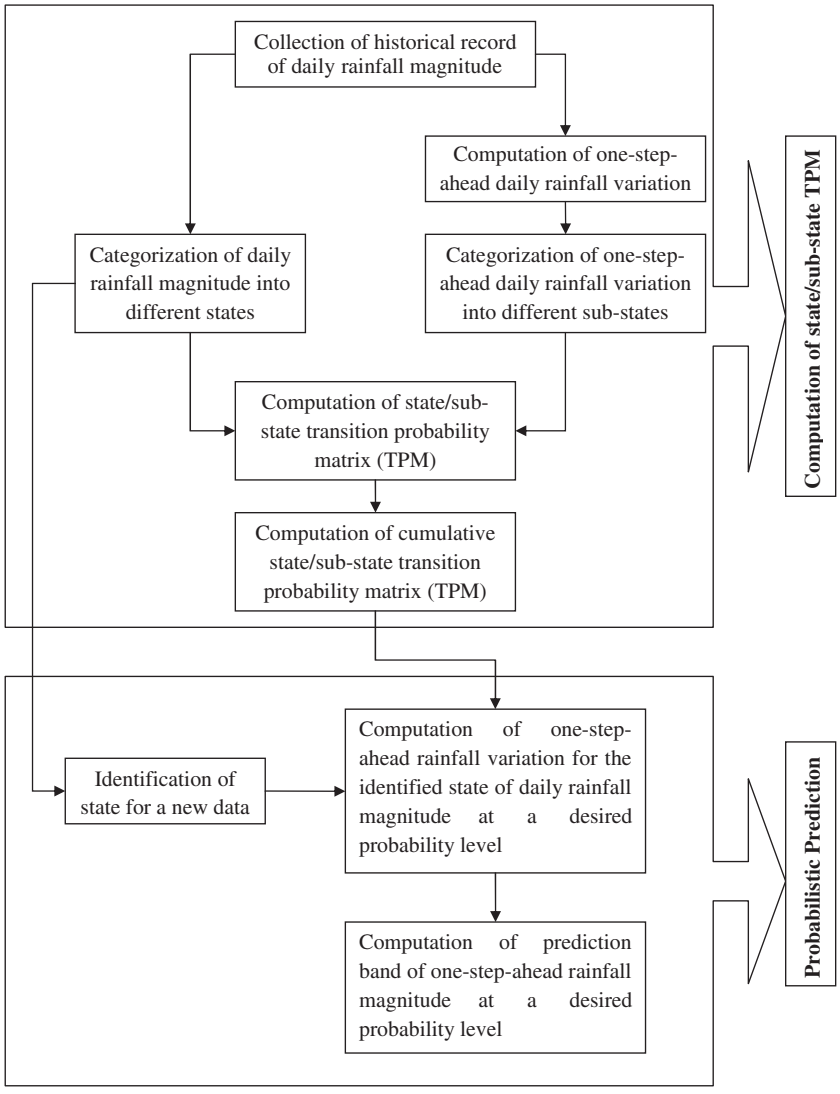

Figure 1. Flowchart showing major steps of Split Markov Process (SMP)

Copyright © 2011 John Wiley & Sons, Ltd.

from a particular state to a particular sub-state is obtained from historical data and denoted as state/sub-state transition probability. The general k^{th} order state/sub-state transition probability is expressed as

$$P^{k}_{S,\bar{s}(j)} = P\left[r_t = \bar{s}_j \middle| R_{t-1} = S_i, R_{t-2} = S_k, \cdots, R_{t-k} = S_l\right] \quad (2)$$

where R denotes the daily rainfall magnitude, and r denotes the change in daily rainfall magnitude. A first-order state/sub-state transition implies that the change in magnitude for the next time step depends on the state of the system at the current time. Thus, a first-order state/sub-state transition probability is expressed as

$$P^{1}_{S(i),\bar{s}(j)} = P\left[r_t = \bar{s}_j \middle| R_{t-1} = S_i\right] \quad (3)$$

The first-order state/sub-state TPM is expressed as (omitting the superscript for clarity)

$$P_{S,\bar{s}} = \begin{bmatrix} P_{S(1),\bar{s}(1)} & P_{S(1),\bar{s}(2)} & \cdots & P_{S(1),\bar{s}(m)} \\ P_{S(2),\bar{s}(1)} & P_{S(2),\bar{s}(2)} & \cdots & P_{S(2),\bar{s}(m)} \\ \vdots & \vdots & \vdots & \vdots \\ P_{S(n),\bar{s}(1)} & P_{S(n),\bar{s}(2)} & \cdots & P_{S(n),\bar{s}(m)} \end{bmatrix} \quad (4)$$

State/sub-state TPM is computed by selecting a particular state and counting the number of transition from that state to a particular sub-state. If a particular state, say $S(j)$, is observed for a total n times, and m is the number of transition from state $S(j)$ to a particular sub-state $\bar{s}(j)$, then the $(i,j)th$ component of the state/sub-state TPM will be

$$P_{S(i),\bar{s}(j)} = \frac{m}{n} \quad (5)$$

The total number of times a particular state is observed and its transition to different sub-states are obtained from sufficiently long record of daily rainfall series.

Once the state/sub-state TPM is obtained, the cumulative state/sub-state TPM is obtained by row wise summation of column-by-column probabilities. A contour plot of this cumulative state/sub-state TPM will represent the nature of possible variation (probabilistically) in the forthcoming step from all possible states at the current time step. Thus, this contour plot can be used for probabilistic prediction of possible range of daily rainfall in the next step. For instance, from a particular state (current step), the possible variation of magnitude of expected change in next day rainfall (at some probability level) is computed using cumulative state/sub-state TPM. For graphical interpretation, one has to start from that particular state to that probability contour (desired probability level), and magnitude of expected change can be computed using a suitable interpolation technique. The minimum and maximum possible changes (with sign) are added to the rainfall magnitude of the current step to obtain the possible range of rainfall in the next time step. If the minimum possible change turned out to be very high negative value, it might be possible to get the lower limit of predicted rainfall range as negative value. However, the lower bound of the predicted range of possible rainfall should be bounded by zero. A numerical example on calculation of the TPM for SMP is illustrated in Appendix A. Further, a typical example on estimation of probabilistic range of daily rainfall using SMP is illustrated in Appendix B.

APPLICATION OF SPLIT MARKOV PROCESS

The methodology is applied to the daily rainfall at four rain gauge stations – Khandwa, Jabalpur, Sambalpur and Puri. Location map of these stations in India is shown in Figure 2. Khandwa rain gauge station is located in the Nimar district in Madhya Pradesh, India. Similar to the major part of Madhya Pradesh, Khandwa is having more or less plain topography. Average altitude of Khandwa is 316 m above mean sea level. Puri is a costal station. It is located on the sea coast of Bay of Bengal and having an almost flat terrain. It is just few meters above the mean sea level. Sambalpur is having an undulating topography with approximate altitude 188 m above the mean sea level. It is about 300 km away from the coastal line. Jabalpur is located on the banks of the perennial Narmada River, and approximate altitude is 393 m above mean sea level. The entire area is low rocky and barren hillocks with slopes differing in grade from 2 to 30%. Jabalpur and Khandwa are far away from the coast and located in the interior part of Indian land.

The daily rainfall data are collected for the period 1901 to 1999 from Indian Meteorological Department, Pune. The data set is complete, and there is no missing data. The data are for the monsoon period (June to September) only as most of the annual rainfall (above 80%) occurs in this period only (4 months). During the remaining eight months, there is almost no rainfall. So there is no change in rainfall depth on a daily scale during this period. Thus, it is excluded from the analysis. Basic statistics for the rainfall data at all these stations are shown in Table I. It is found that the station Sambalpur is having maximum mean rainfall whereas the kurtosis (measure of peakedness) is at maximum for Jabalpur. For Khandwa station, mean rainfall is lowest with the maximum coefficient of variation.

Data for the period 1901 to 1980 are used for the development of state/sub-state TPM, and the data for the period 1981 to 1999 are used to test the performance of SMP. Stationarity of the data set is checked, and the results are shown in Table II. The entire period of the data is divided into five parts, and the mean daily rainfall is computed for each period. Mean is also computed for the entire length of data (1901–1999). The p-value (in parentheses) is obtained for the null hypothesis that the mean is equal to the mean for the entire period (1901–1999) for that station at 5% significance level. It is found that for almost all the cases, the mean does not differ from the overall mean (except on two cases). Thus,

Copyright © 2011 John Wiley & Sons, Ltd.

Figure 2. Location map of rain gauge stations (Source: Google map)

it can be safely assumed that the data are first-order stationary. The methodology of SMP is applied to a specific station, thus the homogeneity of the data is not a necessary requirement. On the other hand, being located over different parts of the country, the daily rainfall characteristics need not be homogeneous. During the model development of SMP, TPM and subsequent cumulative TPM are developed for all the rain gauge stations separately. Thus, the model is able to take care of heterogeneity, if any. Thus, it is found later that the SMP performs almost equally for all these stations.

RESULT AND DISCUSSION

The daily rainfall data (R) are divided into nine different states. The zero rainfall ($R = 0$) is categorized as State 1 and range of other eight states are selected suitably as follows (data in mm):

State1 $R = 0$
State 2 $0 < R \leq 5$
State 3 $5 < R \leq 10$
State 4 $10 < R \leq 20$

Table I. Descriptive statistics of the rainfall data

	Descriptive statistics for daily rainfall data				
Station	**Mean**	**Median**	**CV**	**Skewness**	**Kurtosis**
Khandwa	6.22	0	2.69	5.75	52.58
Jabalpur	9.99	1.00	2.21	6.52	111.45
Sambalpur	11.33	1.60	2.09	4.96	52.87
Puri	7.98	0.10	2.46	4.91	41.42

Table II. Test for stationarity in mean. The p-value (in parentheses) is for the null hypothesis that the mean is equal to the mean for entire period (1901–1999) for that station. The boldface cells indicate that null hypothesis cannot be rejected at 5% significance level

	Mean in mm (p-value)					
Station	**1901–1999**	**1901–1920**	**1921–1940**	**1941–1960**	**1961–1980**	**1981–1999**
Khandwa	6.22	4.97 (0.001)	**6.06 (0.682)**	**7.03 (0.029)**	**6.80 (0.124)**	**6.21 (0.983)**
Jabalpur	9.99	**9.09 (0.060)**	11.30 (0.008)	**10.05 (0.909)**	**9.76 (0.635)**	**9.76 (0.653)**
Sambalpur	11.33	**11.47 (0.787)**	**11.99 (0.209)**	**11.48 (0.777)**	**10.58 (0.147)**	**11.14 (0.724)**
Puri	7.98	**7.62 (0.391)**	**7.88 (0.807)**	**7.88 (0.818)**	**7.79 (0.658)**	**8.78 (0.074)**

Copyright © 2011 John Wiley & Sons, Ltd.

State 5 $20 < R \leq 30$
State 6 $30 < R \leq 45$
State 7 $45 < R \leq 65$
State 8 $65 < R \leq 100$
State 9 $R > 100$

The states are selected in such a way that approximately 70% of the data fall below state 2, 80% of the data are below state 3, 85% of the data are below state 4, 90% of the data are below state 5, 95% of the data fall below state 6, 97.5% of the data fall below state 7 and 99% of the data are below state 8. Thus, it is ensured that the higher the magnitude, the finer the division. However, it is also ensured that a minimum of 50 data should fall in any state.

The changes in magnitude of daily rainfall are computed by taking first order different of the original series. These magnitudes (r) are classified into another set of nine different sub-states. The categorization is as follows (values are in mm):

Sub-state a $r \leq -100$
Sub-state b $-100 < r \leq -50$
Sub-state c $-50 < r \leq -25$
Sub-state d $-25 < r \leq -5$
Sub-state e $-5 < r \leq 5$
Sub-state f $5 < r \leq 25$
Sub-state g $25 < r \leq 50$
Sub-state h $50 < r \leq 100$
Sub-state k $r > 100$

State/sub-state TPM is computed by selecting one particular state and historical transitions from that state to a particular sub-state that are obtained from the available data, as shown in Equation (5) in the methodology. The state/sub-state TPM is shown in Table III. Row wise summation of column-by-column probabilities in the state/sub-state TPM results in cumulative state/sub-state TPM. The cumulative state/sub-state TPM is represented in a contour plot (Figure 3). In this plot, 5, 50 and 95% probability contours are shown in particular.

Three points can be noticed from the contour plot of cumulative state/sub-state TPM. First, the low probability contour line is almost linear whereas the high contour lines are nonlinear. Second, the low probability contours indicate that a lower state can have a larger change in the next time step, particularly for the low probability contours. For example, if the initial state is 2, at 50% probability level, the change magnitude is somewhere in between sub-states d and e, whereas if the initial state is 4, the change magnitude is somewhere in between c and d. However, for high probability contours, change magnitude increases with the relatively higher initial states. This can be observed for states 1 through 4 at 95% probability level. The third point is that, for all the probability lines for high initial states, the probability contours are linearly decreasing. This indicates that an extreme event can be followed by reduction in its magnitude in the next step (at daily scale).

As stated before, the cumulative state/sub-state TPM can be used to probabilistically infer the possible change in rainfall magnitude in the next time step. Being in some particular state at the current time step, computation of the magnitude of expected change in rainfall (at some probability level) in the next time step is carried out using cumulative state/sub-state TPM. Two different values (minimum and maximum possible changes) are computed from the identified state of change by interpolation considering lower and upper boundaries for each sub-state. Results using linear interpolation are presented in this paper. The minimum and maximum possible changes (with sign) are added to the rainfall magnitude of the current step to obtain the possible range of rainfall in the next time step. The prediction performance is investigated for the period 1981 to 1999. The prediction performance varies with the probability level for the next day rainfall. A plot between probability level *versus* mean square error (MSE), root mean square error (RMSE) and mean absolute error is prepared (Figure 4). These measures are computed between the observed and the average of upper and lower limits predicted range. It is found that both the MSE and the RMSE are decreasing or are remaining constant up to 80% probability level. Mean absolute error is found to gradually increase with the increase in probability. However, considering all these measures, the best performance is obtained at 80% probability level in terms of MSE and RMSE. Thus, the predictions are made at

Table III. State/sub-state transition probability matrix using Split Markov Process

	Sub-states								
States	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**k**
1	0.000	0.000	0.000	0.000	0.876	0.088	0.023	0.011	0.002
2	0.000	0.000	0.000	0.001	0.773	0.155	0.048	0.017	0.006
3	0.000	0.000	0.000	0.476	0.291	0.144	0.057	0.020	0.012
4	0.000	0.000	0.000	0.684	0.113	0.112	0.055	0.025	0.011
5	0.000	0.000	0.158	0.648	0.072	0.066	0.033	0.018	0.006
6	0.000	0.000	0.646	0.229	0.044	0.026	0.022	0.026	0.007
7	0.000	0.311	0.500	0.104	0.031	0.018	0.031	0.006	0.000
8	0.000	0.782	0.126	0.058	0.012	0.000	0.000	0.000	0.023
9	0.629	0.258	0.048	0.048	0.000	0.000	0.016	0.000	0.000

Copyright © 2011 John Wiley & Sons, Ltd.

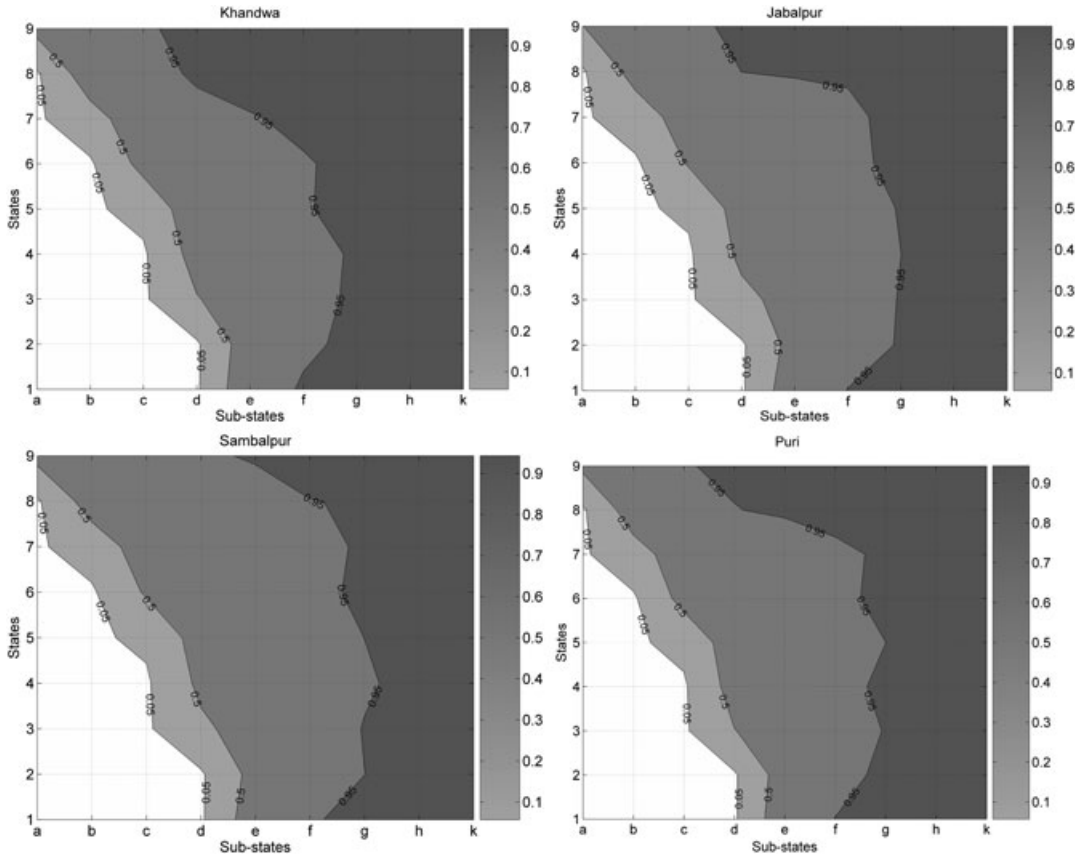

Figure 3. Contour plot of state/sub-state cumulative TPM showing 5%, 50%, and 95% probability contours for different stations as shown in title

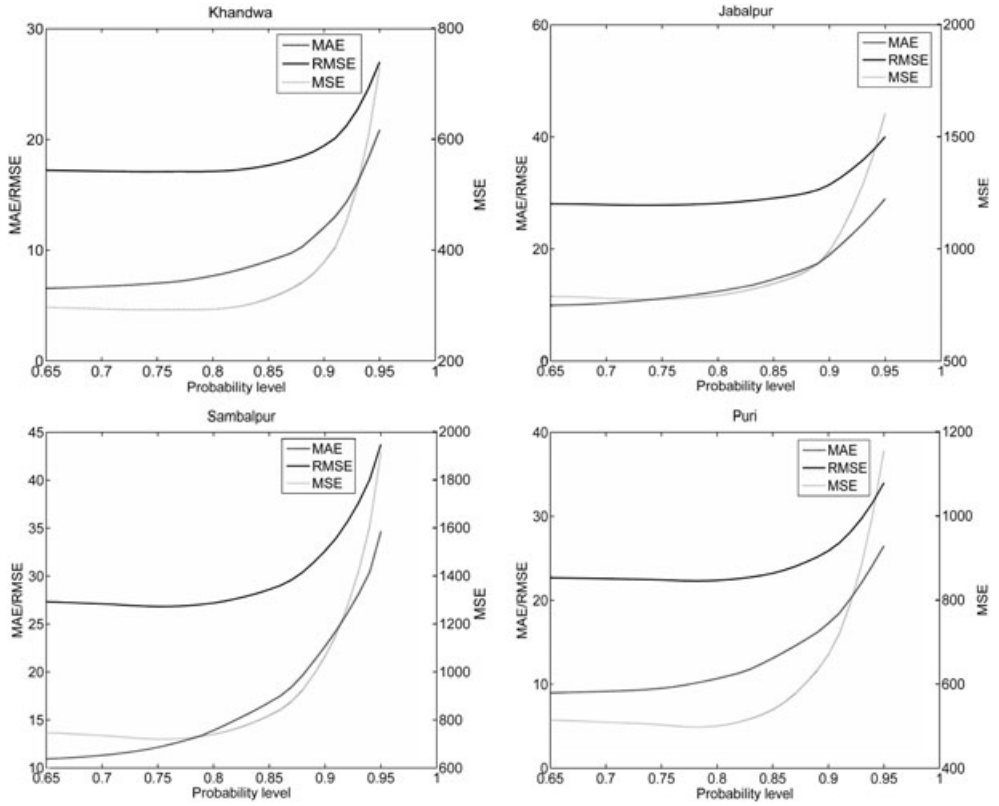

Figure 4. Plot of probability level *versus* mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE)

Copyright © 2011 John Wiley & Sons, Ltd.

80% probability level for the period 1981 to 1999 and shown in Figure 5 for all the rain gauge stations. However, for clarity, the prediction performance for the period 1998 to 1999 is shown in Figure 6 for all the rain gauge stations. Upper and lower limits of possible next day rainfall are shown in these plots along with the actual and predicted rainfall. It is found that most of the observed rainfall either lie within the predicted range or lie close to it. Model performance is assessed in terms of hit rate (*HR*), false alarm rate (*FAR*) and percentage captured (*PC*). *HR* and *FAR* are meant for categorical forecast with dichotomous discrete events. For the purpose of daily rainfall prediction, events are selected as 'rainfall' and 'no rainfall'. *HR* is a measure of probability that a rainfall event is predicted, and it actually occurs. *FAR* is a measure of probability that a rainfall event is predicted, but it does not occur. In this way, four different cases are shown in Table IV. Here *A* denotes 'hits', i.e. number of forecasts when the event is predicted to occur, and it actually occurs; *B* denotes 'false alarms', i.e. number of forecasts when the event is predicted to occur, but it does not occur; *C* denotes the number of 'missed forecasts', i.e. when the event is not predicted to occur, but it actually occurs and finally, *D* denotes the number of all correct no-forecasts, i.e when the event is not predicted to occur, and it does not occur too. Thus, $A+B+C+D=N$ denotes the total number of forecasts during the testing period. Now, *HR* is expressed as

$$HR = \frac{A}{(A+C)} \tag{6}$$

and *FAR* is expressed as

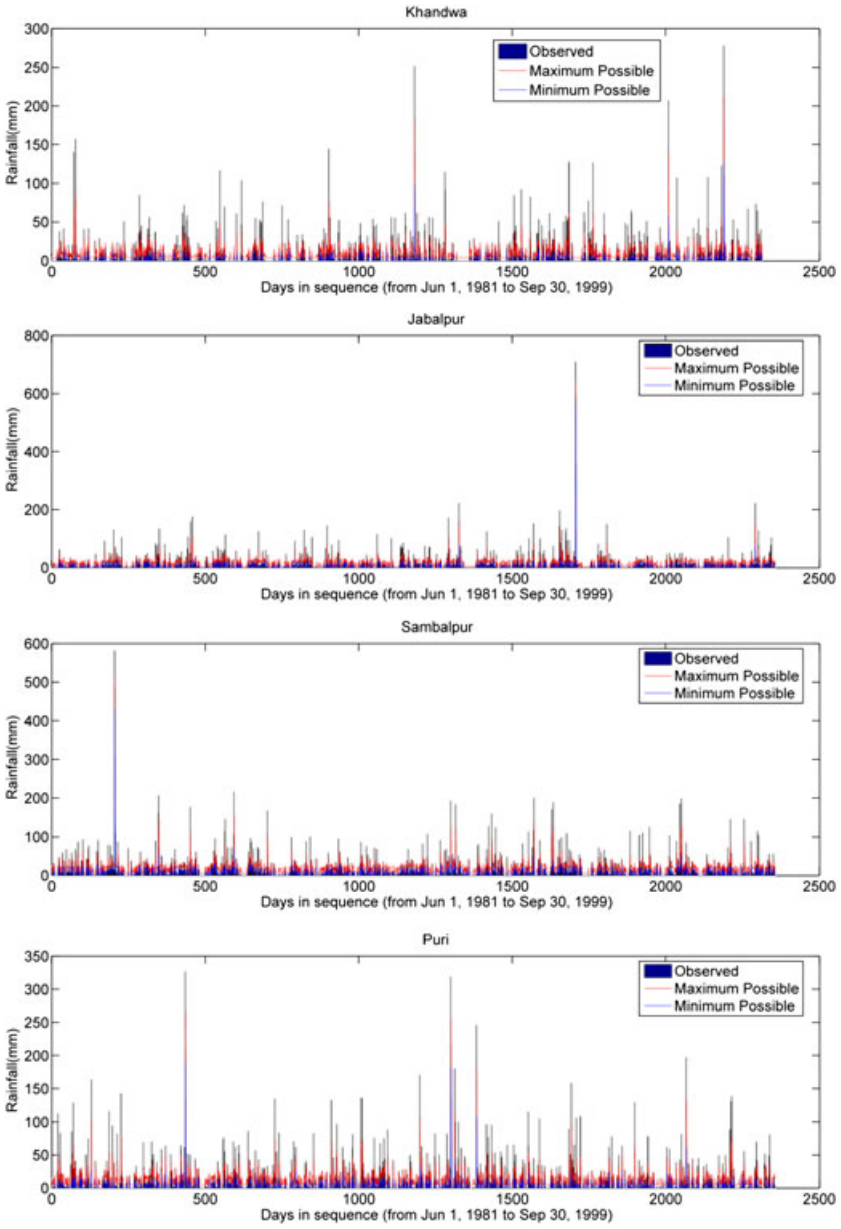

Figure 5. Prediction performance for the period 1 June 1981 to September 1999 for different stations as shown in title

Copyright © 2011 John Wiley & Sons, Ltd.

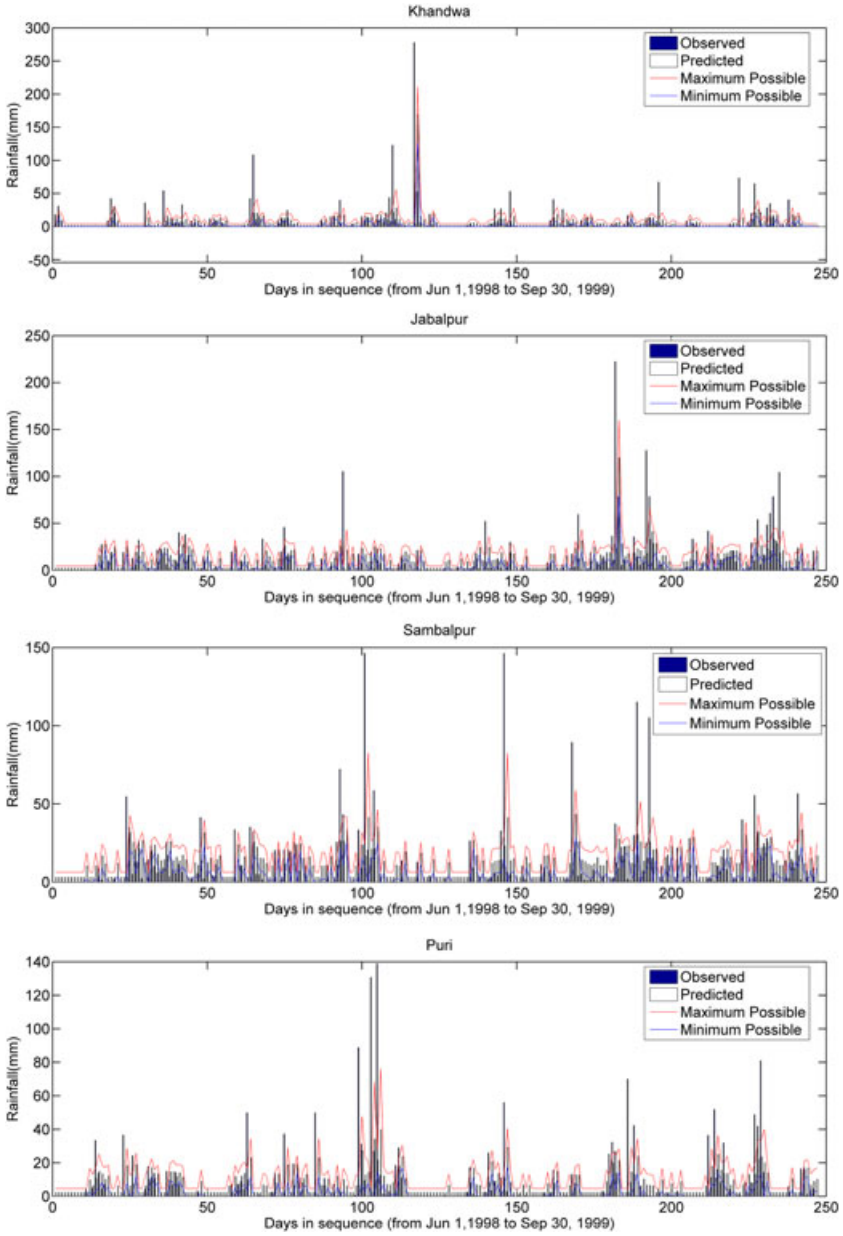

Figure 6. Prediction performance for the period 1 June 1998 to September 1999 for different stations as shown in title

$$FAR = \frac{B}{(A+B)} \tag{7}$$

However, as the basic benefit of SMP is to obtain the prediction range, performance is also assessed through (*PC*) as stated before. *PC* indicates the percentage of actual observations captured within the limits of predicted range. All these measures (*HR*, *FAR* and *PC*) are presented in Table V for all the rainfall stations considered in this study. It is found that *PC* is maximum for Khandwa (70%) and minimum for Sambalpur (44%) whereas *HR* is maximum for Sambalpur and minimum for Khandwa (73%). *FAR* is around or below 30% for Jabalpur, Sambalpur and Puri. *FAR* is maximum for Khandwa (37%). It is further observed that *PC* is towards higher side for Puri and Khandwa (70 and 60%, respectively), which is an indication of a more efficient case. However, for the same stations, *HR* and *FAR* are towards lower and higher side, respectively, which are the indications of less efficient cases as compared with other two rain gauge stations. This is because of the fact that the coefficient of variation is higher for these two stations (Puri and Khandwa) as compared with other stations (Table I). Thus, the rainfall with higher variation is expected to be captured better within predicted range, and rainfall with less variability is expected to have better *HR* (higher) and *FAR* (lower).

As stated before, SMP is an advancement of general MP and specially developed for probabilistic assessment of change in daily rainfall magnitude. The basic difference is in defining another set of sub-states, classifying the changes in magnitude of daily rainfall. On the other hand, MP provides categorical prediction

Copyright © 2011 John Wiley & Sons, Ltd.

Table IV. Four different cases for categorical forecast with dichotomous discrete events – zero rainfall and non-zero rainfall

Prediction status	Actual status	
	Non-zero rainfall	No rainfall
Prediction made for non-zero rainfall	*A*	*B*
Prediction not made (Prediction made for no rainfall)	*C*	*D*

Table V. Model performance in terms of hit rate (*HR*), false alarm rate (*FAR*) and percent captured (*PC*) with and without using sub-states

Station name	Performance measures using both states and sub-states			Performance measures without using sub-states		
	HR	*FAR*	*PC*	*HR*	*FAR*	*PC*
Khandwa	0.63	0.37	0.70	1.00	0.58	0.51
Jabalpur	0.75	0.25	0.48	1.00	0.43	0.40
Sambalpur	0.75	0.25	0.44	1.00	0.40	0.38
Puri	0.69	0.31	0.60	1.00	0.48	0.454

without any information of prediction limits. However, performance of SMP is investigated for possible improvement over a similar approach using MP. In such case (MP), only states are used, i.e. consideration of sub-states is omitted. Thus, nine states as defined earlier are considered. Transition probability matrix and cumulative TPM are then obtained following similar procedure, and finally predictions are obtained. Prediction performance is calculated in terms of *HR*, *FAR* and *PC* as described before, and the results are tabulated in the last column of the Table V. In terms of *FAR* and *PC*, better performance can be concluded for all the stations while both states and sub-states are used as against while sub-states are not in use (only states are used). *HR* apparently looks to be better for the second case, i.e. while sub-states are not in use (only states are used). However, a close insight reveals that when no rainfall is predicted for many cases including all actual no rainfall case. As the events are selected as 'rainfall' and 'no rainfall' cases, prediction of no rainfall case is better in the case when the sub-states are not considered. However, in this case, *FAR* is also more, which indicates a less efficient case. Moreover, from uncertainty quantification point of view, the higher the *PC* is, the better is the performance. Thus, considering all these measures, it can be concluded that better performance can be achieved while both states and sub-states are being used, which are the basic characteristics of SMP.

Stated earlier, there are still few cases when the predicted range fails to capture the observed values. This is particularly following the large rainfall events. This indicates that the historical evidence suggests that high rainfall events generally follow high rainfall events and vice versa. As a result, in such cases, predicted large rainfall events are often one-day later than the observations. Even though this is a shortcoming of the prediction performance, the overall performance is very useful to the community as an early warning to tackle the extreme events, such as flooding, water logging and so on. As the point prediction does not offer any information on uncertainty, it is therefore recommended to provide both predicted rainfall value as well as prediction limits. It is also worthwhile to mention here that one of the most important shortcomings of the SMP is the fact that it needs a long historical record to properly capture the historical behaviour of daily rainfall variation through state/sub-state TPM, which is a general shortcoming for almost all data-driven approaches.

CONCLUSIONS

Daily variation of rainfall is one of the highly complex but most important parameter to tackle various hydrologic problems. SMP is introduced in this paper to assess the daily rainfall variation in a probabilistic way. This study attempts to statistically analyse and predict the probabilistic behaviour of the station rainfall using SMP. SMP investigates the transition between states and sub-states, as against the general MP, which investigates the transition between different states of the system. In order to assess probabilistic range of variation, sub-states are introduced in addition to the states to obtain state/sub-state TPM in SMP. The state/sub-state TPM is generated for daily rainfall data from different rain gauge stations using SMP. The probabilistic behaviour of change in daily rainfall magnitude is captured through state/sub-state cumulative TPM, which is finally used to predict the possible range of daily rainfall in the next time step.

Using SMP, predictions are provided with a possible range of upper and lower limit of rainfall magnitude. Four

Copyright © 2011 John Wiley & Sons, Ltd.

rain gauge stations are selected including one coastal station (Puri) and another station (Sambalpur) within few hundred kilometers from the sea coast. Other two stations (Jabalpur and Khandwa) are located inland. Topography of each station differs from each other. However, the performance of SMP is found to be uniform for all the stations as revealed in the analysis. While investigating the prediction performance in terms of (*HR*) and (*FAR*), it is found that in the cases of Puri and Khandwa, *PC* is towards higher side (70% and 60%, respectively), which is an indication of a more efficient case. However, *HR* and *FAR* are towards lower and higher side, respectively, which are the indications of less efficient cases as compared with the other two rain gauge stations. This is because of the fact that the coefficient of variation is higher for these two stations (Puri and Khandwa) as compared with the other stations (Table I). Thus, the rainfall with higher variation is expected to be captured better within predicted range, and rainfall with less variability is expected to have better *HR* (higher) and *FAR* (lower).

The basic characteristic of SMP that make it different from MP, lies in defining another set of sub-states, classifying the changes in magnitude of daily rainfall. While comparing the performance of SMP for possible improvement over a similar approach without using sub-states (only states are used), it is found that *FAR* and *PC* are better in the case of SMP whereas *HR* apparently looks to be better for the case when sub-states are not in use. However, *PC* is more important from uncertainty quantification point of view, which is associated with the prediction range. Towards this, the results are very useful for the upper range of prediction limits. The early notice for the extreme events is possible to communicate to the concerned community. However, as in the other data-driven methods, the major drawback of the SMP is that it needs a reasonably long historical record to capture the behaviour of daily rainfall variation.

Finally, it is worthwhile to mention here that the illustration of SMP in this paper deals with first-order SMP. The concept can be extended to higher order as well. As explained in Equation 2, in general, previous m states are to be considered for m^{th} order SMP to obtain corresponding TPM. For example, TPM for second-order SMP should consider two previous states. As it is noticed in the analysis, first-order SMP with nine states and nine sub-states constitute a 9×9 TPM, i.e. [*Number of states* $\times$ *Number of sub − states*]. However, for second-order SMP, the size of TPM will be 81×9. Similarly, for third-order SMP, three previous states are to be considered, and the size of TPM will be 729×9. Thus, the number of rows increases by $(\textit{Number of states})^{order}$.

ACKNOWLEDGEMENT

The author wishes to acknowledge the help of three anonymous reviewers through their constructive comments towards the improvement of the manuscript.

APPENDIX A
NUMERICAL EXAMPLE: CALCULATION OF THE TRANSITIONAL PROBABILITY MATRIX FOR SPLIT MARKOV PROCESS

Let us consider that there are 100 data points in a series of observed values. This is pulse data representation (Chow *et al.*, 1988) over discrete time steps (here daily). This means that the daily values are the accumulated depth of rainfall, which has occurred during the entire day. Each observed values can be categorized into different states, thus, there are 100 states. First-order differencing ($r_{t+1} = R_{t+1} - R_t$) is the (next-step) change in rainfall magnitude for the time step t. These changes can also be categorized into different sub-states, and thus, there are 99 sub-states. Finally, paired states and sub-states (one less, i.e. 99) are obtained. Let us further consider that there are five states (I, II, ... V) and five sub-states (a, b, ... e). It may, however, be noted that number of states and sub-states need not be the same). Now, from the record, the numbers of different states are as shown in the second column of Table A1. Again, transition from one particular state to different sub-states is also obtained from the record and shown in the third to seventh column of Table A1.

Now to compute the state/sub-state TPM, each row should be divided by row wise total. In this example, summation of first row is 15 and second row is 45. Thus

Table A1. Number of occurrences of states and its transitions to different sub-states (ref. section 2.3 for the example problem)

State	Number of occurrences (Total = 99)	Number of observed transitions to sub-state				
		a	b	c	d	e
I	15	5	6	3	0	1*
II	45	15	22#	5	3	0
III	18	2	7	6	1	2
IV	12	1	2	5	2	2
V	9	0	1	2	4	2

*This cell should be read as there is 1 occurrence of transition from state I to sub-state e.
#This cell should be read as there are 22 occurrences of transition from state II to sub-state b, and other cells should be read in a similar way.

Table A2. State/sub-state TPM for the example problem shown in Table A1

State	Sub-states				
	a	b	c	d	e
I	0.333	0.400	0.200	0.000	0.067
II	0.333	0.489	0.111	0.067	0.000
III	0.111	0.389	0.333	0.056	0.111
IV	0.083	0.167	0.417	0.167	0.167
V	0.000	0.111	0.222	0.444	0.222

Copyright © 2011 John Wiley & Sons, Ltd.

Table A3. Cumulative state/sub-state TPM for the example problem shown in Table A1 (ref. section 2.4 for the example problem)

	Sub-states				
State	a	b	c	d	e
I	0.333	0.733	0.933	0.933	1.000
II	0.333	0.822	0.933	1.000	1.000
III	0.111	0.500	0.833	0.889	1.000
IV	0.083	0.250	0.667	0.833	1.000
V	0.000	0.111	0.333	0.778	1.000

Table B1. Example of a particular row of the state/sub-state TPM for computation of probabilistic range of predicted rainfall

	Sub-states				
States	a (<−100)	b (−100 to −25)	c (−25 to 25)	d (25 to 100)	e (>100)
State S	0.000	0.291	0.515	0.183	0.011

Table B2. Cumulative state/sub-state TPM following Table B1 for computation of probabilistic range of predicted rainfall

	Sub-states				
States	a (<−100)	b (−100 to −25)	c (−25 to 25)	d (25 to 100)	e (>100)
State S	0.000	0.291	0.806	0.989	1.000

Table B3. Typical example of interpolation to find the limits of changes following Tables B1 and B2

Interpolation for lower limit		Interpolation for upper limit	
a (<−100)	b (−100 to −25)	a (<−100)	b (−100 to −25)
0.291	−100	0.000	−100
0.806	−25	0.291	−25
0.950	**14.34**	0.806	25
0.989	25	0.950	**84.02**
1.000	100	0.989	100

the elements of first and second rows are divided by 15 and 45, respectively. Similarly, summations of third, fourth and fifth rows are 18, 12 and 9, respectively (as shown in Table A1, second column). Hence, elements of these rows are divided by 18, 12 and 9, respectively. Thus, the state/sub-state TPM is as shown in Table A2. Next, the cumulative state/sub-state TPM is obtained as row wise summation of probabilities up to that cell, i.e. cumulative probability of being transited to a particular sub-state or lower than that sub-state. Thus, the cumulative state/sub-state TPM is as shown in Table A3.

APPENDIX B
NUMERICAL EXAMPLE: ESTIMATION OF PROBABILISTIC RANGE OF DAILY RAINFALL USING SPLIT MARKOV PROCESS

Computation of probabilistic range of predicted rainfall is computed from a particular row of the state/sub-state TPM. This row refers to the state at which the previous day rainfall belongs to. Let us consider a row as shown in Table B1, which indicates, without the loss of generality that the rainfall state was 'S' at previous time step. Range of all possible sub-states (a through e for example) is also shown in parentheses.

If we are interested to know the 95% limits of the next day rainfall, we should obtain the lower and upper limits of the predicted change. We should first get the cumulative state/sub-state TPM, which is as shown in Table B2.

The change in magnitude should be in between states c and d. Lower limits of c and d are 25 and 25, respectively, whereas upper limits of sub-states c and d are 25 and 100, respectively. Thus, to find the limits of changes, interpolations are done as shown in Table B3. Interpolated values are shown in boldface.

Thus, the lower and upper limits of the projected change are 14.34 and 84.02 unit. These can be subtracted and added from the actual observed value of the present time step to obtain the limit, subject to the lower bound of the predicted range of possible rainfall should be bounded by zero, as mentioned before. Thus, if today's observed rainfall is 10 unit, then tomorrow's lower and upper limits of the rainfall will be 0 and 94.02 unit.

REFERENCES

Baik HS, Jeong HS, Abraham DM. 2006. Estimating transition probabilities in Markov chain based deterioration models for management of wastewater systems. *Journal of Water Resources Planning and Management* **132**(1): 15–24.

Box G, Jenkins GM, Reinsel G. 1976. *Time Series Analysis: Forecasting and Control*. Prentice-Hall: NJ.

Bohra AK, Basu S, Rajagopal EN, Iyengar GR, Das GM, Ashrit R, Athiyaman B. 2006. Heavy rainfall episode over Mumbai on 26 July 2005: assessment of NWP guidance. *Current Science* **90**(9): 1188–1194.

Chin EH. 1977. Modeling daily precipitation occurrence process in Markov chain. *Water Resources Research* **13**(6): 949–956.

Chow VT, Maidment DR, Mays LW. 1988. *Applied Hydrology*. McGraw-Hill International Edition, ISBN: 978–0071001748, pp. 585.

Deni SM, Jemain AA, Ibrahim K. 2009. Fitting optimum order of Markov chain models for daily rainfall occurrences in Peninsular Malaysia. *Theoretical and Applied Climatology* **97**: 109–121, DOI: 10.1007/s00704-008-0051-3

Fraedrich K, Müller K. 1983. On single station forecasting: sunshine and rainfall Markov chains. *Beitr. Phys. Atmosph.* **56**(1): 108–134.

Fraedrich K, Leslie LM. 1987. Combining predictive schemes in short-term forecasting. *Monthly Weather Review* **115**: 1640–1644.

Gabriel KR, Neumann J. 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society* **88**: 90–95.

Gates F, Tong H. 1976. On Markov chain modeling to some weather data. *Journal of Applied Meteorology* **15**: 1145–1151.

Jimoh OD, Webster P. 1996. Optimum order of Markov chain for daily rainfall in Nigeria. *Journal of Hydrology* **185**: 45–69.

Haan CT. 2002. *Statistical Methods in Hydrology*, 2nd Edition. Iowa State Press: Iowa.

Copyright © 2011 John Wiley & Sons, Ltd.

Haan CT, Allen DM, Street JO. 1976. A Markov chain model for daily rainfall. *Water Resources Research* **12**(3): 443–449.

Hayhoe HN. 2000. Improvements of stochastic weather data generators for diverse climates. *Climate Research* **14**: 75–87.

Kaseke TN, Thompson ME. 1997. Estimation for rainfall-runoff modeled as a partially observed Markov process. *Stochastic Hydrology and Hydraulics* **11**(1): 1–16.

Kottegoda NT, Natale L, Raiteri E. 2004. Some considerations of periodicity and persistence in daily rainfalls. *Journal of Hydrology* **296**: 23–37.

Rajagopalan B, Lall U, Tarbotan DG. 1996. Nonhomogeneous Markov model for daily precipitation. *Journal of Hydrologic Engineering* **1**(1): 33–40.

Sharma A. 2000a. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 – a strategy for system predictor identification. *Journal of Hydrology* **239**(1–4): 232–239. DOI: 10.1016/S0022-1694(00)00346-2

Sharma A. 2000b. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 – a nonparametric probabilistic forecast model. *Journal of Hydrology* **239**(1–4): 249–258. DOI: 10.1016/S0022-1694(00)00348-6

Sharma A, Luk KC, Cordery I, Lall U. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2 – predictor identification of quarterly rainfall using ocean-atmosphere information. *Journal of Hydrology* **239**(1–4): 240–248. DOI: 10.1016/S0022-1694(00)00347-4

Stern RD, Coe R. 1984. A model fitting analysis of daily rainfall data. *Journal of Royal Statistical Society Series A* **147**(Part 1): 1–34.

Weeks WD, Boughton WC. 1987. Tests of ARMA model forms for rainfall-runoff modeling. *Journal of Hydrology* **91**(1–2): 29–47.

Wilks DS. 1999. Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology* **93**(3): 153–169.

Wójcik R, Torfs P, Warmerdam P. 2003. Application of Parzen densities to probabilistic rainfall-runoff modeling. *Journal of Hydrology and Hydromechanics* **51**(3): 175–186.

Copyright © 2011 John Wiley & Sons, Ltd.