Probability Structure and Return Period of Multiday Monsoon Rainfall

Nur S. Muhammad1; Pierre Y. Julien, M.ASCE2; and Jose D. Salas, M.ASCE3

Abstract: The daily monsoon rainfall data recorded at Subang Airport, Malaysia, from 1960 to 2011 is examined in terms of probability structure for the estimation of extreme daily rainfall precipitation during the Northeast (NE) and Southwest (SW) Malaysian monsoons. The discrete autoregressive and moving average [DARMA(1,1)] model is preferable to the first-order Markov chain [DAR(1)] model. The conditional probabilities of $t$ consecutive rainy days are time dependent. Nevertheless, a simple two-parameter gamma distribution appropriately fits the frequency distribution of multiday rainfall amounts. An algorithm is developed by combining the DARMA(1,1) and gamma models to estimate the return period of multiday rainfall. Extensive comparisons showed that the DARMA(1,1)-gamma model gives a reliable estimate of the return period of rainfall for both NE and SW monsoons at Subang Airport. Furthermore, values generated from the models enable the analysis of the frequency distribution of extreme rainfall events. DOI: 10.1061/(ASCE)HE.1943-5584.0001253. © 2015 American Society of Civil Engineers.

Author keywords: Multiday rainfall; Monsoon rainfall precipitation; Return period; Conditional probability; Stochastic modeling.

Introduction

The planning and design of water resources projects require the analysis of reliable, long-term hydrological data such as rainfall and streamflow. The stochastic point process was first introduced by Todorovic (1968) and subsequently used by Todorovic and Yevjevich (1969) and Eagleson (1978) for modeling short-term rainfall. Kavvas and Delleur (1981) were successful in modeling the sequence of daily rainfall in Indiana using the Neyman-Scott (NS) cluster process. They assumed that a rainfall event occurs in midday, to comply with the model order. Multiday rainfall events were treated as a group instantaneous rainfall that occurs once a day, with a 1-day interval. Rodriguez-Iturbe et al. (1987) tested the performance of different types of point process models, i.e., the Poisson and cluster-based models using hourly rainfall data from Denver, Colorado. They concluded that the white-noise Poisson model was unable to produce satisfactory results. Instead, cluster-based models, namely the Neyman-Scott (NS) and Bartlett-Lewis (BL) processes, are more flexible, reliable, and able to represent the actual rainfall scenarios. Since then, other researchers have improved the NS and BL processes to model fine-scale rainfall. Examples of the use of a modified BL model can be found in Rodriguez-Iturbe et al. (1988), Glasbey et al. (1995), Khaliq and Cunnane (1996), Cowpertwait et al. (2007), and Verhoest et al. (2010). In addition, Cowpertwait (1995), Cowpertwait et al. (1996), and Burton et al. (2008, 2010) are studies that utilize the modified NS to model rainfall. However, Rodriguez-Iturbe et al. (1988), Cowpertwait et al. (1996, 2007), and Burton et al. (2008) show that the point-process models were unable to produce extreme values with good accuracy. Furthermore, Obeysekera et al. (1987) applied various types of point-process models for hourly rainfall considering the diurnal cycle that is characteristic at certain locations during some months of the year.

Low-order discrete autoregressive family models, such as the discrete autoregressive [DAR(1)] and discrete autoregressive and moving average [DARMA(1,1)] models, are frequently used for simulating daily rainfall sequences. The DAR(1) model is also equivalent to a first-order Markov chain model. This model assumes that the probability of rain depends only on the current state (wet or dry) and will not be influenced by its past behavior. Haan et al. (1976), Katz (1977), Roldán and Woolhiser (1982), Small and Morgan (1986), Jimoh and Webster (1996), Sharma (1996), Tan and Sia (1997), and Wilks (1998) are among the studies that were successful in modeling the sequence of rainfall and dry days using first-order Markov chains. Wilks (1998) used the first-order Markov chain to simulate the occurrence of daily rainfall based on data from 1951 to 1996 from 25 stations in New York State, USA. The statistical properties such as the joint probabilities for both rainy and dry days, monthly rainfall, and standard deviations of monthly rainfall indicate that the simulated rainfall data reproduce the rainfall data statistics really well. It was concluded that the model was successful in preserving the dependence nature of daily rainfall at these stations. First-order Markov chains are simple and do not require a lot of computational effort. However, Feyerherm and Bark (1965) found that first-order Markov chains are unable to model the scenario of strong dry day persistence. Similar findings were reported by Wallis and Griffiths (1995) and Semenov et al. (1998). The order of a Markov chain may be influenced by seasonal change and location (Chin 1977; Cazacu and Cipu 2005; Deni et al. 2009). Chin (1977) found that the seasonal change has a significant impact in determining the suitable order of a Markov chain in more than 200 stations located throughout the USA. High-order Markov chains are suitable to model the sequence of daily precipitation during winter at most stations, and first-order Markov chains are appropriate for summer.
The physical environmental causes and geography can influence the order of Markov chains. Similar findings were reported by Cazacioc and Cipu (2005) for the simulation of rainfall sequences at several stations in Romania.

For tropical regions, a different approach was used by Deni et al. (2009) in the analysis of Malaysian daily rainfall data based on the Markov chain model. The objective of their study was to find the optimum order of a Markov chain for daily rainfall during the Northeast (NE) and Southwest (SW) monsoons using two different thresholds, i.e., 0.1 and 10.0 mm, where NE and SW are the directions from which the monsoons are coming. The Akaike information criterion (AIC) and Bayesian information criteria were used to determine the appropriate order of the Markov chain models. The study used the available data from 18 rainfall stations located in various parts of Peninsular Malaysia. They concluded that the optimum order of a Markov chain varies with the location, monsoon season, and the level of threshold. For example, the occurrence of rainfall (threshold level 10.0 mm) for the NE and SW monsoons at stations located in the northwestern and eastern regions of Peninsular Malaysia can be represented using a first-order Markov chain. Additionally, higher order Markov chain models are suitable to represent rainfall occurrence, especially during the NE monsoon, for both levels of threshold. Other examples of the use of a high-order Markov chain to simulate the rain and dry day sequence are reported by Mimikou (1983), Dahale et al. (1994), Katz and Parlange (1998), and Dastidar et al. (2010).

Even though higher order Markov chain models may be used to overcome the lack of persistence of the simple Markov chain, more parameters have to be used, which increases the model uncertainty (Jacobs and Lewis 1983) and also makes the calculations more complex. Jacobs and Lewis (1978) and Kedem (1980) discuss the concept of the stationary DARMA model, which is intended to be simpler for modeling stationary sequences of dependent discrete random variables with specified marginal distribution and correlation structure. Buishand (1977, 1978) modeled the sequence of daily rainfall using DARMA(1,1) at several stations in the Netherlands, Suriname, India, and Indonesia. Since DARMA(1,1) is a stationary model, the data for each station were divided into their respective seasons in order to consider the rainfall seasonal variations. The results have shown that the DARMA(1,1) model is successful in simulating the daily rainfall in tropical and monsoon areas, where prolonged dry and wet seasons may occur. The DARMA(1,1) model provides longer persistence than the DAR(1) model does. Other studies that use the DARMA(1,1) model to simulate sequences of daily rainfall include Chang et al. (1982, 1984b, a), Delleur et al. (1989), and Cindrić (2006). In addition, DARMA models have been applied for the analysis of droughts (Chung and Salas 2000; Salas et al. 2005; Cancelliere and Salas 2010). For example, Chung and Salas (2000) analyzed the annual streamflow time series of the Niger River in Africa and concluded that the drought occurrence can be successfully simulated using the DARMA(1,1) model. The results showed long periods of low flows (drought) and high flows, and the DARMA(1,1) model was suitable for simulating streamflows with a longer memory as compared to the DAR(1) model.

Return periods are useful in hydrology to measure the severity of an event. Various definitions of return period have been reported in the literature, such as first arrival time and interarrival time or recurrence interval. These definitions give different values when the events are dependent in time. However, for single and independent events, the first arrival time and recurrence interval give the same value (Fernández and Salas 1999a). Extensive theories and applications on the return period definitions and serial dependence are discussed in Fernández and Salas (1999a, b), Woodyer et al. (1972), Kite (1978), Lloyd (1970), Loaiciga and Mariño (1991), and Shen (1999) defined recurrence interval as the average elapsed time between the occurrences of critical events, such as earthquakes of high magnitude and extreme floods or droughts. In addition, Vogel (1987) and Douglas et al. (2002) used the return period as the average number of trials required to the first occurrence of a critical event. This definition may be more useful in relation to reservoir operation because knowing the first time that the reservoir is at risk of failure is of greater interest than the average time between failures (Douglas et al. 2002). Furthermore, Goel et al. (1998), Shih and Shen (2001), Kim et al. (2003), González and Valdés (2003), Salas et al. (2005), and Cancelliere and Salas (2004, 2010) reported studies on the calculation of return period and risk that include both the amount and duration of extreme hydrological events.

This study concentrates on the occurrence of multiday rainfall events in Malaysia. The country experiences two major seasons classified as the Northeast (NE) and Southwest (SW) monsoons. The NE monsoon typically occurs from November to March, while the SW monsoon is from May to September. April and October are known as intermonsoon months. Both monsoons bring lots of moisture and as a result, Malaysia receives between 2,000 to 4,000 mm of rainfall with 150 to 200 rainy days annually (Suhaila and Jemain 2007). One of the most devastating recent multiday rainfall events resulted in the Kota Tinggi flood in December 2006 and January 2007. These two extreme monsoon events resulted in more than 350 and 450 mm of cumulative rainfall in less than a week. The estimated economic loss reached half a billion US dollars and more than 100,000 local residents had to be evacuated (Abu Bakar et al. 2007). Even though it is well known that multiday events are the main cause of flooding in Malaysia, the topic has received little attention from local researchers.

This paper discusses various aspects of Malaysian monsoons, including the probability distribution and probability structure of multiday monsoon rainfall events, the modeling and simulation of daily rainfall sequences, the estimation of extreme rainfall quantiles, and the estimation of the return period of multiday rainfall. The occurrences of daily rainfall are characterized and simulated using the discrete autoregressive and moving average [DARMA(1,1)] model. These approaches were tested using the observed daily rainfall measurements collected from Subang Airport near Kuala Lumpur, Malaysia.

**Summary of DAR(1) and DARMA(1,1) Models**

This study uses the DAR(1) and DARMA(1,1) models to simulate the occurrence of daily rainfall. The DAR(1) model is represented as (Jacobs and Lewis 1978)

\[ A_i = V_i A_{i-1} + (1 - V_i) Y_i \]

with \[ A_i = \begin{cases} A_{i-1} & \text{with probability } \lambda \\ Y_i & \text{with probability } (1 - \lambda) \end{cases} \] (1)

where \( V_i \) is an independent random variable taking values of 0 and 1 such that

\[ P(V_i = 1) = \lambda = 1 - P(V_i = 0) \] (2)

and \( \lambda \) is a parameter. The variable \( Y_i \) is another independent and identically distributed (i.i.d.) random variable, with a common probability \( \pi_k = P(Y_i = k) \), \( k = 0,1 \).

It should be noted that \( A_i \) is a first-order Markov chain and the process of simulation is assumed to start at \( A_{-1} \)
The autocorrelation function of the DAR(1) model is (Jacobs and Lewis 1978)

\[ \text{cor}(A_i, A_{i-k}) = r_k(A) = \lambda^k, k \geq 1 \]  

where \( r_k \) is the lag-\( k \) (days) autocorrelation coefficient. The autocorrelation function \( r_k \) is estimated based on the sequences of dry and rainy days, i.e., 0 and 1, and not the rainfall amounts (Delleur et al. 1989) as

\[ r_k = \left[ \frac{N-k}{N} \sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x}) \right] \left[ \frac{N}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right]^{-1} \]

where \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \)

The transitional probability matrix is given by (Jacobs and Lewis 1978) as follows:

\[ P(i, j) = \begin{cases} \lambda + (1 - \lambda)\pi_j, & \text{if } i = j \\ (1 - \lambda)\pi_j, & \text{if } i \neq j \end{cases} \]

where \( \pi_0 = \frac{T_0}{T_0 + T_1} \) and \( \pi_1 = 1 - \pi_0 \)

The one-step transitional probability, \( p(i, j) = P(A_{i+1} = j|A_i = i) \) is given by (Jacobs and Lewis 1978) as follows:

\[ p(i, j) = \begin{cases} \lambda + (1 - \lambda)\pi_j, & \text{if } i = j \\ (1 - \lambda)\pi_j, & \text{if } i \neq j \end{cases} \]

Equation (8) can also be represented in terms of the transitional probability matrix, as shown in Eq. (9)

\[ P = \begin{bmatrix} \lambda + (1 - \lambda)\pi_0 & (1 - \lambda)\pi_1 \\ (1 - \lambda)\pi_0 & \lambda + (1 - \lambda)\pi_1 \end{bmatrix} \]

The probability distribution of wet and dry run lengths can be obtained from Eqs. (10) and (11) as derived by Chang et al. (1984b)

\[ P(T_1 = n) = n^{n-1}(1, 1)[1 - p(1, 1)] \]

\[ P(T_0 = n) = n^{n-1}(0, 0)[1 - p(0, 0)] \]

The DARMA(1,1) model is represented as (Jacobs and Lewis 1978)

\[ X_t = U_tY_t + (1 - U_t)A_{t-1} \]

with \( X_t = \begin{cases} Y_t, & \text{with probability } \beta \\ A_{t-1}, & \text{with probability } (1 - \beta) \end{cases} \)

\[ Y_t \] is another i.i.d. random variable having a common probability \( \pi_k = P(Y_t = k), k = 0,1, \) and \( A_t \) is an autoregressive component given by

\[ A_t = \begin{cases} A_{t-1}, & \text{with probability } \lambda \\ Y_t, & \text{with probability } (1 - \lambda) \end{cases} \]

The three parameters of the DARMA(1,1) model need to be estimated, namely \( \pi_0, \pi_1, \lambda, \) and \( \beta \). The parameters \( \pi_0 \) or \( \pi_1 \) may be estimated from Eqs. (6) to (7). The estimation of \( \lambda \) may be determined by minimizing Eq. (16) using the Newton-Raphson iteration techniques. Buishand (1978) suggested using the ratio of the second to the first autocorrelation coefficients as an initial estimate for \( \lambda \), as shown in Eq. (17)

\[ \lambda = r_2 \]

\[ \beta = \frac{(3\lambda - 1) \pm \sqrt{(3\lambda - 1)^2 - 4(2\lambda - 1)(\lambda - \bar{c})}}{2(2\lambda - 1)} \]

The probability distributions of the wet and dry run lengths for the DARMA(1,1) model are well known in the literature (e.g., Jacobs and Lewis 1978), see the Appendix for more details.

**Probability Distribution and Return Period of Multiday Rainfall Events**

In this section, the probability distribution and return period for multiday rainfall events are investigated considering that the occurrence of daily rainfall is correlated. The return period of multiday rainfall events is based on the number of trials between two successive occurrences of the same event. Multiday rainfall events occur frequently during the NE and SW monsoons. Therefore, it is appropriate to estimate the return period as the average time (in days) between the occurrences of specific events. It may also be referred to as the recurrence interval. The most important parameters that hydrologists and water resources specialists are concerned about when analyzing a multiday rainfall event are the duration and the amount of cumulative rainfall. Hence, this study considers both parameters in formulating the estimation of the return period.
Muhammad (2013) found that the two-parameter gamma function is most suitable for representing the rainfall amount distribution for specified durations at Subang Airport. The method of moments is used to estimate the parameters and the formulas can be found in Mood et al. (1974) or Yevjevich (1984). It should be noted that (e.g., Mood et al. 1974) if two independent gamma variables are added, for example, \( X = R_1 + R_2 \) where both \( R_1 \) and \( R_2 \) are \( \text{gamma}(\alpha, \beta) \), then \( X \) is also \( \text{gamma}(\alpha, \beta) \). Likewise, if you add \( t \) independent \( \text{gamma}(\alpha, \beta) \) variables then \( X = R_1 + R_2 + \ldots + R_t \) is \( \text{gamma}(\alpha, t\beta) \). The data analysis showed that the two-parameter gamma distribution was best for describing the distribution of 1-day and multiday rainfall events at Subang Airport. The empirical representation of rainfall amount distribution in \( t \) consecutive rainy days is given in Eq. (19)

\[
f(x) \approx \frac{1}{24.0} f(0.6t) \left( \frac{x}{24.0} \right)^{0.6t-1} \exp\left( -\frac{x}{24.0} \right) \quad (19)
\]

where \( x = \) total amount of rainfall for \( t \) consecutive rainy days (mm) and \( t = \) number of consecutive rainy days.

To determine the return period of rainfall events, \( E \), of a given duration, \( t \), an approach used previously for determining the return period of droughts (e.g., Cancelliere and Salas 2002; Gonzáles and Valdès 2003; Salas et al. 2005) were followed. It follows that the return period of multiday rainfall events can be determined as

\[
T = \frac{T_t}{P(E|t)}\quad (21)
\]

where \( T_t = \) mean run length for wet days; \( T_0 = \) mean run length for dry days; and \( P(E|t) = \) probability of a rainfall event given by Eq. (20).

Results and Discussion

Probability Distribution of Daily Rainfall

The daily rainfall measurements at Subang Airport (3°31.20’N, 101°33.00’E) were used in this study. A long and reliable record of 52 years for the period 1960 to 2011 was provided by the Department of Meteorology, Malaysia.

Fig. 1 shows the cumulative distribution function (CDF) for 1-day and multiday rainfall at Subang Airport. The figure shows that the two-parameter gamma distribution function given by Eq. (19) fits reasonably well the historical CDF for 1 through 6 days of rainfall duration. The CDF plot shows that for a single rainy day, there is about 60% chance that the rainfall amount will be less than 10 mm, and there is a less than 5% chance that the rainfall amount will exceed 50 mm. The multiday rainfall events resulted in a significant amount of rainfall to the study area. The CDF plot also shows that there is a nonnegligible probability that 2 and 3 consecutive rainy days may produce more than 100 mm of rain. Further, there is 50% of chance of 4, 5, and 6 consecutive rainy days yielding more than 55, 65, and 85 mm of rainfall, respectively. The probability of rainfall events with more than 100 mm of rain increases as the number of consecutive rainy days increases.

![Fig. 1. Cumulative distribution function of 1-day and multiday rainfall events at Subang Airport](image)

results illustrate the important need for a detailed analysis of both duration and magnitude of multiday rainfall events.

Probability Structure of Rainfall Occurrence

When rain occurs on a given day, it is called a wet day while the absence of rain on a given day is called a dry day. In this study, a wet day is indicated for rainfall amounts of more than 0.1 mm, while a dry day is assumed for amounts less than or equal to 0.1 mm. The threshold amount was determined based on the Von Neumann (1941) ratio and a more detailed analysis is presented in Muhammad (2013).

The analysis of the probability structure of rainfall events at Subang Airport shows that more than 50% of the events observed at the study site are rainy days. The estimated probability of rain on any given day is 0.53. If day-to-day rainfall events were independent, the probability of rain on any day would remain constant at 0.53 [shown by a triangle in Fig. 2(a)]. However, Fig. 2(a) shows that the field rainfall measurements of the conditional probability increases significantly as the number of consecutive rainy day increases, i.e., from 0.53 for a single rainy day to about 0.80 for 15 consecutive rainy days. For example, the estimated conditional probability of a fourth rainy day, given that it follows 3 consecutive rainy days, is 0.68. This probability is far greater than the estimated probability of the first day of rain, i.e., 0.53. The occurrence of rain on a given day affects the probability of rain the following days. Thus, the conditional probabilities estimated from the historical data show that the events are dependent.

Likewise, Fig. 2(b) gives the estimated conditional probabilities of \( n \) consecutive dry days at Subang Airport. The estimated probability that any given day is dry is 0.47, which increases significantly to 0.72 after 15 consecutive dry days. For example, the estimated conditional probability for a second consecutive dry day is 0.58, and the estimated probability for the third dry day increases to 0.63. Thus, the probability structure of \( n \) consecutive dry days is also dependent as is the case for rainy days. Table 1 gives the details of the frequency and the estimated conditional probabilities of 1 to 15 consecutive wet and dry days.

Modeling the Occurrence of Daily Rainfall

In this study, the NE and SW monsoons are considered as the daily rainfall recorded during the months of October to March and April
Table 1. Frequencies and Estimated Conditional Probabilities of $t$ Consecutive Wet and Dry Days

<table>
<thead>
<tr>
<th>$t$ Consecutive wet days</th>
<th>Frequency</th>
<th>Estimated conditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10,092</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>6,366</td>
<td>0.63</td>
</tr>
<tr>
<td>3</td>
<td>4,226</td>
<td>0.66</td>
</tr>
<tr>
<td>4</td>
<td>2,875</td>
<td>0.68</td>
</tr>
<tr>
<td>5</td>
<td>2,009</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>1,432</td>
<td>0.71</td>
</tr>
<tr>
<td>7</td>
<td>1,050</td>
<td>0.73</td>
</tr>
<tr>
<td>8</td>
<td>778</td>
<td>0.74</td>
</tr>
<tr>
<td>9</td>
<td>582</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>444</td>
<td>0.76</td>
</tr>
<tr>
<td>11</td>
<td>345</td>
<td>0.78</td>
</tr>
<tr>
<td>12</td>
<td>266</td>
<td>0.77</td>
</tr>
<tr>
<td>13</td>
<td>207</td>
<td>0.78</td>
</tr>
<tr>
<td>14</td>
<td>162</td>
<td>0.78</td>
</tr>
<tr>
<td>15</td>
<td>129</td>
<td>0.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$t$ Consecutive dry days</th>
<th>Frequency</th>
<th>Estimated conditional probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8,901</td>
<td>0.47</td>
</tr>
<tr>
<td>2</td>
<td>5,174</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>3,236</td>
<td>0.63</td>
</tr>
<tr>
<td>4</td>
<td>2,148</td>
<td>0.66</td>
</tr>
<tr>
<td>5</td>
<td>1,455</td>
<td>0.68</td>
</tr>
<tr>
<td>6</td>
<td>1,006</td>
<td>0.69</td>
</tr>
<tr>
<td>7</td>
<td>700</td>
<td>0.70</td>
</tr>
<tr>
<td>8</td>
<td>485</td>
<td>0.69</td>
</tr>
<tr>
<td>9</td>
<td>335</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>236</td>
<td>0.71</td>
</tr>
<tr>
<td>11</td>
<td>161</td>
<td>0.68</td>
</tr>
<tr>
<td>12</td>
<td>117</td>
<td>0.73</td>
</tr>
<tr>
<td>13</td>
<td>85</td>
<td>0.73</td>
</tr>
<tr>
<td>14</td>
<td>61</td>
<td>0.72</td>
</tr>
<tr>
<td>15</td>
<td>44</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Fig. 2. Plot of conditional probability of $t$ consecutive (a) wet days and (b) dry days.
In addition, Fig. 4 gives the wet run length distributions during NE (SW) monsoons for the DARMA(1,1) model is 0.0015 consecutive wet days. Likewise, the sum of squared errors for probabilities for the DAR(1) model are 0.224 (0.243) for 2 probability of 0.207 (0.218). On the other hand, the corresponding (1,1) is 0.197 (0.206), while the observed rainfall data give a distribution for 2 consecutive wet days estimated for the DARMA example, during the NE (SW) monsoon, the theoretical probability amount of error, as compared to the DAR(1) model. For (1,1) model is able to generate the probabilities with the least error, as compared to the DAR(1) model, particularly those that cannot be derived in analytical form. This analysis may also indicate whether some statistics obtained from the limited historical sample may show some evidence of departures or bias.

The statistics estimated in both simulations include the mean and standard deviation of the amount of rainfall, maximum rainfall in a day, lag-1 autocorrelation coefficient, and the maximum wet and dry run lengths. The mean, standard deviation, and the maximum daily rainfall were included to evaluate the statistics of the generated rainfall amounts, while the lag-1 autocorrelation and maximum wet and dry run lengths are used to evaluate the statistics of the simulated sequences of the occurrence of daily rainfall. Table 2 summarizes the statistics of the observed and simulated daily rainfall events at Subang Airport during the NE and SW monsoons.

Generally, the statistics of generated rainfall from simulation A show good results for both monsoons. For NE monsoons, all statistics derived from the observed data fall within two standard deviations relative to the mean calculated from the simulated samples. The results are similar for SW monsoon except for the mean, which falls within three standard deviations. The coefficient of variation obtained from the generated samples are generally small, i.e., on the order of 0.0021 for the mean and standard deviation, 0.065 for the lag-1 correlation, and about 0.16 for the statistics related to the maximums.

For the NE and SW monsoons, the results obtained from simulation B regarding the mean, standard deviation, and lag-1 correlation are similar to those obtained from simulation A. The maximum rainfall obtained from simulation B is about 70% higher than that obtained from the historical sample for NE monsoon and about 100% higher for SW monsoon. The higher results are expected because of the longer sample considered for simulation B. Likewise, the maximum wet run length and the maximum dry run length obtained from simulation B are higher than those obtained from the historical data and the differences are more noticeable for the SW monsoon.

Further verification was made by comparing the probabilities of wet and dry run lengths obtained from the observed data and from simulations A and B. Overall, the analyses show reasonable results. For example, for the NE (SW) monsoon the estimated probabilities for 5 consecutive rainy days derived from the observed data and

![Fig. 3. Observed (dependent) and theoretical ACF for (a) NE monsoons; (b) SW monsoons](image)

![Fig. 4. Probability distribution of wet run lengths for NE monsoons](image)
simulations A and B, respectively, are 0.0588 (0.0469), 0.0550 (0.0458), and 0.0519 (0.0434). Likewise, the estimated probabilities for 7 consecutive dry days obtained from the observed data and simulations A and B, respectively, are 0.0130 (0.0216), 0.0134 (0.0216), and 0.0135 (0.0209). In addition, Fig. 5(a) gives the plot of wet probability distributions obtained from the observed data and simulations A and B for NE monsoons and Fig. 5(b) show results for dry run lengths. Therefore, the results obtained suggest that the DARMA(1,1) model for representing the rainfall occurrence and the two-parameter gamma model for representing the distribution of the rainfall amount for a given rainfall duration give reasonable results for simulating the sequences of daily rainfall for the monsoons at Subang Airport.

**Return Period Curves**

Eqs. (19)–(21) were used to calculate the theoretical return periods, which were then compared with those obtained from the observed data. The threshold amounts of rainfall, $x_0$ (in mm), considered for this analysis are 1, 13, 30, 60, 90, 120, and 150. The smallest amount of 1 mm is selected to represent the majority of rainfall events and 13 mm is the average daily rainfall. The other amounts are selected because they represent significant values of rainfall, especially during multiday events. The return periods were estimated separately for the NE and SW monsoons.

The observed return period, $T$, was calculated for a given duration, e.g. 2 days, for precipitation exceeding a specified threshold, say $x_0$ (in mm). Empirical frequency analysis was made using the Weibull (plotting position) formula and $T = 1/p$, where $p$ is the exceeding probability.

The comparison of the estimated return periods for NE monsoons is shown in Fig. 6(a). In general, the theoretical return period curves obtained from the DARMA(1,1)-gamma model show good agreement with those of the observed data. Fig. 6(a) shows the complex behavior of the return period curves for various rainfall durations and rainfall threshold amounts of the corresponding events. For the smallest amounts, the return periods increase as the rainfall durations increase (e.g., the estimated return periods for multiday rainfall events for amounts >1 mm are higher as compared to the 1-day event). It can be said that the 1-day events occur more often as compared with 2 consecutive days or more. However, for higher rainfall amounts (in mm), e.g., 30, 60, 90, 120, and 150, the return periods decrease as the rainfall duration $t$ increases, reach a minimum, then increase steadily after that. Similar patterns of return period curves are shown in Fig. 6(b) for the case of SW monsoons. Both Figs. 6(a and b) indicate that for the frequent events where there are a large number of historical observations, the return periods estimated from the fitted DARMA(1,1)-gamma models (theoretical) correspond reasonably well with those estimated from the observations. However, in those cases where a very small number of observations are available because of the rarity of the events (e.g., for large rainfall thresholds or large rainfall durations) some significant departures may occur. In addition, the comparison of Figs. 6(a and b) show that while the return period patterns for the NE and SW monsoons are similar, some differences can be noted particularly for rainfall events longer than 4 days.

**Extreme Rainfall Events**

A sample of 1,000,000 days (more than 2,000 years) of daily rainfall was generated to further verify the applicability of Eqs. (19)–(21) for estimating the return period. In this case, the estimations were performed for significant rainfall threshold values, i.e. 50, 100, 150, 200, 250, 300, and 350 mm. Values in

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Observed data NE monsoon (SW monsoon)</th>
<th>Simulation A—Simulated daily rainfall (based on 100 samples, each 9,600 days)</th>
<th>Simulation B—Simulated daily rainfall (based on one sample of 1,000,000 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (mm)</td>
<td>13.4 (12.0)</td>
<td>12.9 (12.9)</td>
<td>12.9 (12.9)</td>
</tr>
<tr>
<td>Standard deviation (mm)</td>
<td>17.6 (16.8)</td>
<td>0.3 (0.3)</td>
<td>0.3 (0.3)</td>
</tr>
<tr>
<td>Maximum rainfall in a day (mm)</td>
<td>171.5 (158.3)</td>
<td>17.2 (17.2)</td>
<td>17.3 (17.2)</td>
</tr>
<tr>
<td>Lag-1 correlation</td>
<td>0.196 (0.192)</td>
<td>0.4 (0.4)</td>
<td>0.4 (0.4)</td>
</tr>
<tr>
<td>Maximum wet run length (days)</td>
<td>31 (17)</td>
<td>24.6 (26.4)</td>
<td>292.2 (325.1)</td>
</tr>
<tr>
<td>Maximum dry run length (days)</td>
<td>21 (20)</td>
<td>3 (3)</td>
<td>34 (27)</td>
</tr>
</tbody>
</table>

**Fig. 5.** Probability distribution of (a) wet run lengths for NE monsoons generated from simulations A and B; (b) dry run lengths for NE monsoons generated from simulations A and B.
excess of 150 mm are considered to represent extreme to rare events and can cause devastating floods on large watersheds.

Fig. 7 shows the comparison between calculated return periods based on the generated sample and the theoretical equations corresponding to NE monsoons. The 2,000-year rainfall was generated using the fitted DARMA(1,1)-gamma models as described earlier. Generally, the return period curves for most rainfall threshold amounts show excellent agreement, which further verifies that Eqs. (19)–(21) are reliable to estimate the return periods for multiday events. Some departures are noted for the very extreme amounts, i.e. 300 and 350 mm, which is attributed to the variability of the generated samples.

Furthermore, it was desired to illustrate the applicability of the fitted DARMA(1,1)-gamma models for obtaining via stochastic rainfall generation the variability of the $T$-year rainfall quantiles, i.e., the annual frequency distribution of the maximum daily rainfall at Subang Airport. Fig. 8 shows the computed annual frequency distribution of maximum daily rainfall obtained from the 52-year historical records and from the 100 samples of 2,000 years of rainfall derived from the generated daily rainfall based on the fitted DARMA(1,1)-gamma models. This result may be particularly useful in cases of short records that may be available at a given site where the occurrence of daily rainfall and the variability of rainfall amount can be modeled using the procedures described in this paper and the variability of the annual frequency distribution obtained based on data generation.

**Summary and Conclusions**

The analysis of both the Northeast (NE) and Southwest (SW) monsoon rainfall precipitation events at Subang Airport in Malaysia from 1960–2011 demonstrates the following:

1. The majority (57%) of rainfall events are multiday events;
2. The distribution of daily rainfall is well reproduced (Fig. 1) with a gamma distribution. Likewise, the distribution of multiday rainfall events is also well reproduced with a gamma distribution. Considering the well-known properties of the sum of independent gamma variables enables the derivation of a simple two-parameter gamma distribution to fit the distribution of daily and multiday rainfall;
3. As expected, the probability of rainfall occurrence (or nonoccurrence) on a given day is not independent, but depends on whether the previous day was dry or wet. The conditional probabilities increase with the number of consecutive rainy (or dry) days (Fig. 2);
4. The rainfall occurrence for both NE and SW monsoons at Subang Airport can be well represented by the DARMA(1,1) model. It reproduces reasonably well a number of key statistics, such as the autocorrelation function (Fig. 3) and run lengths;
5. A simple algorithm has been suggested for estimating the return period for multiday rainfall events defined by combining the DARMA(1,1) model and the gamma distribution. The resulting DARMA(1,1)-gamma model yields good agreement (Fig. 6) between the return periods estimated from the observed historical sample and those estimated by the proposed method. As expected, some departures occur in cases of rare rainfall extremes for which very few observations are available; and

6. The proposed DARMA(1,1)-gamma model also enables the estimation of the variability in T-year daily maximum rainfall, which could be especially useful for the analysis of extreme rainfall precipitation in areas with short historical records.

Appendix. Probability Distributions of the Wet and Dry Run Lengths for DARMA(1,1) Model

The procedures to estimate the probability distributions of wet and dry run lengths are given in this section.

The one-step transitional probabilities \( H_k(u, v) \) can be written as (Jacobs and Lewis 1978)

\[
H_k(u, v) = P(X_{t+1} = k, A_{t+1} = v|X_t = m, A_t = u) = P(X_{t+1} = k, A_{t+1} = v|A_t = u)
\]

where \((X_{t+1}, A_{t+1})\) is independent of \(X_t\) and \(u, v, k,\) and \(m\) are 0, 1 values.

The transition probability matrices are

\[
H_0(u, v) = \begin{bmatrix}
\lambda(1-\beta) + [1 - \lambda(1-\beta)]\pi_0 & (1 - \beta)(1 - \lambda)\pi_1 \\
\beta(1 - \lambda)\pi_0 & \beta\lambda\pi_0
\end{bmatrix}
\]

\[
H_1(u, v) = \begin{bmatrix}
\beta\lambda\pi_1 & (1 - \beta)(1 - \lambda)\pi_1 \\
(1 - \beta)(1 - \lambda)\pi_0 & \lambda(1 - \beta) + [1 - \lambda(1-\beta)]\pi_1
\end{bmatrix}
\]

Lloyd and Salem (1979) introduced the use of label variable \( W_t = 2X_t + A_t \) to convert the first-order bivariate Markov chain \((X_t, A_t)\) into a four-state simple Markov chain. \((X_t, A_t)\) can have values of 0 or 1, so there are four possibilities for the value of \( W_t \); i.e., \(0, 1, 2, 3\). Table 3 summarizes the \( W_t \) values.

The value of 0 and 1 for \( W_t \) corresponds to the state of 0 in \( X_t \), which implies a dry day. In the same manner, a wet day is represented as 1 in \( X_t \), which gives the value of 2 and 3 for \( W_t \).

The transition probabilities are given as

\[
p_w(0,1) = P(X_{t+1} = 0, A_{t+1} = 1|X_t = 0, A_t = 0) = P(X_{t+1} = 0, A_{t+1} = 1|A_t = 0) = H_0(0,1)
\]

\[
p_w(0,2) = P(X_{t+1} = 1, A_{t+1} = 0|X_t = 0, A_t = 0) = P(X_{t+1} = 1, A_{t+1} = 0|A_t = 0) = H_1(0,0)
\]

\[
p_w(0,3) = P(X_{t+1} = 1, A_{t+1} = 1|X_t = 0, A_t = 0) = P(X_{t+1} = 1, A_{t+1} = 1|A_t = 0) = H_1(0,1)
\]

\[
p_w(1,0) = P(X_{t+1} = 0, A_{t+1} = 0|X_t = 0, A_t = 1) = P(X_{t+1} = 0, A_{t+1} = 0|A_t = 1) = H_0(1,0)
\]

Transition probability matrix, \( Q \), of the univariate Markov chain \( W_t \) is

\[
Q = \begin{bmatrix}
0 & 1 & 2 & 3 \\
H_0(0,0) & H_0(0,1) & H_1(0,0) & H_1(0,1) \\
H_0(1,0) & H_0(1,1) & H_1(1,0) & H_1(1,1) \\
H_0(2,0) & H_0(2,1) & H_1(2,0) & H_1(2,1) \\
H_0(3,0) & H_0(3,1) & H_1(3,0) & H_1(3,1)
\end{bmatrix}
\]

and its marginal distribution is

\[
P[W_t = 0] = P(X_t = 0, A_t = 0) = P(X_t = 0, A_t = 0|A_{t-1} = 0)P(A_{t-1} = 0) \\
+ P(X_t = 0, A_t = 0|A_{t-1} = 1)P(A_{t-1} = 1) = H_0(0,0)\pi_0 + H_0(1,0)\pi_1
\]

\[
P[W_t = 1] = P(X_t = 0, A_t = 1) = P(X_t = 0, A_t = 1|A_{t-1} = 0)P(A_{t-1} = 0) \\
+ P(X_t = 0, A_t = 1|A_{t-1} = 1)P(A_{t-1} = 1) = H_0(0,1)\pi_0 + H_0(1,1)\pi_1
\]

\[
P[W_t = 2] = P(X_t = 1, A_t = 0) = P(X_t = 1, A_t = 0|A_{t-1} = 0)P(A_{t-1} = 0) \\
+ P(X_t = 1, A_t = 0|A_{t-1} = 1)P(A_{t-1} = 1) = H_1(0,0)\pi_0 + H_1(1,0)\pi_1
\]

\[
P[W_t = 3] = P(X_t = 1, A_t = 1) = P(X_t = 1, A_t = 1|A_{t-1} = 0)P(A_{t-1} = 0) \\
+ P(X_t = 1, A_t = 1|A_{t-1} = 1)P(A_{t-1} = 1) = H_1(0,1)\pi_0 + H_1(1,1)\pi_1
\]

Probability distributions of wet and dry run lengths of \( t \) consecutive days for the DARMA(1,1) model, denoted by \( P(T_t = t) \) and \( P(T_0 = t) \), respectively, can be calculated using conditional probabilities, as given by Chang et al. (1984b)

\[
P(T_t = t) = P(X_0 = 0, X_1 = 1, \ldots, X_t = 1) \\
= P(X_0 = 0, X_1 = 1)P(X_1 = 1, X_2 = 1)\cdots P(X_{t-1} = 1, X_t = 1) = \frac{P(X_0 = 0, X_1 = 1)}{P(X_0 = 0, X_1 = 1)}
\]

Note that

\[
P(X_0 = 0, X_1 = 1, \ldots, X_t = 1, X_{t+1} = 0) = \{P[W_0 = 0]|H_0(0)(0) \}
\]

\[
- H_1^{(n+1)}(0)\} + P[W_0 = 1]|H_1(1)(1) - H_1^{(n+1)}(1)\}
\]

where \( H_1^{(n)}(j) = H_1^{(n)}(j, 0) + H_1^{(n)}(j, 1); j = 0, 1 \)

Both \( H_1(j, 0) \) and \( H_1(j, 1) \) are elements of the \( n \) step transition probability matrix

\[
P(X_0 = 0, X_1 = 1) = \sum_{k=0}^{n} \sum_{j=0}^{1} H_1(j, k)\left[\pi_j - \sum_{l=0}^{1} H_1(l, j)\pi_l\right]
\]

Table 3. Four State Markov Chain, \( W_t \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_t )</td>
<td>0</td>
</tr>
<tr>
<td>( A_t )</td>
<td>0</td>
</tr>
<tr>
<td>( W_t )</td>
<td>0</td>
</tr>
</tbody>
</table>
\[ P(T_0 = t) = P(X_0 = 1, X_1 = 0, \ldots, X_t = 0, \ldots, X_t = 0, X_{t+1} = 1) = P(X_0 = 1, X_1 = 0, X_2 = 0, X_3 = 0) \]

\[ P(X_0 = 1, X_1 = 0, \ldots, X_t = 0, X_{t+1} = 1) = \left\{ P[W_0 = 2] \left[ H_0^{(n)}(0) - H_0^{(n+1)}(0) \right] \right. \\
+ \left. P[W_0 = 3] \left[ H_0^{(n)}(1) - H_0^{(n+1)}(1) \right] \right\} \]

where \( H_0^{(j)}(j, j) = H_0^{(j, 0)}(j, 0) + H_0^{(j, 1)}(j, 1) \), \( j = 0, 1 \), are elements of the \( n \) step transition probability matrix.

\[ P(X_0 = 1, X_1 = 0) = \sum_{k=0}^{1} \sum_{l=0}^{1} H_k^{(n)}(j, k) \left[ \pi_j - \sum_{l=0}^{1} H_0^{(l, j)} \pi_l \right] \]

where \( H_k^{(n)}(j, k), H_0^{(n)}(l, j) \) are elements of the \( n \) step transition probability matrix.

**Acknowledgments**

This study has been carried out at Colorado State University during the Ph.D. studies of the first author. Financial support for the first author from the Ministry of Education, Malaysia and Universiti Kebangsaan Malaysia is gratefully acknowledged.

**References**


