Evaluation of Stochastic Daily Rainfall Data Generation Models

J. Jaafar, A. Baki, I.A. Abu Bakar, W. Tahir, H. Awang and F. Ismail

Abstract In developing countries, data is usually a scarce resource as data collection is an expensive exercise. Therefore, analytical method is required to simulate the actual situations and provide synthetic data for forecasting purposes. This paper will compare several methods of synthetically generating rainfall data based on available data. Several models will be used, including lag-one Markov chain model, two-step model, and transition probability model to generate stochastic daily rainfall data of long-term duration, using data from a catchment in Australia. Three variations of lag-one Markov chain models were used: untransformed, logarithmic transformation, and square root transformation. Two-step model uses Markov chain to model rainfall occurrences and gamma distribution to model rainfall depths. Six variations of the Transition Probability Matrices were used, 3 using Shifted Exponential Distribution and 3 using Box–Cox Power Transformation was adopted to predict the high rainfall depths, and the parameters are determined using maximum-likelihood method on the available rainfall data. The models' results were tested by comparing the statistics of the generated data against those of the available data. Direct comparisons of the means, standard deviations, and skews show satisfactory results. Further comparisons of monthly means, standard deviations, skews, maxima and minima, as well as the lengths of wet and dry spells had also shown satisfactory results. In conclusion, all the models have produced synthetic rainfall data, which are statistically similar to those of the available data. In comparison, the TPM model gave the most accurate results. Therefore, this model may be utilised for synthetic rainfall data generations, which can then be used for forecasting.

Keywords Markov chain model · Rainfall modelling · Stochastic modelling · Transition probability matrices and two-step model

J. Jaafar (✉) · I.A. Abu Bakar · W. Tahir · H. Awang · F. Ismail
Faculty of Civil Engineering, Universiti Teknologi MARA, Shah Alam, Malaysia
e-mail: jurina1106@gmail.com

A. Baki
Faculty of Engineering, Al-Madinah International University, Shah Alam, Malaysia

A. Baki
Envirab Services, P.O. Box 7866, Shah Alam 40730, Malaysia

W. Tahir et al. (eds.), *ISFRAM 2015*, DOI 10.1007/978-981-10-0500-8_17

1 Introduction

Long-term data is desirable to enable the asset managers to sufficiently simulate the many possibilities, including flooding and extreme droughts. In developing countries, data is usually a scarce resource as data collection is an expensive exercise. Generation of synthetic data is one of the methods to enable forecasting to be made. One of the techniques available to produce the synthetic data is the stochastic data generation. Rainfall is regarded as the most basic weather variable, independent of temperature and evaporation [16]. Hence, generation of long-term synthetic rainfall data can provide basic set of weather variables for long-term forecasting.

The hydrological time series consists of two contributing factors: random factors and persistence (stochastically deterministic factor) [26]. Stochastic modelling used the stochastic properties of observed time series to generate long-term time series. The statistical and stochastic properties of the observed time series are assumed to represent the population properties, and the synthetic long-term time series is assumed to come from the same population [10].

There are many stochastic data generation models. This paper compared several models including lag-one Markov chain model, two-step model, and the transition probability Matrices model (TPM).

Lag-one Markov chain models are the most popular variations of rainfall data generation models [2, 24, 31]. Higher-order Markov chain models have also been utilised satisfactorily [12]. The major problem in daily rainfall generation using a single-step runoff generation type model (Lag-one Markov Chain Model by [2]) is the large number of zero values of daily rainfall. Richardson [22] used square root transformation and a multivariate normal distribution truncated at zero to overcome the zeros problems. Baki [4] used logarithmic and square root transformation to overcome this problem. Nevertheless, there is an inherent problem of large number of zeros in the historical data, which introduced skews. Nevertheless, Malek and Baki [17] successfully forecasted stochastic data for Gombak River in Malaysia using non-transformed data.

The two-step model was developed by various researchers to separate the analysis between the occurrences of rainfall and the rainfall depth. Jones et al. [16] and Adam [1] modelled occurrences of daily rainfall using a Markov chain. The wet spells, which is a series of rainfall occurrences, and the dry spells, which is a series of non-occurrences of rainfall, have also been satisfactorily modelled using Markov chains [19, 21, 25]. Baki [5] has modelled rainfall data generation using the two-step model using Markov chain for rainfall occurrences and gamma distribution for rainfall depth. Generally, Baki [5] has achieved satisfactory results, where the statistics of the generated data is comparable to those of the recorded data.

Haan et al. [14] and Taewechit et al. [29] used a multistate Markov chain approach to model the distribution of rainfall. Haan et al. [14] used seven states to describe rainfall behaviour based on rainfall depths. The first state is dry (no rain), and six others are wet (with rainfall). Uniform distributions were assumed for states 2–6 and a shifted exponential distribution for the seventh state (unbounded).

A modified TPM model was developed by Srikanthan and McMahon [26] based on the TPM model of Haan et al. [14]. The exception was that the daily rainfall data is transformed using the Box–Cox power transformation [8] instead of a shifted exponential distribution for the last class. Srikanthan and McMahon [27] used TPM model in their development of automatic evaluation of stochastically generated rainfall data. Srikhathan et al. [28] also used TPM model in their comparison of daily rainfall data generation models. Baki [6] found that in general, all six variations used (three sets of matrices using shifted exponential and three sets of matrices using Box–Cox power transformation) were equally satisfactory as the differences between the six variations are minimal. This was consistent with the past research as Haan et al. [14] found that the number of classes did not affect the accuracy of the TPM model to a great extent. Therefore, the selection between the six variations is not very critical.

The objective of this paper is to compare the performance of those models in generating daily rainfall. Apart from comparing the daily statistics of the generated data to those of the recorded data, further comparisons will also be carried out using monthly and annual statistics, daily maxima, and average lengths of dry and wet spells. The comparison will enable identification of the model that will give the most accurate statistical comparisons between recorded and generated rainfall statistics.

2 Data and Methods

2.1 Data

The catchment selected for this study is Kangaroo Valley, which is located about 150 km south of Sydney and about 50 km west of the east coast of New South Wales, Australia. The map is shown in Fig. 1, and catchment characteristics are as listed below [3]:

- The National Index reference is 215,220.
- The catchment area is 330 km^2.
- The length of the stream (Kangaroo River) is 34.5 km.
- The average slope of the Kangaroo River is 1.35 % or 135 in 10,000.
- The annual rainfall for Kangaroo Valley is 1629.0 mm.
- The annual runoff from the catchment is 934.2 mm.
- The annual pan evaporation is 1773.4 mm.
- The climatic condition for this catchment is temperate.
- The vegetation in the area is a mixture of rainforest, hedgeland, sedgeland, and grassland.

A total of 80 years of daily rainfall data were used. Both regionalised and single-site approaches have been satisfactorily used in rainfall data generation.

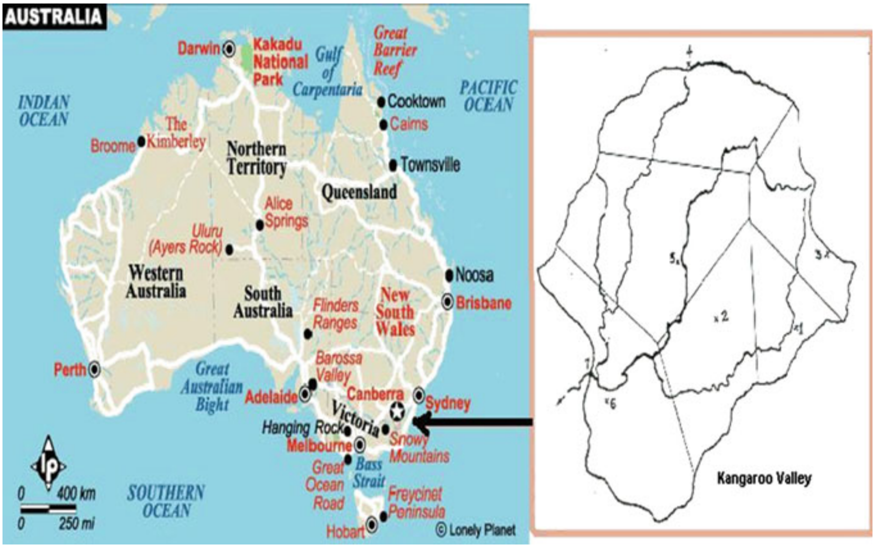


Fig. 1 Catchment (after, [3])

Benson and Matalas [7] used regionalised parameters in stochastic runoff data generation. Solomon [23] used regionalised parameters as he found that regionalised parameters were more suitable than single-site parameters because regionalisation reduced operational bias. Baki [4] found that by using the average rainfall for the catchment, continuity of data could be obtained. Hernáez and Martin-Vide [15], Mehrotra et al. [18], and Camberlin et al. [9] had used regionalised approach to satisfactorily model rainfall data. However, Mhanna and Bauwers [20] had satisfactorily generated rainfall data using single-site approach. In this study, the regionalised approach had been adopted using catchment daily average rainfall. Therefore, the use of catchment average rainfall instead of individual stations allows for better approximations of rainfall stochastic properties and processes.

The location of the catchment is shown in Fig. 1. The locations of the rainfall stations are shown in the enlarged inset of Fig. 1. Catchment average rainfall was computed using the Thiessen polygons [30] of available data for the day. For the day with available data from all rainfall stations, the Thiessen [30] polygons will be computed using 6 rainfall stations (as shown in the inset of Fig. 1). For days that have missing data (e.g. if station 1 data is missing), the Thiessen polygons [30] will be computed using the available data only, namely stations 2, 3, 4, 5, and 6. Similarly, if data from stations 1 and 2 are missing, then the Thiessen polygons [30] will be computed using the available data from stations 3, 4, 5, and 6. There are different polygons for different sets of missing data.

Statistics of daily rainfall for this catchment are shown in Table 1. Table 1 shows that the overall means, standard deviations, skews, and coefficient of variations of daily rainfall for this catchment are 4.4, 15.6, 8.1, and 3.5 mm, respectively.

Table 1 Recorded daily rainfall statistics (after [4])

Month	Mean (mm)	Std. Dev. (mm)	Skew (γ)	Coeff. Var. (C_v)	γ/C_v
Jan	4.8	16.4	11.3	3.4	3.3
Feb	5.5	17.5	7.4	3.2	2.3
Mar	5.8	18.2	6.3	3.1	2.0
Apr	4.9	16.5	7.0	3.4	2.1
May	4.8	17.7	7.9	3.7	2.1
Jun	6.1	19.3	5.5	3.1	1.8
Jul	4.6	18.0	8.3	3.9	2.1
Aug	3.1	11.4	8.2	3.7	2.2
Sep	3.1	9.9	6.5	3.2	2.0
Oct	3.7	15.0	9.2	4.1	2.3
Nov	2.9	8.9	6.8	3.0	2.2
Dec	4.1	12.2	7.3	3.2	2.3
Overall	**4.4**	**15.6**	**8.1**	**3.5**	**2.3**

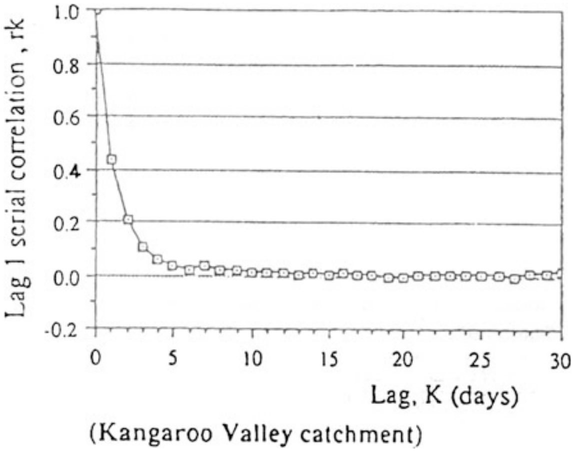


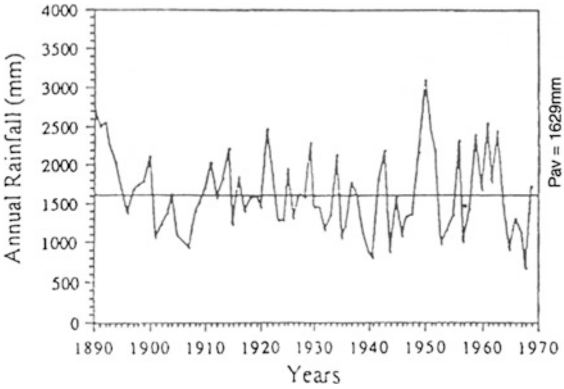
Fig. 2 Plot of serial correlation against lag [4]

The ratio of skew to coefficient of variation is 2.3, which is close to 2, indicating that gamma distribution can be used to approximate the rainfall distribution [5].

Figure 2 shows a plot of serial correlation coefficient (r_k) plotted against the corresponding lag (k). The lag-one value is $r_1 = 0.436$, while the other r_k values are less than half r_1 [4]. Fisher [13] suggested a value of r_k of 0.349 as the conventional minimum value for stochastic analysis of time series. The lag-one serial correlation coefficient ($r1$) was shown to be satisfactory for this catchment, while the other r_k values are much lower than the suggested conventional minimum value. Lag-one correlation was adopted for this paper [4].

Figure 3 shows the plot of annual rainfall values [4]. No apparent trend can be observed in the values of annual rainfall for this catchment. Therefore, the random

Fig. 3 Plot of annual catchment rainfall [4]



variations are assumed to continue in the future. The stochastic rainfall data generation is therefore assumed to be able to reproduce these random variations [4].

2.2 Lag-One Markov Chain Model

Baki [4] used lag-one Markov chain model in modelling daily rainfall. Earlier applications of lag-one Markov chain model were by Adamowski and Smith [2] and Richardson [22].

In the study by Baki [4], the daily recorded rainfall values were standardised as follows:

$$z_i = \frac{(x_i - \overline{x_i})}{\sigma_i} \tag{1}$$

where z_i is the standardised daily rainfall (mm) for day i, with zero mean and unit standard deviation; x_i is the daily rainfall (mm) for day i; σ_i is the standard deviation (mm) for day i; is the average daily rainfall (mm) for day i, where i ranges from 1 to 366 (including leap years).

The generated rainfall data is given by:

$$z_i = r_i z_{i-1} + t_i \sqrt{(1 - r_i^2)} \tag{2}$$

which gives:

$$x_i = \overline{x_i} + \sigma_i \left[r_i z_{i-1} + t_i \sqrt{(1 - r_i^2)} \right] \tag{3}$$

where x_i is the generated rainfall on day i (mm); $\overline{x_i}$ is the mean recorded daily rainfall of day i (mm); σ_i is the standard deviation of recorded daily rainfall on day i (mm); r_i is the lag-one serial correlation for the whole record; z_{i-1} is the standardised rainfall on day $i - 1$; and t_i is the normally distributed random numbers with zero mean and unit variances.

Baki [4] used three variations of the lag-one Markov chain model, untransformed data (referred to as QT), logarithmically transformed data (referred to as LOG), and square root transformation (referred to as SQR). All these three results will be used in the comparison.

2.3 Two-Step Model

The large number of zero values of daily rainfall caused problems to single-step runoff generation type of model to generate daily rainfall data. The two-step model was developed to separate the analysis between the occurrence of rainfall and the rainfall depth. Baki [5] used the two-step model by modelling the occurrences of rainfall using transition probabilities between two classes of events (dry days and wet days). The transition probabilities between the two classes are according to Markov chain probabilities.

The gamma distribution can be used to model rainfall depths during wet days. Table 1 shows that the ratio of daily skew coefficients to coefficient of variation (γ/C_v) of the recorded data is 2.3, which is close to 2. Baki [5] adopted the gamma distribution since the data he used had a ratio (γ/C_v) close to 2. This distribution is also utilised by Jones et al. [16] and Carey and Haan [11].

The gamma distribution is given by:

$$F(x\|k) = \int_{o}^{x} \frac{(\lambda_{ik})\eta_{ik}}{\Gamma(\eta_{ik})} U^{(\eta_{ik}=1)} \exp(-\lambda_{ik}U)\mathrm{d}u \tag{4}$$

where U is the uniformly distributed random number between 0 and 1.

In order to find the parameters, λ and η, maximum likelihood can be used. Carey and Haan [11] used maximum likelihood to find the parameters in their study. For example,

$$\eta^{*} = \frac{0.5000876 + 0.164852y + 0.0544274y^{2}}{y} \tag{5}$$

in which

$$y = \mathrm{In}\left(\sum_{i=1}^{n} \frac{v_i}{n}\right) - \sum_{i=1}^{n} \frac{\ln v_i}{n} \tag{6}$$

v_i = ith observation from a sample of n observations.

Correction for small sample bias can then be made as follows:

$$\eta = \frac{(n-3)\eta^*}{n} \tag{7}$$

The estimate for λ can then be made:

$$\lambda = \frac{\eta}{\sum_{i=1}^{n} \frac{v_i}{n}} \tag{8}$$

Baki [5] used the two-step model, using a first-order Markov chain to model occurrences of rainfall and a gamma distribution to generate rainfall depths during wet days. The parameters of the gamma distribution will be estimated from the recorded wet days. The results from this study will be used in the comparison (referred to as TS).

2.4 Transition Probability Matrices Model

Haan et al. [14] mentioned that persistence and periodicities can be observed in daily weather patterns. The persistence is modelled by a Markov chain. Consider

$$P\left(E_{nj}\middle|E_{n-1j_{n-1,\ldots,}}E_{1j_1}\right) = P\left(E_{nj}\middle|E_{n-1j_{n-1}}\right) \tag{9}$$

where for $x_1, x_2, \ldots$ as the observations of daily rainfall, then $E_{i,j}$ (i = 1, 2,…, c, and j = 0, 1, …, c), where c is the number of classes or states, and if $P(E_{nj}|E_{n-1j})$ does not depend on n, then these transition probabilities (denoted P_{ij}), and the Markov chain is stationary. The transition probability matrices (TPM) is the collection of P_{ij} between classes in (c + 1) × (c + 1) matrices.

Periodicities mean that the weather pattern undergoes a cyclical behaviour within a year. Within a season, the weather pattern can be assumed to be stationary. Therefore, the TPM can be assumed to be stationary within each season:

$$P_{ij}^{(k)}(i.j = 0, 1, \ldots, c) \quad \text{and} \quad (k = 1, \ldots, s) \tag{10}$$

where k denotes the kth season and s is the total number of seasons.

The probability distributions had to be fitted to each class. It was assumed that the same set of distributions would model each season. Therefore, (c + 1) cumulative distribution functions are used:

$$F_m(x)(m = 0, \ldots, c) \tag{11}$$

where $F_m(x) = P$ (rainfall < x | rainfall belongs to class m).

A uniform distribution was assumed for all wet states, except for the last one. For the highest class, a shifted exponential distribution was found by Haan et al. [14] to be the most suitable:

$$F_{\text{last}(x)} = \exp\left({}^{(x - \text{ncl})}/_{\eta}\right) \tag{12}$$

where ncl is the lower boundary of the last class and η is a constant found by maximum likelihood:

$$\eta = \bar{x} - \text{ncl} \tag{13}$$

where $\bar{x}$ is the mean daily rainfall greater than ncl.

Haan et al. [14] adopted the months to be the seasons. Seasons follow an annual cycle, and by using months to represent seasons, the cyclical pattern can be satisfactorily represented. Hence, the TPM can be assumed to be stationary within a month. They also adopted 7 classes of daily rainfall states after testing up to 12 classes. These values were found to be satisfactory for the Kentucky basin. Therefore, twelve sets of (7 × 7) matrices needed to be found from the recorded data.

Baki [6] tested six variations of the TPM model: 6 × 6 TPM (called SE6), 7 × 7 TPM (called SE7), and 8 × 8 TPM (called SE8), all three with shifted exponential distribution for the last class and linear distribution for the other classes, and 6 × 6 TPM (called BC6), 7 × 7 TPM (called BC7) and 8 × 8 TPM (called BC8), all three with Box–Cox power transformation for the last class and linear distribution for the other classes. The last (highest) class has closed lower bound and open upper bound. The class boundaries are shown in Table 2. The results from Baki [6]'s study will be used in the comparison.

Table 2 Class boundaries for TPM model

Class	Lower limit (mm)	Upper limit (mm)		
		6 × 6	7 × 7	8 × 8
1	0.0	0.0	0.0	0.0
2	0.1	0.9	0.9	0.9
3	1.0	2.9	2.9	2.9
4	3.0	6.9	6.9	6.9
5	7.0	14.9	14.9	14.9
6	15.0	∞	30.9	30.9
7	31.0 (for 7 × 7 and 8 × 8)	N/A	∞	62.9
8	63.0 (for 8 × 8)	N/A	N/A	∞

3 Results and Discussion

Ten replicates of generated data were made, each with the same length as the recorded data. The average of the statistical measures of the generated data was compared to those of the recorded data.

Table 3 shows the average of ten replicates of daily means of all models used compared to the recorded data. In general, the statistical measures of daily means of the generated data for all models are satisfactory except for untransformed lag-one Markov chain model (QT). In comparison, all the six variations of the TPM and TS models are more accurate compared to those of the lag-one Markov chain model, in respect to the daily means of the recorded data. In terms of accuracy of the daily means, 7 × 7 TPM (SE7 with 8 accurate daily means followed by BC7 with 7) gave the best results compared to others (as highlighted in Table 3).

Table 4 shows the average of ten replicates of daily standard deviations compared to the recorded data. Again, the statistics of the generated data for all six variations of the TPM and TS models are more accurate compared to those of the lag-one Markov chain model. In terms of accuracy, the standard deviations for 6 × 6 and 7 × 7 TPM (SE6, SE7, BC6, BC7) tend to be lower, indicating that the data generated by the model tend to be more normally distributed, while 8 × 8 TPM (SE8 and BC8) can generate data that are less normally distributed compared to the recorded data as some of the standard deviations exceeded those of the recorded data. Furthermore, SE8 and BC8 both have 4 accurate daily standard deviations, which are much better than others (SE7 and BC6 both have 2, as highlighted in Table 4).

Table 5 shows the average of ten replicates of daily skews compared to the recorded data. Once again, the statistics of the generated data for all six variations of the TPM and TS models are satisfactory compared to those of the lag-one Markov

Table 3 Mean daily rainfall statistics' comparison

Month	Daily means for recorded and average for the generated data (mm)										
	Rec	Lag-one Markov			2-step	TPM model					
		QT	LOG	SQR	TS	SE6	SE7	SE8	BC6	BC7	BC8
Jan	**4.8**	9.2	5.1	5.4	5.0	4.7	**4.8**	5.0	4.9	**4.8**	4.9
Feb	**5.5**	10.3	6.9	6.2	5.6	5.6	5.4	5.4	5.6	**5.5**	5.7
Mar	**5.8**	11.5	6.7	6.5	6.1	5.9	6.1	6.1	5.6	**5.8**	5.9
Apr	**4.9**	9.7	3.8	5.4	5.2	4.8	**4.9**	**4.9**	4.7	5.0	**4.9**
May	**4.8**	10.3	2.8	5.3	**4.8**	5.1	**4.8**	5.0	4.7	**4.8**	5.0
Jun	**6.1**	11.9	4.9	7.0	6.4	**6.1**	6.2	**6.1**	6.2	**6.1**	6.7
Jul	**4.6**	10.0	2.7	5.2	4.9	4.5	**4.6**	**4.6**	4.7	5.0	4.7
Aug	**3.1**	6.6	2.0	3.5	3.3	3.2	**3.1**	3.2	3.3	3.2	**3.1**
Sep	**3.1**	6.1	2.2	3.5	3.4	3.2	**3.1**	3.2	3.2	**3.1**	3.3
Oct	**3.7**	8.1	2.5	4.1	3.6	3.8	**3.7**	3.9	3.6	3.6	**3.7**
Nov	**2.9**	5.5	2.5	3.3	3.3	**3.0**	3.1	**3.0**	**3.0**	**3.0**	**3.0**
Dec	**4.1**	7.9	3.9	4.6	4.4	4.2	**4.1**	4.4	**4.1**	**4.1**	4.4

Table 4 Daily standard deviations' comparison

Month	Daily standard deviations for recorded and average for the generated data (mm)										
	Rec	Lag-one Markov			2-step	TPM model					
		QT	LOG	SQR	TS	SE6	SE7	SE8	BC6	BC7	BC8
Jan	**16.4**	12.5	21.5	7.2	12.3	13.2	14.3	**15.8**	13.5	14.1	15.4
Feb	**17.5**	13.5	28.7	8.5	12.6	15.3	16.4	16.8	16.0	16.7	**17.4**
Mar	**18.2**	13.6	28.9	8.4	14.4	16.2	**17.9**	19.0	16.3	17.3	17.7
Apr	**16.5**	12.1	16.9	7.0	12.6	14.6	16.2	**16.3**	14.9	16.1	15.6
May	**17.7**	12.9	13.0	6.9	12.5	17.0	16.6	18.4	15.7	16.5	**17.3**
Jun	**19.3**	14.5	22.3	9.2	16.7	18.3	18.9	**19.5**	18.3	18.7	20.0
Jul	**18.0**	13.5	15.3	7.5	13.5	15.4	16.6	17.4	16.0	17.6	**18.2**
Aug	**11.4**	8.7	11.0	5.0	8.8	10.7	10.9	**11.5**	11.1	10.9	10.8
Sep	**9.9**	7.4	9.8	4.6	9.5	9.6	**9.8**	10.3	**9.8**	9.7	10.2
Oct	**15.0**	11.2	11.8	5.7	9.2	12.5	14.3	16.7	13.0	14.4	**14.7**
Nov	**8.9**	7.0	11.3	4.5	**9.0**	8.4	9.5	9.2	8.4	**8.8**	8.7
Dec	**12.2**	10.0	17.4	6.1	11.0	12.4	12.6	14.1	**12.1**	12.7	14.4

Table 5 Daily skews' comparison

Month	Daily skews for recorded and average for the generated data (mm)										
	Rec	Lag-one Markov			2-step	TPM model					
		QT	LOG	SQR	TS	SE6	SE7	SE8	BC6	BC7	BC8
Jan	**11.3**	2.7	**9.3**	2.6	5.5	4.9	6.5	8.1	4.7	5.8	7.4
Feb	**7.4**	2.0	8.3	2.6	4.4	5.1	6.2	**7.3**	4.8	5.8	6.4
Mar	**6.3**	1.4	9.0	2.2	4.9	4.8	5.7	6.7	4.7	5.6	**6.0**
Apr	**7.0**	1.6	10.5	2.2	4.3	5.7	6.4	**7.5**	5.2	5.7	6.4
May	**7.9**	1.7	12.1	2.3	4.3	6.3	6.4	**8.2**	5.3	6.0	7.0
Jun	**5.5**	1.5	9.8	2.3	5.4	4.9	**5.4**	6.1	4.4	4.9	5.0
Jul	**8.3**	2.0	14.0	3.0	4.9	6.2	6.8	8.1	5.3	6.3	**8.2**
Aug	**8.2**	1.9	14.8	2.7	4.8	6.1	7.0	**7.7**	5.7	6.4	7.1
Sep	**6.5**	1.5	11.3	2.2	**6.5**	5.5	6.3	**6.5**	5.2	5.9	6.1
Oct	**9.2**	2.1	12.4	2.6	4.2	6.1	8.8	11.5	7.2	**9.3**	9.0
Nov	**6.8**	2.1	12.1	2.8	**6.7**	5.2	6.5	6.6	4.8	5.9	6.6
Dec	**7.3**	1.8	10.4	2.3	5.5	5.5	6.4	**6.9**	5.0	5.9	**6.9**

chain model, in comparison with the recorded data. In terms of accuracy, the skews for 6 × 6 and 7 × 7 TPM (SE6, SE7, BC6, BC7) tend to be lower, indicating that the data generated by the model tend to be more normally distributed. The 8 × 8 TPM (SE8 and BC8) can generate data that are less normally distributed compared to the recorded data, since some of the skews exceeded those of the recorded data. SE8 has 6 accurate daily skews, followed by BC8 with 4 (as highlighted in Table 5).

By comparing the daily statistics (means, standard deviations, and skews), the TPM models gave the most accurate results compared to the two-step (TS) model and both TPM and TS are more accurate than the lag-one Markov chain models (QT, LOG, and SQR), especially the untransformed (QT). Within the TPM models, the 7 × 7 TPM (both SE7 and BC7) gave the best estimates of daily means, but the 8 × 8 TPM (SE8 and BC8) gave best estimates of daily standard deviations and daily skews. However, the differences between the variations (SE6, SE7, SE8, BC6, BC7, and BC8) are not significant. In general, all six variations were equally satisfactory as the differences between the six variations are minimal. Thus, the findings of Baki [6] were consistent with the past research as Haan et al. [14] found that the number of classes did not affect the accuracy of the TPM model to a great extent. Therefore, the selection between the six variations is not very critical.

In all daily statistical measures, i.e. means, standard deviations, and skews, Tables III, IV, and V show that the trend of the figures given by the TPM model (SE6, SE7, SE8, BC6, BC7, and BC8) follows the trend of the recorded data better than the other models (TS and lag-one Markov). In overall considerations, the TPM is proven to be the most satisfactory model. This finding is consistent with other researches, such as by Srikanthan et al. [28].

Apart from comparing the daily statistical measures (as carried out by [4–6]), other measures were also necessary to be compared. As discussed above, selection between the TPM variations is not critical, and thus, SE8 and BC8 are adopted for further comparison. Since TS has no variations, it is also adopted for further comparison. For the three variations of the lag-one Markov chain model, Baki [4] found that LOG was the most satisfactory variation, and thus, it is adopted for further comparison. Hence, further comparisons were made between SE8, BC8, TS, and LOG.

Figure 4 shows the comparison of daily maxima between recorded data and 4 adopted models. For daily maxima, SE8 was found to be most satisfactorily as it is

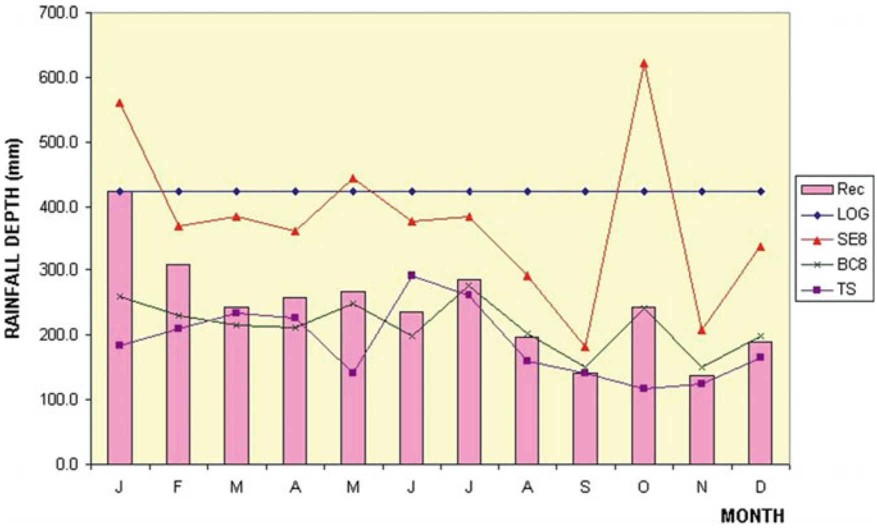


Fig. 4 Comparison of daily maxima

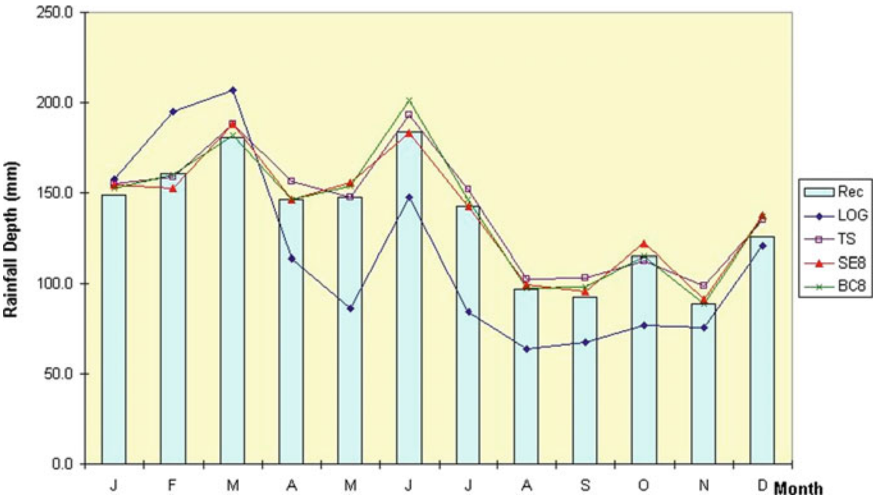


Fig. 5 Monthly means

capable of generated daily maxima greater than the recorded maximum daily rainfall of 423.5 mm. BC8 and TS were also satisfactory in generating the trend, but they tend to have values slightly lower than the recorded maximum. Nevertheless, SE8, BC8, and TS are satisfactory in generating similar trend of daily maxima to the recorded data, hence satisfactory in generating extreme rainfall events. LOG seems to be overestimating the occurrences of daily maxima, with the model generating daily maxima with higher magnitude at higher frequencies compared to other models and also compared to the recorded data.

Figures 5, 6, 7, and 8 show the comparison of monthly statistics. The daily data (recorded and generated) were accumulated on monthly basis, and statistical comparisons were made between the cumulative monthly figures. Figure 5 shows

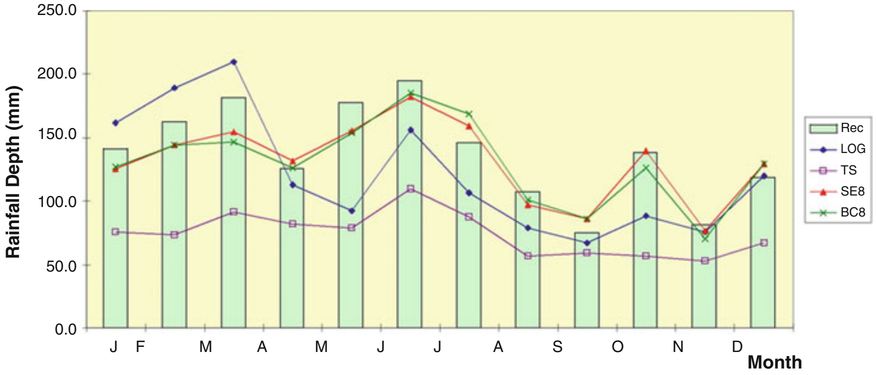


Fig. 6 Monthly standard deviations

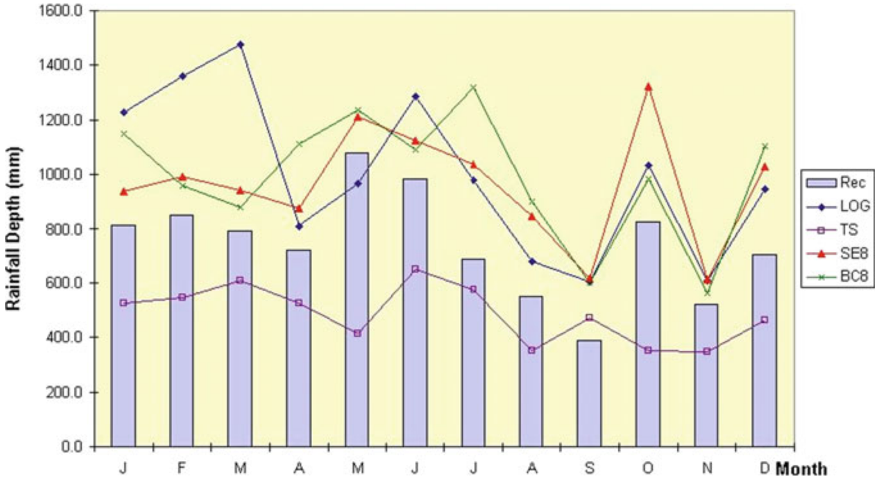


Fig. 7 Monthly maxima

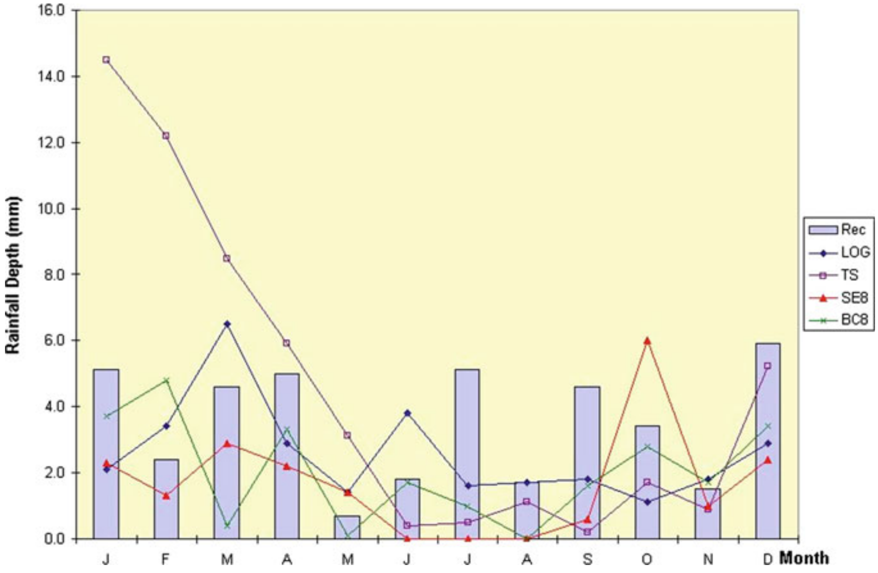


Fig. 8 Monthly minima

that for monthly means, SE8, BC8, and TS were most satisfactory in generating monthly means. Figure 6 shows that SE8 and BC8 were most satisfactory in generating monthly standard deviations, followed by LOG, as TS tends to underestimate the monthly standard deviations. Figure 7 shows that for monthly maxima, SE8, BC8, and LOG were most satisfactory, while TS tends to generate lower

Table 6 Annual statistical comparison

Measures	Recorded	LOG	TS	SE8	BC8
Mean (mm)	1629.2	1394.8	1703.4	1669.4	1678.1
Std. Dev.(mm)	515.0	457.2	263.6	474.0	499.6
Skew	0.5	0.7	0.3	0.5	0.6
Maximum (mm)	3103.3	3222.5	2617.8	3643.0	3736.5
Minimum (mm)	684.9	358.5	1029.2	574.2	632.8

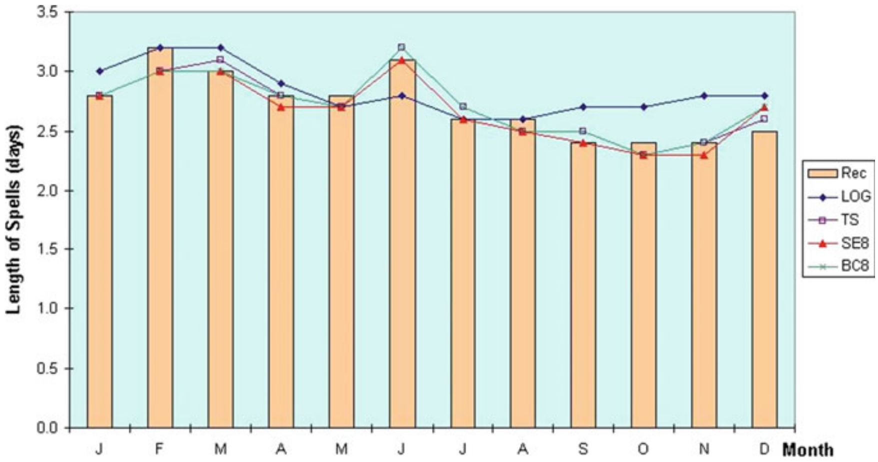


Fig. 9 Average length of wet spells

maxima. Figure 8 shows that for monthly minima, SE8, BC8, and LOG were most satisfactory, while TS tends to generate higher minima during the first quarter. Thus, TS is not able to generate extreme rainfall or drought events.

Table 6 shows the comparison of annual statistics. For annual statistics, SE8 and BC8 are satisfactory in generating annual means, standard deviations, skews, and maxima and minima. TS is only satisfactory in generating annual means, but tends to underestimate the standard deviations, skews, and maxima and overestimate the minima. Thus, TS is unable to reproduce the variations in the recorded data. LOG generated data with lower annual means (14.4 % lower than the annual recorded rainfall), satisfactory standard deviations, skews, and maxima and minima. LOG had the tendency to underestimate the annual rainfall figures.

Figure 9 shows the comparison of average length of wet spells. In terms of sequences of rainfall events, all models were generally satisfactory in reproducing the average lengths of wet spells. Figure 10 shows the comparison of average length of dry spells. LOG tends to underestimate the average lengths of dry spells, while the other three models (SE8, BC8, and TS) are satisfactory. Thus, LOG is unable to model drought events satisfactorily.

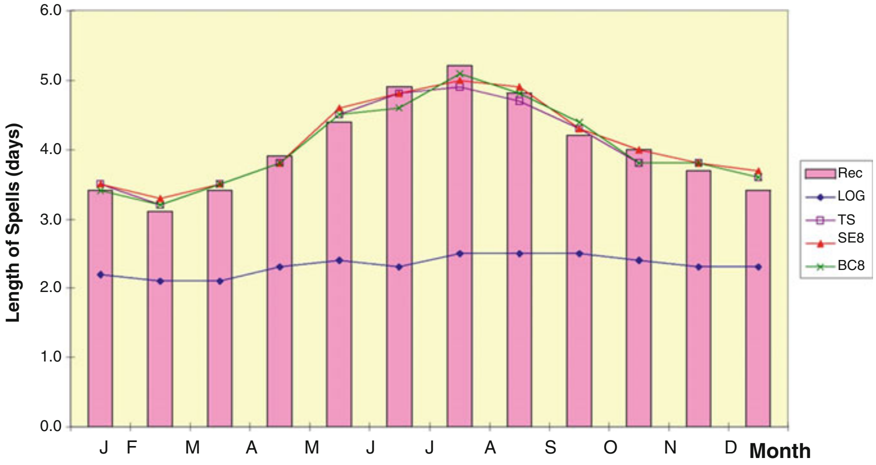


Fig. 10 Average length of dry spells

After further comparisons were made, findings are consistent with the daily statistical comparison (Tables 3, 4, and 5). It is also indicated that the TPM is the most satisfactory model. This finding is consistent with earlier discussions on Tables 3, 4, and 5 and also with other researches, such as by Srikanthan et al. [28]. Thus, TPM can be used to generate stochastic daily rainfall data, which will give synthetic data that is statistically similar to the recorded data.

4 Conclusions

In conclusion, except for QT, all the other models have produced synthetic rainfall data, which are statistically similar to those of the available data. The data generated have similar stochastic properties compared to the recorded data, and statistically, it can be deduced that both samples (recorded and generated sets) come from the same statistical population.

In comparison, the most accurate model is the TPM model for this particular case. It is able to generate data with the closest statistical measures to those of the recorded data. As the data in this case is shown to be persistent over the whole 80-year period, the model can be assumed to be able to forecast the variation in rainfall data. Therefore, this model may be utilised for synthetic rainfall data generations. These synthetic data can then assist in giving possible variations of rainfall over longer period, which would be useful for forecasting.

References

1. Adam RY (2012) Stochastic model for rainfall occurrence using markov chain model. PhD Thesis, Sudan University of Science and Technology, Khartoum, Sudan, unpublished
2. Adamowski K, Smith AK (1972) Stochastic generation of rainfall. J Hydraul Div Am Soc Civil Eng 98(HY11):1935–1945
3. Baki ABM (1996) Objective functions in the optimisation of daily rainfall-runoff modelling. JURUTERA: Mon Bull Inst Eng Malays 9:11–15 (September)
4. Baki ABM (1997) Stochastic rainfall data generation using lag-one markov chain model. J Inst Eng Malays 58(3):55–61
5. Baki ABM (2002) Stochastic rainfall data generation using two-step markov chain model: a case study. In: Proceedings of the 20th conference of ASEAN federation of engineering organisations (CAFEO20), Phnom Penh, Cambodia, vol 1, pp 85–92, 2–4 Sept 2002
6. Baki ABM (2005) Stochastic rainfall data generation using transition probability model. In: Proceedings of the seventh annual IEM water resources colloquium 2005, The Institution of Engineers Malaysia, Petaling Jaya, Malaysia, pp 9-1–9-9, 18 June 2005
7. Benson MA, Matalas NC (1967) Synthetic hydrology based on regional statistical parameters. Water Resour Res 3(4):931–935
8. Box GEP, Cox OR (1964) The analysis of transformations. J Roy Stat Soc B 26(2):211–252
9. Camberlin P, Gitau W, Oettli P, Ogallo L, Bois B (2014) Spatial interpolation of daily rainfall stochastic generation parameters over East Africa. Clim Res 59(1):39–60
10. Campo MA, Lopez JJ, Rebole JP (2012) Rainfall stochastic models, EGU general assembly 2012, held 22–27 Apr 2012 in Vienna, Austria, p 13458
11. Carey DI, Haan CT (1978) Markov processes for simulating daily point rainfall. J Irrig Drai Div Am Soc Civil Eng 104(IR1):111–125
12. Dartidar AG, Gosh D, Dasgupta S, De UK (2010) Higher order markov chain model for monsoon rain over West Bengal, India. Ind J Radio Space Phys 39:39–44 (Febraury)
13. Fisher RA (1958) Statistical methods for research workers, 13th edn. Oliver & Boyd, London
14. Haan CT, Allen DM, Street JO (1976) A Markov chain model of daily rainfall. Water Resour Res 12(3):443–449
15. Hernáez PF-A, Martin-Vide J (2011) Regionalization of the probability of wet spells and rainfall persistence in the Basque Country (Northern Spain). Int J Climatol 32(Issue 12): 1909–1920 (October 2012)
16. Jones JW, Colwick RD, Threadgill ED (1972) A simulated environmental model of temperature. Evaporation Rainfall Soil Moisture Trans Am Soc Agric Eng 15(2):366–372
17. Malek MA, Baki AM (2014) Forecasting of hydrological time series data with lag-one markov chain model. ASEAN J Sci Technol Dev 31(1): 31–37. ISSN: 0217-5460
18. Mehrotra R, Westra SP, Sharma A, Srikanthan R (2012) Continuous rainfall simulation: 2 a regionalized daily rainfall generation approach. Water Resour Res 48:W01536
19. Meshram S, Bisen Y, Kant S, Singh G, Nema AK (2013) Markov chain model probability of dry wet weeks and statistical analysis of weekly rainfall for agricultural planning at Jabalpur. J Environ Ecol 31(3):1250–1254
20. Mhanna M, Bauwens W (2011) Stochastic single-site generation of daily and monthly rainfall in the Middle East. Meteorol Appl 19(1):111–117 (March 2012)
21. Nema AK, Bisen Y, Singh SR, Singh T (2013) Markov chain approach—dry and wet spell rainfall probabilities in planning rainfed rice based production system. Ind J Dryland Agric Res Dev 28(2):16–20
22. Richardson C (1978) Generation of daily precipitation over an area. Water Resour Bull 1 (5):1035–1047
23. Solomon S (1976) Parameter regionalisation and network design. In: Shen HW (ed) Stochastic approaches to water resources. Colorado State University Press, Fort Collins, pp 12.1–12.37
24. Sonnadara DUJ (2012) Modeling daily rainfall using markov chains, annual research sympsium 2012. University of Colombo, Sri Lanka

25. Sonnadara DUJ, Jayawardene DR (2014) A Markov chain probability model to describe wet and dry patterns of weather at Colombo. Theor. Appl Climatol, February 2014. doi:10.1007/s00704-014-1117-z (February)
26. Srikanthan R, McMahon TA (1983) Stochastic simulation of daily rainfall for australian stations. Trans Am Soc Agric Eng 26:754–759
27. Srikanthan R, McMahon TA (2005) Automatic evaluation of stochastically generated rainfall data. Aust J Water Resour 8(2):195–201
28. Srikanthan R, Siriwardena L, McMahon TA (2005) Comparison of two daily rainfall data generation models. Aust J Water Resour 8(2):203–212
29. Taewechit S, Soni P, Salokhe VM, Jayasuriya HPW (2011) Optimal stochastic multi-states first-order Markov chain parameters for synthesizing daily rainfall data using multi-objective differential evolution in Thailand. Meteorol Appl 20(1): 20–31, March 2013 (March 2013)
30. Thiessen AH (1911) Precipitation averages for large areas. Mon Weather Rev 1082 pp
31. Yusuf AU, Adamu L, Abdullahi M (2014) Markov chain model and its application to annual rainfall distribution for crop production. Am J Theor Appl Stat 3(2):39–43