RAINFALL DATA SIMULATION BY THE HIDDEN MARKOV MODEL AND DISCRETE WAVELET TRANSFORMATION

A. W. Jayawardena[1]

International Centre for Water Hazard and Risk Management, Public Works Research Institute, Tsukuba, Japan

P. C. Xu

Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, China

W. K. Li

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

ABSTRACT: In a specific region, monthly (or bimonthly) rainfall data can be considered as deterministic while daily rainfall data are almost completely random. As a result, deterministic models cannot fit the daily data because of the strong stochastic nature while stochastic models cannot fit into daily rainfall time series because of the deterministic nature in the large scale. Although there are different approaches for simulating daily rainfall, mixing of deterministic and stochastic models has not hitherto been proposed. An attempt is made in this study to simulate daily rainfall data by utilizing a discrete wavelet transformation (DWT) and Hidden Markov model (HMM). We use a deterministic model to obtain large-scale data, and a stochastic model to simulate the wavelet tree coefficients. The simulated daily rainfall is obtained by inverse transformation. We then compare the accumulation of simulated and observed data from the Chao Phraya Basin in Thailand. Because of the stochastic nature in the small scale, the simulated daily rainfall would on a point to point comparison show deviations with the observed data. However the accumulations of simulated data do show some level of agreement with the observed data.

KEYWORDS: Daily rainfall; discrete wavelet transformation; hidden Markov model; EM algorithm; false nearest neighbour; phase space reconstruction

[1] Corresponding author.
E-mail address: hrecjaw@hkucc.hku.hk (A. W. Jayawardena)

INTRODUCTION

Rainfall data constitute one of the most important time series in hydrology. There are two approaches of simulating rainfall time series, namely deterministic and stochastic. Because of the strong stochastic nature of daily rainfall, stochastic methods are more widely used in hydrology than deterministic methods. Salas (1993), Wilks (1998), and Sharma and Lall (1999) have suggested different approaches of simulating daily rainfall data (see also a review of daily rainfall models for Australia by Chapman (1994)). However, because of the strong deterministic character of the large-scale rainfall data and the strong stochastic character of the small-scale data, such models cannot produce successful simulations. In this paper, we introduce a combination of deterministic and stochastic models, and try to simulate rainfall data of such characteristics.

In the proposed approach, a signal (daily rainfall in this case) is decomposed into sub-signals with different scales, i.e., a large-scale signal and several small-scale signals. The approach is an alternative to the frequency domain analysis of a signal via Fourier transforms by which the frequency content of the signal is obtained. The Discrete Wavelet Transform (DWT) approach proposed in the study is capable of not only providing the frequency content in the signal but also the times of occurrences of each frequency component thereby giving a multiple resolution of the signal. The approach also has the added advantage of not requiring the assumptions of stationarity and periodicity of the time series. Related studies on multi-scale analysis in hydrology have been carried out to characterize daily stream flow (Smith et al., 1998), monthly reservoir inflows (Coulibaly et al., 2000), and to generate streamflow data (Bayazit and Aksoy, 2001). The DWT approach has also been used to generate rainfall data (Unal et al., 2004). Such studies however have not addressed the issue of the mixed-scale composition (large and small) of daily rainfall data.

In this study, the low frequency component of the rainfall series is considered to emanate from a deterministic system and therefore is represented by a deterministic model. Several types of deterministic models are available in the literature, but a local linear model is adopted in this study. The high frequency component of the series which is considered to emanate from a stochastic system is analyzed using stochastic and statistical methods. The methods developed and applied include the Independent Mixture Model (Chipman, 1997; Wong and Li, 2000), the Hidden Markov Chain Model (Rabiner, 1989; Aksoy and Bayazit, 2000), and the Hidden

Markov Tree Model (Rabiner, 1989). The high frequency wavelet signal which is considered as a random variable is assumed to follow a mixture model that consists of the weighted combination of several Gaussian distributions, whose weights themselves are stochastic and are functions of a pre-assigned number of hidden states, and, state and transition probabilities. The state probabilities are treated as Markov processes. By assuming that the transition probabilities are the same for the same period of time in different years, the EM algorithm (Crouse *et al.*, 1998; McLachlan, 1997; Ronen *et al.*, 1995; Dempster *et al.*, 1997) is applied to estimate the state and transition probabilities (Rabiner, 1989) for the Markov model. Once the state probabilities are estimated, the wavelet coefficients are simulated by Monte Carlo method. After the decomposition, and together with the low-frequency component, the daily rainfall data can be simulated via inverse transformation.

The approach is then applied to simulate three daily rainfall time series from the Chao Phraya Basin (CPB) in Thailand. The first data series is from the gauging station No.111 (CPB111) for the period April 1, 1980 to March 31, 1994, the second from the gauging station No.112 (CPB112) for the period April 1, 1980 to July 31, 1994, and the third from the gauging station No.117 (CPB117) for the period April 1, 1980 to July 31, 1994. The statistics of the three data sets are given in Table 1.

The flow of the paper is as follows: Since several mathematical/statistical tools, some not so familiar in the hydrological context, have been used in the study, a brief account of them is first presented. Attempts will be made to explain these ideas and techniques briefly and provide references for detailed analysis and explanations. It is followed by a section on discussing a tree model used in the paper. As there are many parameters (e.g. state and transition probabilities, number of states (M) and number of scales (N)) in the model, their methods of estimation will then be given. When the model is well established, it is examined and applied to simulate rainfall data for 3 gauging stations. Finally, a brief analysis of the simulated data and the limitations of the approach will be given.

SOME PRELIMINARIES

In this section, a brief introduction to wavelet decomposition, discrete wavelet transformation, hidden Markov tree model, and Monte Carlo method will be given. It is by no means complete. A more rigorous treatise can be found in standard textbooks (e.g. Daubechies (1992) and Sobol' (1994)).

2.1 Introduction to wavelet decomposition

Wavelet decomposition, or wavelet transform, is a relatively new technique of signal processing whereby a time series can be viewed in multiple resolutions with each resolution reflecting a different frequency. It has several advantages over the traditional method of frequency domain analysis by Fourier transforms. For example, wavelet decomposition can handle situations where the signal has sharp peaks or discontinuities, which the Fourier transform approach cannot. It is also capable of giving the time and frequency information simultaneously (subject to the limitations imposed by the Heisenberg Uncertainty Principle), i.e. the time-frequency representation of the signal, whereas the Fourier transform approach can give only the frequency information. The times of occurrences of the frequencies remain unknown. Fourier transform approach can be used in situations where the times of occurrences of various frequency components are not of interest but only interested in what frequency components exist. The Short Time Fourier Transform (STFT) or, Windowed Fourier Transform, which is a variant of the normal Fourier transform, and which is capable of giving the times of occurrences of a band of frequencies rather than the exact frequency, can be considered as an improved version which still has problems. Wavelets, which can be thought of as a spinoff from STFT overcome many such problems. Unlike in Fourier analysis, the wavelet analysis does not require assumptions about stationarity and periodicity of data.

The basic approach in Fourier as well as wavelet decomposition is to convolute the signal function by a basis function. In the case of Fourier approach the basis function is a combination of Sines and Cosines. In the case of wavelet approach, there exist a number of different basis functions such as for example, the Harr (Harr, 1910) wavelet, the Daubechies (Daubechies, 1992) wavelet, the Mexican Hat (normalised second derivative of a Gaussian function) wavelet among others. Of these, the Harr wavelet has the major advantages of been conceptually simple, computationally fast and exactly reversible. Among the many popular uses of the Harr wavelet is its application in the JPEG format of digital image compression.

Each step of a wavelet transformation produces a set of averages and a set of differences thereby halving the size of the input data. With an input data size of 2^N (almost all wavelet algorithms work with data expressed as a power of two) recursive repetition of this process leaves with one (2^0) sum and 2^N -1 differences. The differences are referred to as wavelet coefficients.

2.2 Discrete Wavelet Transformation

For a given signal with 2^N samples, $\{C_{0,1}, C_{0,2}, \ldots, C_{0,2^N}\}$, in which the subscript 0 refers to the finest scale, i.e. 0-order (also known as 0-scale), the following two equations can be introduced on the assumption that these numbers are not random and have some correlation structure (for $1 \le n \le 2^{N-1}$):

$$C_{1,n} = \frac{C_{0,2n-1} + C_{0,2n}}{2}, \tag{1}$$

$$D_{1,n} = \frac{C_{0,2n} - C_{0,2n-1}}{2C_{1,n} + 1}. \tag{2}$$

It should be noted that the decomposition according to Eqs. (1) and (2) is slightly different from that given for Haar Wavelet base (Mallet, 1997; Daubechies, 1992). The proposed decomposition confines the wavelet data to the interval $(-1, 1)$, thus making it easier for stochastic simulation. In the Harr wavelet, the wavelet coefficients are taken as half the differences of a pair of consecutive data values, and the averages or the smoothed value as half the sum. The second term in the denominator of Eq. 2 is introduced to avoid division by zero.

The 0-order data is now transformed into 1-order data which carries some information of the original signal:

$$\{C_{1,1}, C_{1,2}, \ldots, C_{1,2^{N-1}}, D_{1,1}, D_{1,2}, \ldots, D_{1,2^{N-1}}\}.$$

It should be noted that, the 1-order data can be inversely transformed back to 0-order data, as follows:

$$C_{0,2n-1} = \frac{2C_{1,n} - D_{1,n}(2C_{1,n} + 1)}{2}, \tag{3}$$

$$C_{0,2n} = \frac{2C_{1,n} + D_{1,n}(2C_{1,n} + 1)}{2}. \tag{4}$$

Next, we keep $\{D_{1,1}, D_{1,2}, \ldots, D_{1,2^{N-1}}\}$ fixed and process the 2^{N-1} data points

$\{C_{1,1}, C_{1,2}, \ldots, C_{1,2^{N-1}}\}$ by using the more general formulae (for $1 \le k \le N$, $1 \le n \le 2^{N-k}$, where k can be considered as the scale level in the wavelet tree (Section 1), while n can be considered as the position of nodes:

$$C_{k,n} = \frac{C_{k-1,2n} + C_{k-1,2n-1}}{2}, \tag{5}$$

$$D_{k,n} = \frac{C_{k-1,2n} - C_{k-1,2n-1}}{2C_{k,n} + 1}, \tag{6}$$

that yield the 2-order data

$$\{C_{2,1}, \ldots, C_{2,2^{N-2}}, D_{2,1}, \ldots, D_{2,2^{N-2}}, D_{1,1}, D_{1,2}, \ldots, D_{1,2^{N-1}}\}.$$

By transforming the data successively, the *N*-order data can be obtained as:

$$\{C_{N,1}, D_{N,1}, D_{N-1,1}, D_{N-1,2}, \ldots, D_{k,1}, \ldots, D_{k,2^{N-k}}, \ldots, D_{1,1}, D_{1,2}, \ldots, D_{1,2^{N-1}}\}.$$

From another point of view, if the information $C_{N,1}$ and all $D_{k,n}$ are known, the original values $\{C_{0,1}, C_{0,2}, \ldots, C_{0,2^N}\}$ can be recovered via inverse transformation. Here, $C_{N,1}$ is the large-scale (coarsest scale) data, while $D_{k,n}$ are known as wavelet coefficients. Although the above analysis is for signals with 2^N samples, signals that have lengths different from a power of 2 can also be transformed into signals of length 2^N by adding zeros to one or both ends.

This decomposition is originated from the input signal $C_{0,n}$, which is first transformed into scale '1' data $\{C_{1,n}, D_{1,n}\}$. Then each scale '*k*' data $\{C_{k,n}, D_{k,n}\}$ is further transformed into scale '$k+1$' data $\{C_{k+1,n}, D_{k+1,n}\}$, until the largest scale (*N* in this case) is reached. Without ambiguity, it should be noted that the scale '*k*' in $C_{k,n}$ is ranging from 0 to *N*, while the scale '*k*' in $D_{k,n}$ is ranging from 1 to *N*.

2.3 Monte Carlo Method

The central theme of Monte Carlo method is to simulate an arbitrary continuous random variable by a random variable γ which is uniformly distributed on $[0,1]$. Suppose we are given a continuous random variable ξ with probability density function $f(x)$, then the probability distribution function of ξ is defined as

$$F(x) = \Pr(\xi \le x) = \int_{-\infty}^{x} f(t)dt\,. \tag{7}$$

As x increases from $-\infty$ to $+\infty$, $F(x)$ is monotonically increasing from 0 to 1. The idea of Monte Carlo method is first determining γ from a uniform distribution on $[0,1]$, then using it to obtain the value ξ by the equation

$$\gamma = \int_{-\infty}^{\xi} f(t)dt = F(\xi)\,. \tag{8}$$

Next, consider a more general example. Suppose we have a mixture of densities,

$$p(x) = \sum_{i=1}^{M} c_i p_i(x) \tag{9}$$

where each $p_i(x)$ is itself a density function, $c_i > 0$ and $\sum_{i=1}^{M} c_i = 1$, then we have $P(x) = \sum_{i=1}^{M} c_i P_i(x)$ where each $P_i(x)$ is the distribution of $p_i(x)$. To simulate a random variable ξ with mixture of densities $p(x)$, we introduce another discrete random variable κ, so that $\Pr(\kappa = i) = c_i$ $(i = 1, \ldots, M)$ and define a two-stage modelling scheme: select 2 random numbers γ_1 and γ_2 from $[0,1]$, then

1) use γ_1 to define a random value $\kappa = i$ and
2) use γ_2 to define ξ.

Explicitly, $\gamma_2 = P_i(\xi)$ if $c_1 + \cdots + c_{i-1} \le \gamma_1 \le c_1 + \cdots + c_i$ (here $c_1 + \cdots + c_{i-1}$ is defined to be 0 when $i = 1$). The distribution function of ξ defined in this way is exactly $P(x)$.

2. 4 False Nearest Neighbourhood (FNN) method for test of determinism

For a dynamical system $\mathbf{x}(t) \to F(\mathbf{x}(t)) = \mathbf{x}(t+1)$, it is possible to define a new d_e-dimensional Euclidean space of vectors $\mathbf{y}(t)$, in which the evolution in time $\mathbf{y}(t) \to \mathbf{y}(t+1)$ follows that of the unknown dynamics $\mathbf{x}(t) \to \mathbf{x}(t+1)$.

If $\mathbf{y}(t) = [x(t), x(t+T), \ldots, x(t+(d_e-1)T)]$ represents a vector point at time level t in a (time-delay) reconstructed phase space of dimension d_e and time delay T, then there must exist another point $\mathbf{z}(s)$, defined as

$$\mathbf{z}(s) = [x(s), x(s+T), \ldots, x(s+(d_e-1)T)] \tag{29}$$

at time level $s \ne t$, that satisfies

$$\| \mathbf{z}(s) - \mathbf{y}(t) \| \leq \| \mathbf{w}(u) - \mathbf{y}(t) \|, \qquad (30)$$

for every point $\mathbf{w}(u)$ (defined in similar fashion as $\mathbf{y}(t)$, and $u \neq t$) in the same reconstructed phase space. Here, $\| \cdot \|$ is the usual Euclidean norm. In other words, $\mathbf{z}(s)$ is the nearest point in the Euclidean space to $\mathbf{y}(t)$. Then $\mathbf{z}(s)$ is called the nearest neighbour of $\mathbf{y}(t)$ and can be written as $\mathbf{y}^{NN}(t)$.

Such a neighbour $\mathbf{y}^{NN}(t)$ is called a true neighbour, if it has come to its neighbourhood through dynamical origins, and is called a false nearest neighbour (FNN) of $\mathbf{y}(t)$ if it arrives in its neighbourhood by projection from a higher dimension. To check whether a nearest neighbour is true or false, we compare the distance between the points $\mathbf{y}^{NN}(t)$ and $\mathbf{y}(t)$ in dimension d_e with those in higher dimension $d_e + 1$. This is checked using the approximate condition that if $\mathbf{y}^{NN}(t)$ is the nearest neighbour of $\mathbf{y}(t)$, and if

$$\frac{| x(t + d_e T) - x^{NN}(t + d_e T) |}{\| \mathbf{y}(t) - \mathbf{y}^{NN}(t) \|} > 15, \qquad (31)$$

then $\mathbf{y}^{NN}(t)$ is a FNN of $\mathbf{y}(t)$ (Abarbanel, 1996). If for a certain d_e, the percentage of FNN’s is less than 5%, then d_e is accepted as the embedding dimension. It is expected that the percentage of FNN’s drops from nearly 100 in dimension 1, to zero when the true parameter value (in this case, value of N) is reached.

3. TREE MODEL FOR DAILY RAINFALL SERIES

For a series of given daily rainfall data r_i $(1 \leq i \leq N_{\max})$; $N_{\max}$ is the length of the rainfall time series, a new time series u_n, which gives the mean rainfalls for every 2^N days can be constructed as

$$u_n = \frac{1}{2^N} \sum_{i=1}^{n2^N} r_{(n-1)2^N + i}, \qquad (10)$$

where N is the given scale (assigned to be 6 in this study, and will be explained in Section 5). The new time series u_n can be considered as deterministic and therefore predictable for sufficiently large scale N (in this study, $N = 6$ means that we are considering the means of 64 days). To estimate the daily rainfall using the predicted value $C_{N,1}$ of the mean rainfall in a 2^N day period, a reasonable thought is to estimate the mean rainfalls $C_{N-1,1}$ and $C_{N-1,2}$ at the $N-1$ scale level, $C_{N-2,1}$, $C_{N-2,2}$, $C_{N-2,3}$ and $C_{N-2,4}$ at the $N-2$ scale level etc. successively, until the daily scale level $C_{0,1}$, $C_{0,2}$, …, $C_{0,2^N}$ is reached. However, to obtain the mean rainfall at the $N-1$ scale level, using the predicted rainfall at the N scale level, more information is needed. Since $C_{N,1}$, and $C_{N-1,1}$ & $C_{N-1,2}$ respectively are the mean rainfalls at N and $N-1$ scales, they satisfy the equation

$$C_{N,1} = \frac{C_{N-1,1} + C_{N-1,2}}{2}. \tag{11}$$

To estimate the $N-1$ scale mean rainfall, another variable is introduced:

$$D_{N,1} = \frac{C_{N-1,2} - C_{N-1,1}}{2C_{N,1} + 1}. \tag{12}$$

Then, the $N-1$ scale mean rainfall can be estimated by known $C_{N,1}$ and $D_{N,1}$ as

$$C_{N-1,1} = \frac{2C_{N,1} - (2C_{N,1} + 1)D_{N,1}}{2}, \tag{13}$$

$$C_{N-1,2} = \frac{2C_{N,1} + (2C_{N,1} + 1)D_{N,1}}{2}. \tag{14}$$

Eqs. (11) and (12), in general are

$C_{k,n} = \dfrac{C_{k-1,2n} + C_{k-1,2n-1}}{2}$, which is the same as Eq. (5), and,

$D_{k,n} = \dfrac{C_{k-1,2n} - C_{k-1,2n-1}}{2C_{k,n} + 1}$, which is the same as Eq. (6),

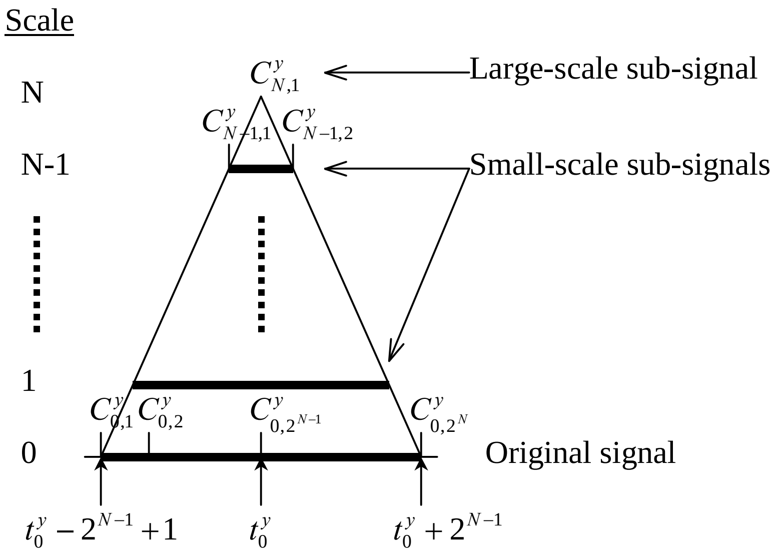
for all the scales $1 \le k \le N$. Then $C_{k-1,2n}$ and $C_{k-1,2n-1}$ can be obtained for all n, $1 \le n \le 2^{N-k}$. The estimation of daily rainfall data, $C_{0,n}$ for all n, can then be obtained by sequential application of this procedure. Therefore, the remaining task is to find the large-scale signal $C_{N,1}$ and the wavelet coefficients $D_{k,n}$. In the above notation, $C_{k,n}$ is the *k*-order “scale” signal at position n, and $D_{k,n}$ is the *k*-order “wavelet” signal at position n in the k^{th} layer (Figure 1). In this study, 0-order scale data correspond to daily rainfall data (2^0), 1-order correspond to 2-day data (2^1), 2-order correspond to 4-day data (2^2), 3-order correspond to 8-day data (2^3), etc., and *N*-order data correspond to 2^N-day data. For example,

$C_{0,n} = r_n$ - daily data

$$C_{1,k} = \frac{1}{2}(r_{2k-1} + r_{2k}) = \frac{C_{0,2k-1} + C_{0,k}}{2} \quad \text{- average of 2 days}$$

$$C_{2,k} = \frac{1}{4}(r_{4k-3} + r_{4k-2} + r_{4k-1} + r_{4k}) = \frac{C_{1,2k-1} + C_{1,2k}}{2} \quad \text{- average of 4 days}$$

For simplicity, the above descriptions are restricted to a single signal. However, several signals from a given historical data set will be used to model the large-scale prediction ($C_{N,1}$) and wavelet tree simulations ($D_{k,n}$). Therefore, the above concept is applied with several signals, by adding a superscript y to denote the data in year y as follows:



Throughout this paper, the notations (for example, $C_{k,n}$ and $D_{k,n}$) without superscript y are refer to the data that are to be estimated, whereas the those (for example, $C^y_{k,n}$ and $D^y_{k,n}$) with a superscript y refer to the historical data that have been used in the simulation process.

By using the above decomposition method the (observed) 0-order scale data $C_{0,n}$ can be decomposed as follows:

$$\begin{tikzcd}
C_{0,n} \arrow[r] \arrow[dr] & C_{1,n} \arrow[r] \arrow[dr] & C_{2,n} \arrow[r] \arrow[dr] & \cdots \arrow[r] \arrow[dr] & C_{N-1,n} \arrow[r] \arrow[dr] & C_{N,n} \\
& D_{1,n} & D_{2,n} & \cdots & D_{N-1,n} & D_{N,n}
\end{tikzcd}$$

Conversely, the (simulated) 0-order scale data can be reconstructed by the wavelet signal and the large-scale signal as:

$$\begin{tikzcd}
C_{0,n} & C_{1,n} \arrow[l] & C_{2,n} \arrow[l] & \cdots\cdots \arrow[l] & C_{N-1,n} \arrow[l] & C_{N,n} \arrow[l] \\
& D_{1,n} \arrow[ul] & D_{2,n} \arrow[ul] & \cdots\cdots \arrow[ul] & D_{N-1,n} \arrow[ul] & D_{N,n} \arrow[ul]
\end{tikzcd}$$

Since the large-scale signal of a daily rainfall data is assumed to be deterministic, the N-order scale data $C_{N,1}$ can be predicted using the historical data for large N. In this study, a local linear model is used for this purpose and a tree model is used to simulate the wavelet signal $D_{k,n}$, $(1 \le k \le N, 1 \le n \le 2^{N-k})$.

In the "wavelet tree" (Figure 2), the node (k,n) where k is the layer number and n is the position number in layer k has the parent $(k+1,[(n+1)/2])$ while the offsprings are $(k-1,2n-1)$ and $(k-1,2n)$. Here the function $[x]$ returns the largest integer smaller than x (the parent-child terminology is used in related papers, e.g. Ronen *et al.* (1995) and Crouse *et al.* (1998)).

The data in the wavelet tree are all stochastic, and therefore a stochastic method must be used for simulation. A simple probability function such as the Gaussian distribution will not be suitable because the data in the wavelet tree which are built by the daily rainfall data, will contain many small values (See Table 1 which gives the dry probability, an indicator of the number of days with zero rainfall), and some large values. In this study, a mixture model, a combination of several Gaussian distribution functions, is used to simulate the wavelet signal:

$$p(D_{k,n}) = \sum_{i=1}^{M} v_{k,n}(i) \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{D_{k,n}{}^2}{2\sigma_i^2}), \qquad (15)$$

where $p(x)$ is the probability distribution of the wavelet signal $D_{k,n}$. The weighted value for each Gaussian distribution function, $v_{k,n}(i)$, is also stochastic. It is simulated by using a new random variable $S_{k,n}$, called the hidden state variable, which has values of $\{1, 2, 3, \ldots, M\}$. The weighted value $v_{k,n}(i)$ is equal to the probability of hidden state variable in state i (i.e. $v_{k,n}(i) = \Pr(S_{k,n} = i)$). In the above definitions, i is the number of hidden states assumed $(i = 1, 2, 3, \ldots, M)$, M is the maximum number of hidden states, and σ_i^2 is the variance of the Gaussian distribution.

It can also be seen that the weighted values for the mixture model for the data points $D_{k,n}$ and

$D_{k-1,2n-1}$ or $D_{k-1,2n}$ are dependent. A large value of $D_{k,n}$ always means that one value of either $D_{k-1,2n-1}$ or $D_{k-1,2n}$ is large. So the weighted value $v_{k-1,2n-1}(i)$, (or, $v_{k-1,2n}(i)$) for the data point $D_{k-1,2n-1}$ (or, $D_{k-1,2n}$) which is equal to the probability of the hidden state variable $S_{k-1,2n-1}$ (or, $S_{k-1,2n}$) depends on the weighted value $v_{k,n}(i)$, the probability of the hidden state variable $S_{k,n}$ equal to *i*. Since probability of transition of the hidden state variable from $S_{k,n}$ to $S_{k-1,2n-l}$ $(l = 0, 1)$ could vary with position (k, n), we introduce the transition probabilities as follows:

$$T_{k-1,2n-l}(i,j) = \Pr(S_{k-1,2n-l} = i \mid S_{k,n} = i) \qquad , \ l = 0, 1. \qquad (16)$$

Then the weighted values $v_{k-1,2n}(i)$, or, $v_{k-1,2n-1}(i)$ satisfy the Markov condition

$$v_{k-1,2n-l}(i) = \sum_{j=1}^{M} T_{k-1,2n-l}(i,j) v_{k,n}(j) \qquad , \ l = 0, 1. \qquad (17)$$

In order to simulate the daily rainfall data by the above approach, we need to know the following parameters: the *N*-order scale data $C_{N,1}$; the number of hidden states M; the variance for each hidden state σ_i^2; the weighted value $v_{N,1}(i)$ for each Gaussian distribution, i.e. the probability of the hidden state random variable $S_{N,1} = i$ and the transition probabilities $T_{k,n}(i,j)$. All the remaining $v_{k,n}(i)$ can then be obtained by Eq. (17) and the estimated $v_{N,1}(i)$. In this study, we fix the number of the hidden states M and their variances σ_i^2, $1 \le i \le M$, *a priori*, and focus on discussing the other parameters.

SIMULATION PROCEDURE

The tree model and the hidden Markov model can now be used to simulate the daily rainfall data. For a given prediction origin t_0, the mean rainfall value for 2^N days $[t_0 - 2^{N-1} + 1, t_0 + 2^{N-1}]$ is estimated and denoted as $C_{N,1}$. Since the mean rainfall data $C_{N-1,1}$ of 2^{N-1} days period $[t_0 - 2^{N-1} + 1, t_0]$ is known, the mean rainfall $C_{N-1,2}$ (in $N-1$ scale level) and $D_{N,1}$ (in *N* scale level) can be obtained by Eqs. (11) and (12). The weighted value $v_{N,1}(i)$ is obtained by the method described in APPENDIX A. Together with the transition probabilities $T_{k,n}(i,j)$ estimated by the EM algorithm using the previous years' daily rainfall data, all the hidden state probabilities can be obtained (APPENDIX B). Using the hidden state probabilities, the remaining values $D_{k,n}$ (other than $D_{N,1}$) are simulated by the Monte Carlo method.

4.1 *N*-order scale data

The *N*-order scale data $C_{N,1}$ can be estimated by using the historical daily rainfall data. For a given prediction origin t_0, we identify the date corresponding to t_0 in the y^{th} year as t_0^y . For example, if $t_0 = 1000$, we set $t_0^1 = 270$ and $t_0^2 = 635$, then decompose the (signals of) daily data in the 2^N days periods $[t_0^y - 2^{N-1} + 1, t_0^y + 2^{N-1}]$ into sub-signals (wavelet tree). If u_0^y and u_{-1}^y respectively denote the mean rainfalls for the y^{th} year for the periods $[t_0^y - 2^{N-1} + 1, t_0^y + 2^{N-1}]$ and $[t_0^y - 3\times 2^{N-1} + 1, t_0^y - 2^{N-1}]$, then by the determinism of the *N*-order scale data, it can be assumed that u_0^y and u_{-1}^y satisfy an evolutionary equation of the form

$$u_0 = h(u_{-1}) \tag{18}$$

where h denotes the evolutionary function.

In this study the function h is assumed to be linear of the form

$$u_0 = w_0 + w_1 u_{-1} + \varepsilon \tag{19}$$

where the parameters w_0 and w_1 are estimated by the least squares method by minimizing

$$\sum_{y=1}^{Y} (u_0^y - w_0 - w_1 u_{-1}^y)^2 \,.$$

as,

$$\hat{w}_1 = \frac{\sum_{y=1}^{Y} (u_{-1}^y - \bar{u}_{-1})(u_0^y - \bar{u}_0)}{\sum_{y=1}^{Y} (u_{-1}^y - \bar{u}_{-1})^2}, \tag{20}$$

$$\hat{w}_0 = \bar{u}_0 - \hat{w}_1 \bar{u}_{-1}, \tag{21}$$

where $\bar{u}_{-l} = \frac{1}{Y}\sum_{y=1}^{Y} u_{-l}^y$, for $l = 0, 1$.

Once the coefficients $\hat{w}_0$ and $\hat{w}_1$ are known, the mean rainfall data u_0 for the period $[t_0 - 2^{N-1} + 1, t_0 + 2^{N-1}]$ can be estimated by $u_0 = \hat{w}_0 + \hat{w}_1 u_{-1}$ (here u_{-1} will be the mean rainfall for the period $[t_0 - 3\times 2^{N-1} + 1, t_0 - 2^{N-1}]$). Since the mean rainfall for the period $[t_0 - 2^{N-1} + 1, t_0]$ denoted by $C_{N-1,1}$ is known, the $(N-1)$-order scale data $C_{N-1,2}$ and N-order wavelet data $D_{N,1}$ can be obtained from Eqs. (11) and (12).

The large-scale simulation by using this simple linear model is shown in Figure 3, and the statistics of the linear model simulation are listed in Table 2.

4.2 Weighted value $v_{k,n}(i)$

The weighted value $v_{k,n}(i)$ for the node $(N,1)$ in the wavelet tree is first estimated using the wavelet value $D_{N,1}$. The transition probabilities for the wavelet tree are estimated under the assumption that different wavelet trees by the rainfall data for the same time period in different years have the same transition probabilities (i.e. for any y, $T_{k-1,2n-l}(i,j) = \Pr(S^y_{k-1,2n-l} = i \mid S^y_{k,n} = j)$ for $l = 0,1$). If the 0-order scale data for the y^{th} year ($1 \le y \le Y$) is $\{C^y_{0,1}, C^y_{0,2}, \ldots, C^y_{0,2^N}\}$, then the Y wavelet trees are constructed by using these 0-order scale data as $D^y_{k,n}$ for $1 \le k \le N$, $1 \le n \le 2^{N-k}$ and $1 \le y \le Y$. All the wavelet trees have the same transition probabilities $T_{k,n}(i,j)$ but different weighted values $v^y_{k,n}(i)$, which could be estimated by the EM algorithm (APPENDIX B). After estimating the weighted value $v_{N,1}(i)$, the other weighted values $v_{k,n}(i)$ are given iteratively by

$$v_{k-1,2n-1}(i) = \sum_{j=1}^{M} T_{k-1,2n-1}(i,j) v_{k,n}(j) \tag{22}$$

$$v_{k-1,2n}(i) = \sum_{j=1}^{M} T_{k-1,2n}(i,j) v_{k,n}(j) \tag{23}$$

for $k = N, N-1, \ldots, 2$; $n = 1, 2, \ldots, 2^{N-k}$ and $i = 1, 2, \ldots, M$.

4.3 Wavelet value $D_{k,n}$ and daily rainfall data

The Monte Carlo method is used to simulate the wavelet coefficients $D_{k,n}$ using the weighted value $v_{k,n}(i)$. For a pair of (k,n), with $k = 1, 2, \ldots, N-1$ and $n = 1, 2, \ldots, 2^{N-k}$ if

$$\delta_{k,n}(i) = \sum_{j=1}^{i} v_{k,n}(j), \quad 1 \le i \le M \tag{24}$$

then, we have

$$0 = \delta_{k,n}(0) \le \delta_{k,n}(1) \le \delta_{k,n}(2) \le \cdots \le \delta_{k,n}(M) = 1. \tag{25}$$

For two random numbers, γ_1 and γ_2 chosen from $[0,1]$, the value of the wavelet coefficient $D_{k,n}$ at node (k,n) is simulated as

$$\begin{aligned} \gamma_2 &= \int_{-\infty}^{D_{k,n}} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{{D_{k,n}}^2}{2\sigma_i^2} \right) dx \\ &= \frac{1}{2} + \int_0^{D_{k,n}} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{{D_{k,n}}^2}{2\sigma_i^2} \right) dx \\ &= \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{D_{k,n}}{\sqrt{2\sigma_i}} \right) \end{aligned} \tag{26}$$

Here, $\operatorname{erf}(x)$ is the error function and is defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt \tag{28}$$

and $\operatorname{erfinv}(x)$ is its inverse – the inverse error function.

The random number $D_{k,n}$ is defined as

$$D_{k,n}^m = \sigma_i\left(\sqrt{2}\operatorname{erfinv}(2\gamma_2^m - 1)\right), \text{subject to}\, \gamma_1 \in [\delta_{k,n}(i-1), \delta_{k,n}(i-1)]\,, \tag{27}$$

and its expected value is defined as

$$D_{j,k} = \frac{1}{NUM}\sum_{m=1}^{NUM} D_{k,n}^m$$

This is equivalent to choosing two random numbers x and y, where x is normally distributed with zero mean and unit variance, and y is uniformly distributed in the range $[0,1]$, then $D_{k,n}$ is simulated as $D_{k,n} = \sigma_i x$, if $y \in [\delta_{k,n}(i-1), \delta_{k,n}(i)]$ (by Monte Carlo method, the formula $\sqrt{2}\operatorname{erfinv}(2\gamma_2 - 1)$ is indeed a simulation of standard normal distribution with zero mean and unit variance). The command ERFINV in MATLAB can be used for executing Eq. 27.

The daily rainfall data, $C_{0,n}$ for $t_0 < n \le t_0 + 2^{N-1}$, can then be obtained using the wavelet tree and the scale data $C_{N-1,2}$ of equation (12).

APPLICATION

The proposed method is applied to three rainfall data sets from the Chao Phraya River Basin in Thailand. As mentioned earlier, some parameters needed to be determined *a priori*. They include the number of layers N in the wavelet tree, the number of hidden states M and the variances σ_i for each hidden state.

The number of layers (or, the scale) in the wavelet tree, N, is determined using the False Nearest Neighbours (FNN) method that has been proposed for finding the embedding dimension d_e of a deterministic system (Abarbanel, 1996; Jayawardena *et al.*, 2002). In this study, we extend the same concept to determine the best scale order N which will ensure that the data in the *N*-order scale are deterministic. In other words, the FNN method is used as a test for determinism of a time series. It should be mentioned that, N is chosen to be the minimum, so that the data of scale less than N are considered as stochastic, and therefore the wavelet tree coefficients are stochastic.

In this study, we use the FNN method to detect determinism for mean rainfall for different scales by fixing the time delay and comparing the percentage of FNN at different embedding dimensions. The time delay T is fixed as unity and the best embedding dimension, by trial and error, has been found to be 3. The analysis shows (Figure 4) that when the scale is 6, almost all the points in the reconstructed phase space have no False Nearest Neighbours (FNN), and therefore the mean rainfall series with scale 6 can be considered as a deterministic series.

The second parameter to be assigned *a priori* is the number of hidden states which has been set at 3. The third is the variances. Since all the wavelet signals lie in the interval $(-1, 1)$, their variances will be within the range $(0, 1)$. Therefore, they are assigned the values $\sigma_1 = 0.1$, $\sigma_2 = 0.4$, $\sigma_3 = 0.7$ representing large, medium and small hidden states.

The simulation procedure for fixed *N*, *M* and prediction origin *t*$_0$, involves four steps: The first step is to estimate $C_{N,1} = C_{6,1} = u_0$ - the mean of rainfall data in the interval $[t_0 - 2^{6-1} + 1, t_0 + 2^{6-1}]$ by the linear model. We determine the mean of corresponding 2^6 days in the previous years, i.e. the mean u_0^y in the interval $[t_0^y - 2^{6-1} + 1, t_0^y + 2^{6-1}]$, and the mean

u_{-1}^{y} in the interval $[t_0^y - 3\times 2^{6-1} + 1, t_0^y - 2^{6-1}]$, for $1 \le y \le Y$. By using the linear model, u_0 is simulated. Since the rainfall data of the period $[t_0 - 2^{6-1} + 1, t_0]$ is known, $C_{5,2}$ is obtained and thus $D_{6,1}$ is evaluated by Eq. (12). The second step is to estimate the weighted value $v_{6,1}(i)$ for $1 \le i \le 3$ using $D_{6,1}$ obtained in step one and the algorithm in APPENDIX A. The third step is to estimate the transition probabilities $T_{k,n}(i, j)$ ($1 \le k \le 5$, $1 \le n \le 2^{6-k}$ and $1 \le i, j \le 3$) of the wavelet trees constructed by the historical data of the period $[t_0^y - 2^{6-1} + 1, t_0^y + 2^{6-1}]$. The last step is to simulate the wavelet coefficients $D_{k,n}$ (other than $D_{6,1}$) by the Monte Carlo method. The daily rainfall data for the 2^5 days following the prediction origin t_0 can then be calculated by inverse transformation.

RESULTS

The results of daily simulation of the rainfall data in the Chao Phraya Basin are shown in Figure 5 for different origins of prediction. Since the model that has been used is a mixed one, it has inherently some randomness built into it. Therefore a deterministic comparison alone is not expected to give a one to one match. Nevertheless, the direct comparisons show some level of agreement as can be seen in the dry period of the simulated and observed time series.

On the other hand, as illustrated in Figure 6, comparisons of the observed and simulated rainfall accumulation seem to give a better interpretation. It can be seen that the accumulated rainfall of the simulated one is close to the observed one in between the scale from 1 to 16. This observation is anticipated, since the mean of 2^6 days rainfall data is of deterministic nature, the mean of simulated rainfall of a number of days less than 64 should be close to the mean of original rainfall, and because of the stochastic nature of the model the period that simulated mean (equivalently accumulated rainfall) equals to original one is varying.

Another point to be mentioned is that the accumulation of 32 days rainfall data does not always give a good prediction. This is mainly due to the error in simulating the large-scale data by the linear model. A small error ε in the large-scale simulation could result in a large error, namely $2^6 \times \varepsilon$, in the large scale. It is also known that there are only limited successes of deterministic modelling of rainfall. It seems that some other models could also be employed in this situation;

however, the estimation of parameters may lead to difficulties.

The results are summarized in Table 3.

7. CONCLUSION

A novel approach for the simulation of daily rainfall data has been presented. The model is motivated by the notion that in some regions, large-scale rainfall data exhibit a deterministic character while small-scale data exhibit strong stochastic character. In this study, we aimed at simulating the rainfall data sets with such properties, thus requiring many assumptions for different modelling levels. A brief account of these assumptions is summarized in Table 4.

Although the idea of the model is quite clear, the methodology is rather complex, thereby imposing several limitations:

1) Linear model

In the study, the FNN method is employed to find the scale in which the underlying dynamical system is deterministic. We are able to explore the original system by using the simple transformed phase space. However, the original dynamic is still unknown. In other words, the system $\mathbf{x}(t) \to F(\mathbf{x}(t)) = \mathbf{x}(t+1)$ remains unresolved and the underlying dynamics could be very complicated. There is no special theory to justify the assumption that the linear model used in this study could fit into the large-scale data. However it gives an acceptable primary simulation (see Figure 3). In fact, some statistical methods or some deterministic methods can also be used to predict the monthly rainfall accumulation. Here we employ a simple deterministic model – the linear model – to do it. For different rainfall series at different regions, different methods can be used for prediction purpose. For example, we may use a similar 3-parameter linear model, $u_0 = w_0 + w_1 u_{-1} + w_2 u_{-2}$, to do the simulation. The results are shown in Figure 7, and summarized in Table 5. It gives a very similar estimation as the 2-parameter linear model in this study.

Many fitting tools, for example, Fourier series and polynomial interpolation are widely used in the analysis of periodic time series. However such methods are not suitable in the present case. For example, if we have historical rainfall data up to the date t_0, and we need to simulate the daily rainfall starting from t_0 using the proposed model, the mean of 2^N days data centered at t_0 would be necessary; i.e. the rainfall of the period $[t_0 - 2^{N-1} + 1, t_0 + 2^{N-1}]$, which include a prediction of 2^{N-1} days rainfall data (for the period $[t_0 + 1, t_0 + 2^{N-1}]$). When this N is small (e.g. 1, 2), such fitting tools may be suitable, however it will give unpredictable error when N is large (e.g. 6).

2) Transition probabilities

We have assumed that the rainfall data has same statistical properties in the same time region in different years. This assumption is crucial for simulating the transition probabilities by EM algorithm (provided in APPENDIX B).

3) Negative values in large-scale predictions

Some large deviations or negative values are predicted using the 2-parameter (see Figure 3(c, e)) and 3-parameter (see Figure 7(a, b, c, e)) linear model. These errors were essentially due to insufficient quantity of data for the estimation of parameters $\hat{w}_i$ (Eqs. (20), (21)). Our large-scale simulation, ranging from $t_0 = 4349$ to $t_0 = 5078$, can only provide 12 or 13 data points for the least-square fitting. Therefore, in some cases, the least square method may not give a precise estimation for the parameters $\hat{w}_i$, and thus resulted in large error and causing the large-scale prediction to be negative.

8. ACKNOWLEDGEMENTS

The initial part of the work presented in this paper was carried out while the first author was in the Department of Civil Engineering of the University of Hong Kong and the second author was visiting.

9. REFERENCES

Abarbanel, H. D. I., 1996. *Analysis of observed chaotic data*. New York: Springer-Verlag, 1996.

Aksoy, H., and Bayazit, M. (2000): *A model for daily flows of intermittent streams.* Hydrological Processes 14, 1725-1744.

Aksoy, H. and Unal, N.E. (2007): *Discussion of 'Comparison of two nonparametric alternatives for stochastic generation of monthly rainfall'.* ASCE Journal of Hydrologic Engineering 12, 699-702.

Bayazit, M. and Aksoy, H. (2001): *Using wavelets for data generation*. Journal of Applied Statistics 28, 157-166.

Bayazit, M., Onoz, B. and Aksoy, H. (2001): *Nonparametric streamflow simulation by wavelet or Fourier analysis*. Hydrological Sciences Journal 46, 623-634.

Chapman, T. G., 1994. *Stochastic models of daily rainfall.* Proceedings of Water Down Under 94, National Conference Publications, Institution of Engineers, Canberra, ACT, Australia, vol. 3, pp 7-12.

Chipman, H. A., Kolaczyk, E. D., McCulloch, R. E., 1997. *Adaptive Bayesian wavelet shrinkage.* J. Amer. Stat. Assoc. 92 (1997) 1413-1421.

Chow, C. K., Liu, C. N., 1968. *Approximating discrete probability distributions with dependence trees.* IEEE Trans. Inform. Theory 14 (1968) 462-467.

Coulibaly, P., Anctil, F., Bobee, B., 2000. *Daily reservoir inflow forecasting using artificial neural networks with stopped training approach*. Journal of Hydrology 230, 244–257.

Crouse, M. S., Nowak, R. D., Baraniuk, R. G., 1998. *Wavelet-based statistical signal processing using hidden Markov models.* IEEE Transactions on Signal Processing 46 (1998) 886-902.

Daubechies, I., 1992. *Ten lectures on wavelets.* New York: SIAM, 1992.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. *Maximum likelihood from incomplete data via the EM algorithm.* J. Royal Stat. Soc. B. 39 (1977) 1-38.

Haar A. (1910): *Zur Theorie der orthogonalen Funktionensysteme*, Mathematische Annalen, 69, pp 331-371.

Hamilton, J. D., 1994. *Time series analysis.* Princeton: Princeton University Press, 1994.

Jayawardena, A. W., Li, W. K., Xu, P., 2002. *Neighbourhood selection for local modelling and prediction of hydrological time series.* J. Hydrol. 258 (2002) 40-57.

Labat, D., Ababou, R., Mangin, A., 1999. *Wavelet analysis in Karstic hydrology 2nd Part:*

rainfall–runoff cross-wavelet analysis. Comptes Rendus de l'Academie des Sciences Series IIA Earth and Planetary Science 329, 881–887.

Mallat, S., 1997. *A wavelet tour of signal processing.* Academic Press, 1997.

McLachlan, G. J., Krishnan, T., 1997. *The EM algorithm and extensions.* New York: John Wiley, 1997.

Rabiner, L. R., 1989. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proc. IEEE 77 (1989) 257-286.

Ronen, O., Rohlicek, J. R., Ostendorf, M., 1995. *Parameter estimation of dependence tree models using the EM algorithm.* IEEE Signal Proc. Lett. 2 (1995) 157-159.

Saco, P., Kumar, P., 2000. *Coherent modes in multiscale variability of streamflow over the United States.* Water Resources Research 36, 1049–1068.

Salas, J. D., 1993. *Analysis and modelling of hydrologic time series.* In: Handbook of Hydrology (edited by David R. Maidment; New York: McGraw-Hill, 1993) 19.1-72.

Sharma, A., Lall, U., 1999. *A nonparametric approach for daily rainfall simulation.* Math. and Comp. in Simulation 48 (1999) 361-371.

Smith, L.C., Turcotte, D.L., Isacks, B.L., 1998. *Stream flow characterization and feature detection using a discrete wavelet transform*. Hydrological Processes 12, 233–249.

Smyth, P., Hecherman, D., Jordan, M. I., 1997. *Probabilistic independence networks for hidden Markov probability models.* Neural comp. 9 (1997) 227-269.

Sobol', I. M., 1994. *A primer for the Monte Carlo method.* CRC Press, 1994.

Unal, N.E., Aksoy, H. and Akar, T. (2004): *Annual and monthly rainfall data generation schemes*. Stochastic Environmental Research and Risk Assessment 18, 245-257.

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

Wong, C. S., Li, W. K., 2000. *On a mixture autoregressive model.* J. Royal Stat. Soc. B. 62 (2000) 91-115.

APPENDIX A

The algorithm provided below is used to estimate the parameters in a mixture of Gaussian distributions. It turns out to be a special case of the EM algorithm developed by Dempster, Laird, and Rubin (1997). Quick reference could also be made to the book by Hamilton (1994).

Let $v_{N,1}^{t}(i)$ be an arbitrary initial guess of $v_{N,1}(i)$ under the probability rule $\sum_{i=1}^{M} v_{N,1}^{t}(i) = 1$ (Here the superscript t is an iteration counter). For $i = 1, 2, \ldots, M$, we let

$$\alpha_{N,1}(i) = v_{N,1}^{t}(i) \tag{32}$$

and

$$\beta_{N,1}(i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{D_{N,1}^2}{2\sigma_i^2} \right). \tag{33}$$

Then a new estimation of the probability $v_{N,1}(i)$ is obtained as

$$v_{N,1}^{t+1}(i) = \frac{\alpha_{N,1}(i)\beta_{N,1}(i)}{\sum_{j=1}^{M} \alpha_{N,1}(j)\beta_{N,1}(j)} \tag{34}$$

for $i = 1, 2, \ldots, M$. The iteration stops when the convergence criterion

$$\sum_{i=1}^{M} | v_{N,1}^{t+1}(i) - v_{N,1}^{t}(i) | < \varepsilon \tag{35}$$

($\varepsilon = 1.0E-6$ in this study) is satisfied. Thus the estimation of the probability $v_{N,1}(i)$ can be obtained.

APPENDIX B

The EM algorithm for estimating the transition probabilities closely follows the paper by Crouse *et al.* (1998). The EM algorithm for dependent tree models first appeared in Chow and Liu (1968). Ronen *et al.* (1995) expanded the work to include the condition that some components of the tree are unobserved. However, the EM steps have been derived only for discrete-valued random variables. Crouse *et al.* (1998) generalized the algorithm so that it can be applied to the hidden Markov tree models, which are of discrete and continuous valued nodes. We slightly modified the algorithm in Crouse's paper, so that some of the parameters (means and variances) are fixed. The steps of the EM algorithm for this model are as follows:

Initialization

Given arbitrary initial assignments of the transition probabilities $T^t_{k,n}(i,j)$ and hidden state probabilities $v^{y,t}_{k,n}(i)$ for different wavelet trees (Here the superscript t is an iteration counter):

$T^t_{k,n}(i,j)$: for $1 \le i, j \le M$, $k = 1,\ldots,N-1$, and $n = 1,\ldots,2^{N-k}$.

$v^{y,t}_{k,n}(i)$: for $1 \le i \le M$, $k = 1,\ldots,N$, $n = 1,\ldots,2^{N-k}$ and $y = 1,\ldots,Y$.

under the probability rules: $\sum_{i=1}^{M} v^{y,t}_{k,n}(i) = 1$ and $\sum_{i=1}^{M} T^t_{k,n}(i,j) = 1$ for any choice of j (since for any fixed j, the state variable $S^y_{k,n}$ takes a value from $\{1,\ldots,M\}$).

Expectation step (E-step)

For each wavelet tree, that is for $y = 1,\ldots,Y$, we apply the "upward-downward" algorithm:

A. The upward algorithm.

1) Initialization: assign the values of β at the "leaves" of the wavelet trees. For $i = 1, 2,\ldots, M$ and $n = 2^{N-2},\ldots,2^{N-1}$, let

$$\beta^y_{1,n}(i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(D^y_{1,n})^2}{2\sigma_i^2} \right). \tag{36}$$

2) Step upward: calculate all the values of β by the following formulas. For $k = 2,\ldots,N$, $n = 2^{N-2} \ldots,2^{N-1}$, $l = 0,1$ and $i = 1,\ldots,M$,

$$\phi^y_{k-1,2n-l}(i) = \sum_{j=1}^{M} T^t_{k-1,2n-l}(j,i)\beta^y_{k-1,2n-l}(j), \tag{37}$$

$$\beta^y_{k,n}(i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(D^y_{k,n})^2}{2\sigma_i^2}\right)\prod_{l=0}^{1}\phi^y_{k-1,2n-l}(i)\text{, and let} \tag{38}$$

$$\varphi^y_{k-1,2n-l}(i) = \frac{\beta^y_{k,n}(i)}{\phi^y_{k-1,2n-l}(i)}. \tag{39}$$

B. The downward algorithm.

1) Initialization: To get the values of α at the “root” of the wavelet trees, we set for $i = 1, 2, \ldots, M$

$$\alpha^y_{N,1}(i) = v^{y,t}_{N,1}(i). \tag{40}$$

2) Step downward: obtain the remaining values of α . For $k = N, \ldots, 2$, $n = 2^{N-k-1}, \ldots, 2^{N-1}$, $l = 0, 1$ and $i = 1, 2, \ldots, M$,

$$\alpha^y_{k-1,2n-l}(i) = \sum_{j=1}^{M} T^t_{k-1,2n-l}(i, j)\alpha^y_{k,n}(j)\varphi^y_{k-1,2n-l}(j). \tag{41}$$

Maximization step (M-step)

A. Update the state probabilities in the wavelet trees.

For $k = 1, \ldots, N$, $n = 2^{N-k-1}, \ldots, 2^{N-1}$ and $i = 1, \ldots, M$, the new iteration is

$$v^{y,t+1}_{k,n}(i) = \frac{\alpha^y_{k,n}(i)\beta^y_{k,n}(i)}{\sum_{j=1}^{M}\alpha^y_{k,n}(j)\beta^y_{k,n}(j)}. \tag{42}$$

B. Renew the transition probabilities.

For $k = 2, \ldots, N$, $n = 2^{N-k-1}, \ldots, 2^{N-1}$, $l = 0, 1$ and $i = 1, \ldots, M$, the following simplified equation (to facilitate computations) is used:

$$T^{t+1}_{k-1,2n-l}(i, j) = \frac{1}{Y}\sum_{y=1}^{Y}\frac{T^t_{k-1,2n-l}(i, j)\beta^y_{k-1,2n-l}(i)}{\phi^y_{k-1,2n-l}(j)}. \tag{43}$$

Convergence checking

- For $k = 1, \ldots, N-1$, $n = 2^{N-k-1}, \ldots, 2^{N-1}$ and $1 \le i, j \le M$, set

$$\varepsilon_1 = \max(| T^{t+1}_{k,n}(i, j) - T^t_{k,n}(i, j) |). \tag{44}$$

- For $y = 1, \ldots, Y$, $k = 1, \ldots, N$, $n = 2^{N-k-1}, \ldots, 2^{N-1}$ and $1 \le i \le M$, set

$$\varepsilon_2 = \max(| v^{y,t+1}_{k,n}(i) - v^{y,t}_{k,n}(i) |). \tag{45}$$

- Set $\varepsilon = \max(\varepsilon_1, \varepsilon_2)$.

If $\varepsilon < 1.0E - 6$ (convergence criterion), then we STOP the algorithm and set all

$$T_{k,n}(i, j) = T_{k,n}^{t+1}(i, j) . \tag{46}$$

Otherwise, we need to set $v_{k,n}^{y,t}(i) = v_{k,n}^{y,t+1}(i)$ and $T_{k,n}^{t}(i, j) = T_{k,n}^{t+1}(i, j)$, and do the EM algorithm steps again until the convergence criterion is met.

TABLES

Table 1. Data summary for the Chao Phraya

Regions	Gauging station	Number of data points	Dry probability	Average annual rainfall (mm)
Chao Phraya Basin	No. 111 (CPB111)	5110	0.7102	1052.15
	No. 112 (CPB112)	5232	0.8249	868.60
	No. 117 (CPB117)	5232	0.8562	1007.38

Table 2. Data summary for the large-scale simulation.

Gauging station	Mean		Standard deviation	
	Observed	Simulated	Observed	Simulated
CPB111	2.58	2.87	2.24	2.14
CPB112	2.04	2.38	1.80	1.79
CPB117	2.36	2.68	3.08	2.02

Table 3. Mean and standard deviation of daily rainfall for the simulated and observed data.

Gauging station	Prediction origin	Mean		Standard deviation	
		Observed	Simulated	Observed	Simulated
CPB111	$t_0 = 4800$	2.14	5.41	5.90	11.40
	$t_0 = 4850$	4.67	8.48	10.65	7.70
	$t_0 = 4900$	11.05	10.83	16.78	22.10
CPB112	$t_0 = 4800$	3.37	5.76	11.30	12.05
	$t_0 = 4850$	2.82	7.26	9.53	8.91
	$t_0 = 4900$	8.72	9.18	15.41	17.10
CPB117	$t_0 = 4800$	0.45	6.69	4.07	13.27
	$t_0 = 4850$	2.24	5.18	8.21	6.83
	$t_0 = 4900$	3.32	8.09	9.73	8.94

Table 4. Summary of assumptions of the proposed simulation method.

Modelling assumptions	Parameter assumptions
Monthly rainfall data are deterministic	$N = 6$
Linear regression model is used for the deterministic system	$M = 3$
Wavelet coefficients $D_{k,n}$ are assumed to follow mixture Gaussian distribution and are stochastic	$\sigma_1 = 0.1$ $\sigma_2 = 0.4$ $\sigma_3 = 0.7$
Transition probabilities for the same period of time in different years are the same	

Table 5. Data summary for the large-scale simulation by the linear model with 3 parameters.

Gauging station	Mean		Standard deviation	
	Observed	Simulated	Observed	Simulated
CPB111	2.58	2.75	2.24	2.12
CPB112	2.04	2.43	1.80	1.78
CPB117	2.36	2.43	3.08	1.91

FIGURE CAPTIONS

Figure 1. Wavelet tree.

Figure 2. Part of the wavelet tree.

Figure 3. Large-scale simulation using 2-parameter linear function: from $t_0 = 4349$ (Day 1) to $t_0 = 5078$ (Day 730) of (a) CPB111, (b) CPB112, (c) CPB117

Figure 4. Percentage of FNN for the rainfall data with embedding dimension (a) $d_e = 1$; (b) $d_e = 2$; (c) $d_e = 3$. (These figures are not as identical to the percentage-dimension representations encountered in standard nonlinear time series literatures (e.g. Abarbanel, 1996). The percentage-scale graphs are provided instead, in order to illustrate the embedding dimension for the new time series constructed by Eq. (10))

Figure 5. Daily rainfall simulation of (a) CPB111; (b) CPB112; (c) CPB117 at different prediction origins t_0.

Figure 6. Rainfall accumulation for the data sets (a) CPB111; (b) CPB112; (c) CPB117 at different prediction origins t_0.

Figure 7. The large-scale simulation using 3-parameter linear function: from $t_0 = 4349$ (Day 1) to $t_0 = 5078$ (Day 730) of (a) CPB111, (b) CPB112, (c) CPB117

FIGURES

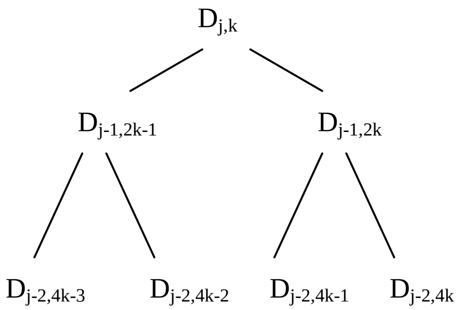


Figure 1a

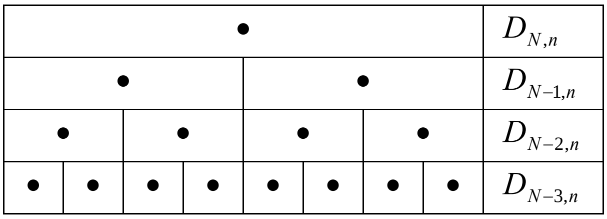


Figure 1b.

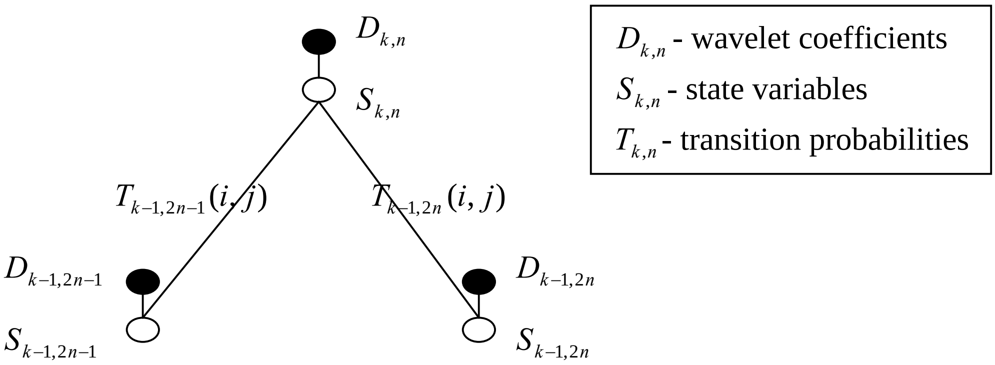


Figure 2.

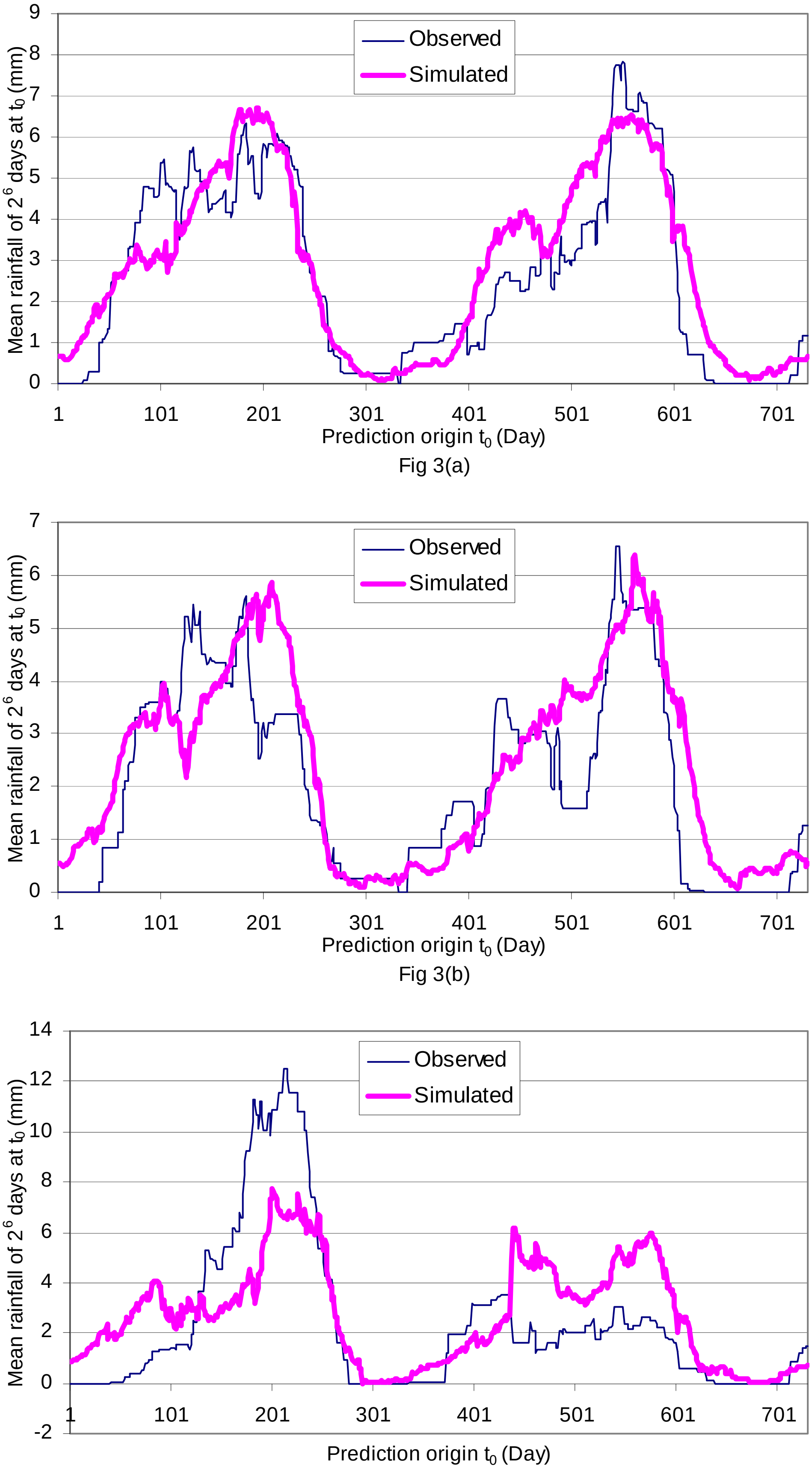


Fig 3(a)

Fig 3(b)

Fig 3(c)

Figure 3.

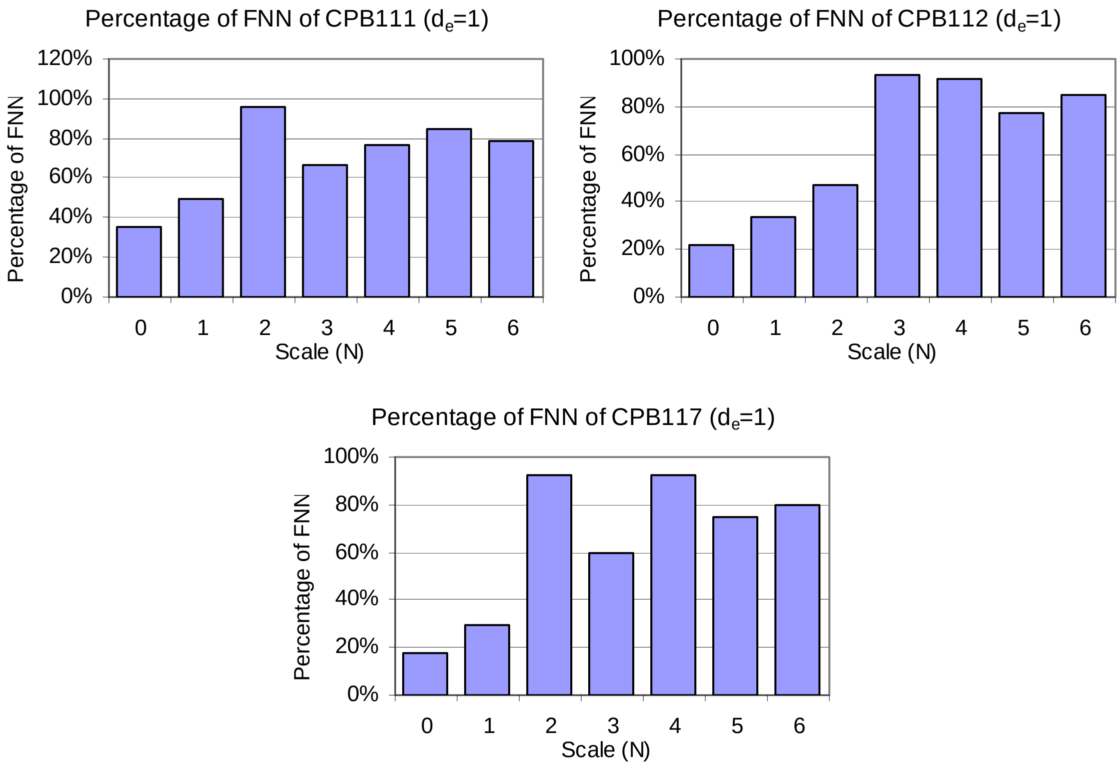
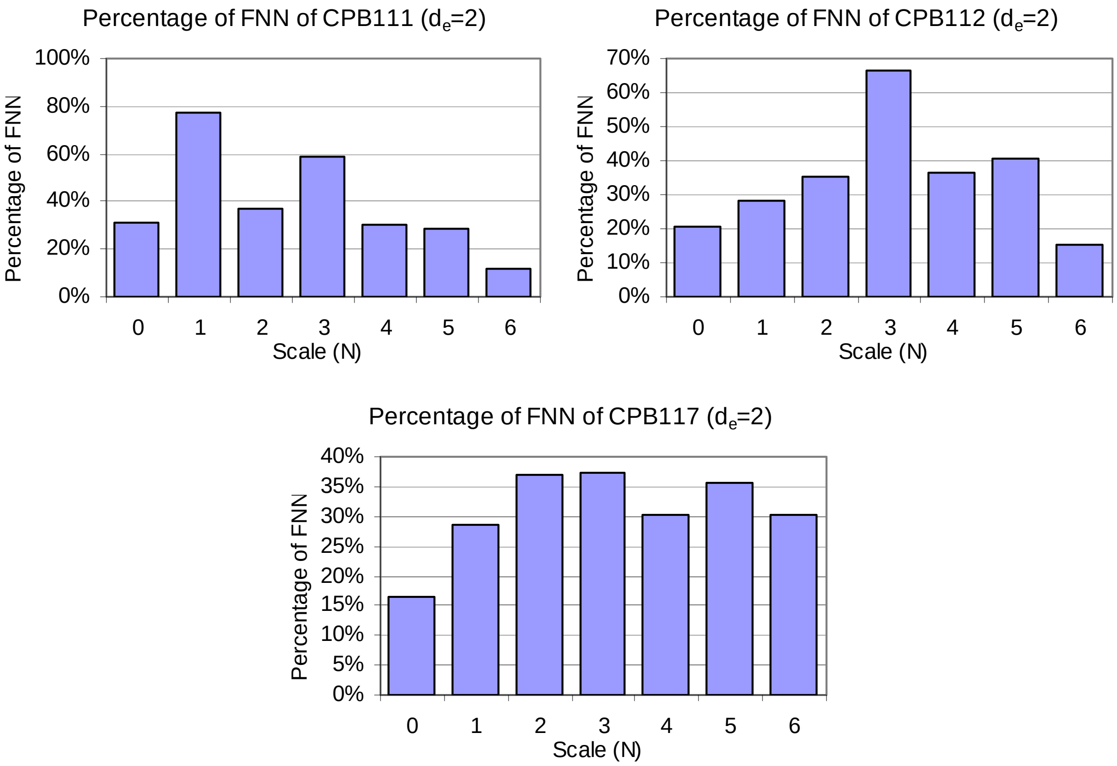

Percentage of FNN of CPB111 (d_e=1)
Percentage of FNN
120%
100%
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
Scale (N)
Percentage of FNN of CPB112 (d_e=1)
Percentage of FNN
100%
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
Scale (N)
Percentage of FNN of CPB117 (d_e=1)
Percentage of FNN
100%
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
Scale (N)

Figure 4(a).

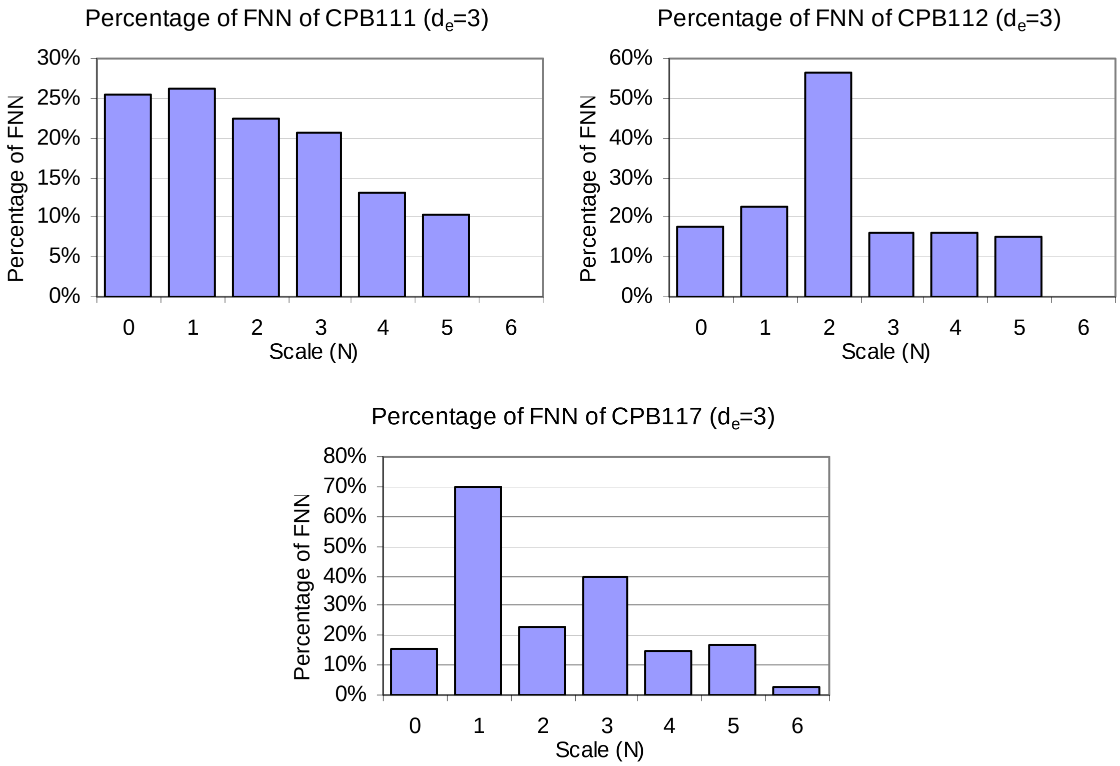


Percentage of FNN of CPB111 (d_e=2)

Percentage of FNN of CPB112 (d_e=2)

Percentage of FNN of CPB117 (d_e=2)

Figure 4(b).



Percentage of FNN of CPB111 (d_e=3)

Percentage of FNN of CPB112 (d_e=3)

Percentage of FNN of CPB117 (d_e=3)

Figure 4(c).

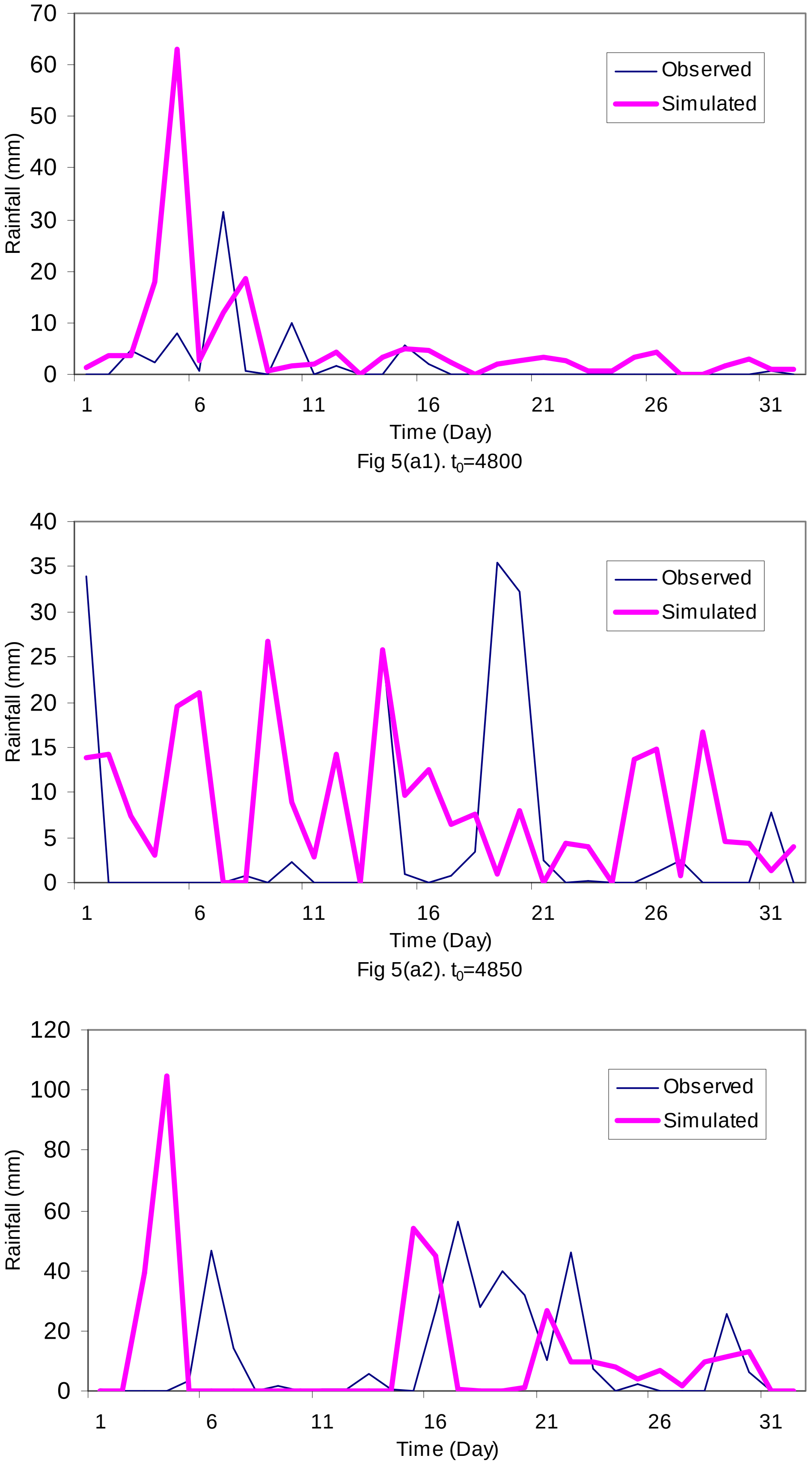


Fig 5(a1). t_0=4800

Fig 5(a2). t_0=4850

Fig 5(a3). t_0=4900

Figure 5(a).

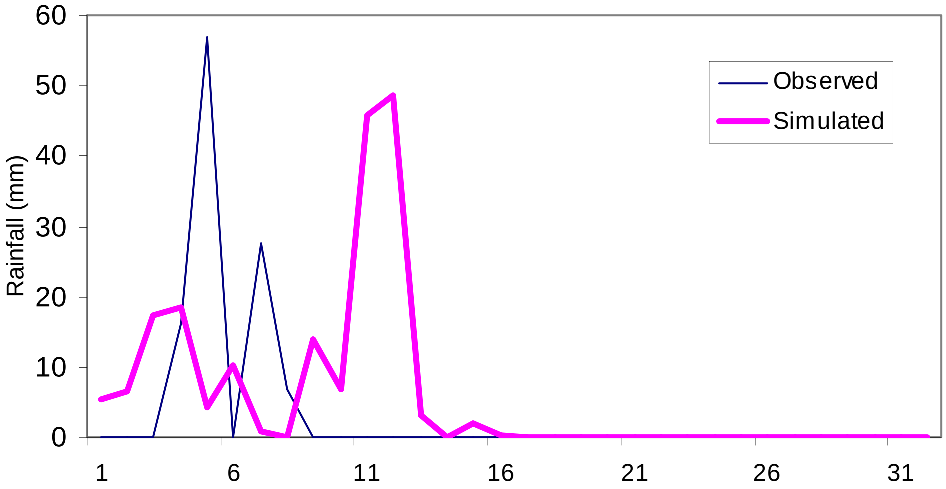


Fig 5(b1). t_0=4800

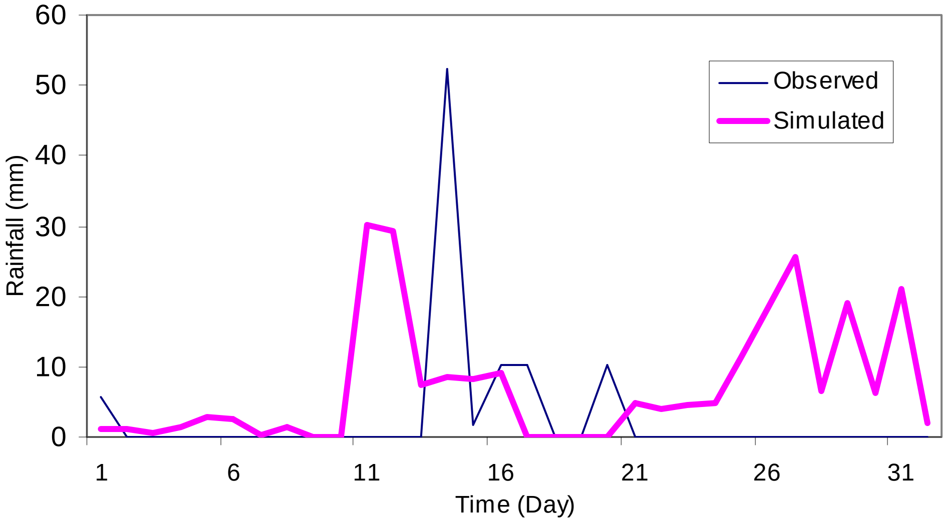


Fig 5(b2). t_0=4850

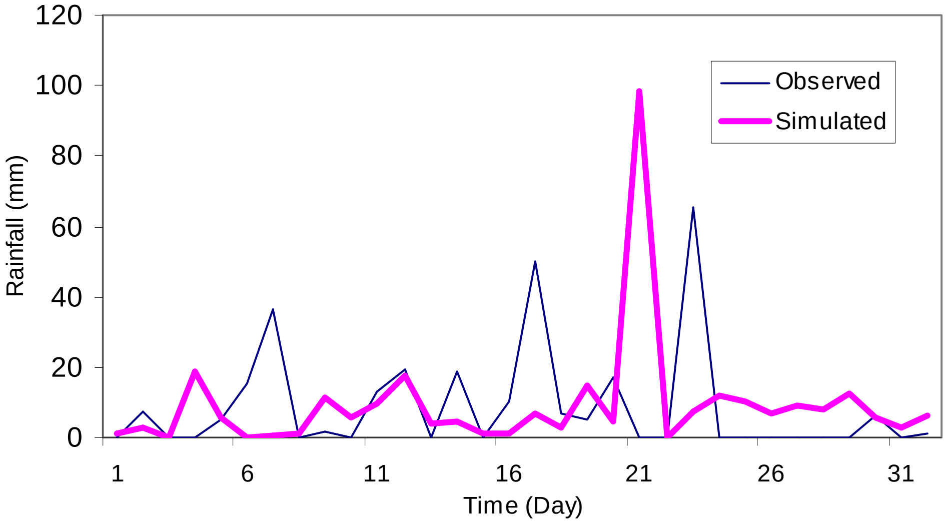


Fig 5(b3). t_0=4900

Figure 5(b).

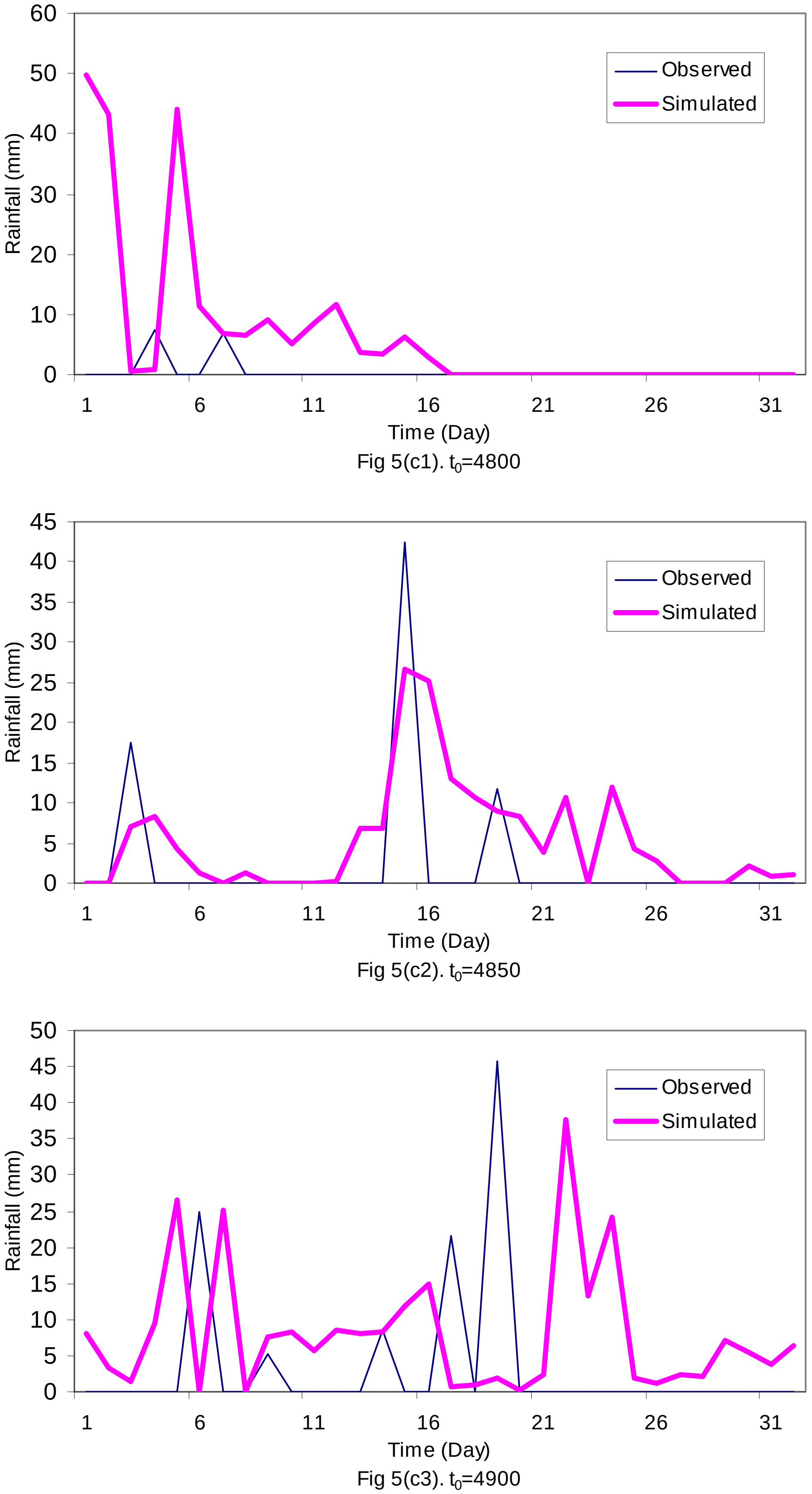


Fig 5(c1). t_0=4800

Fig 5(c2). t_0=4850

Fig 5(c3). t_0=4900

Figure 5(c).

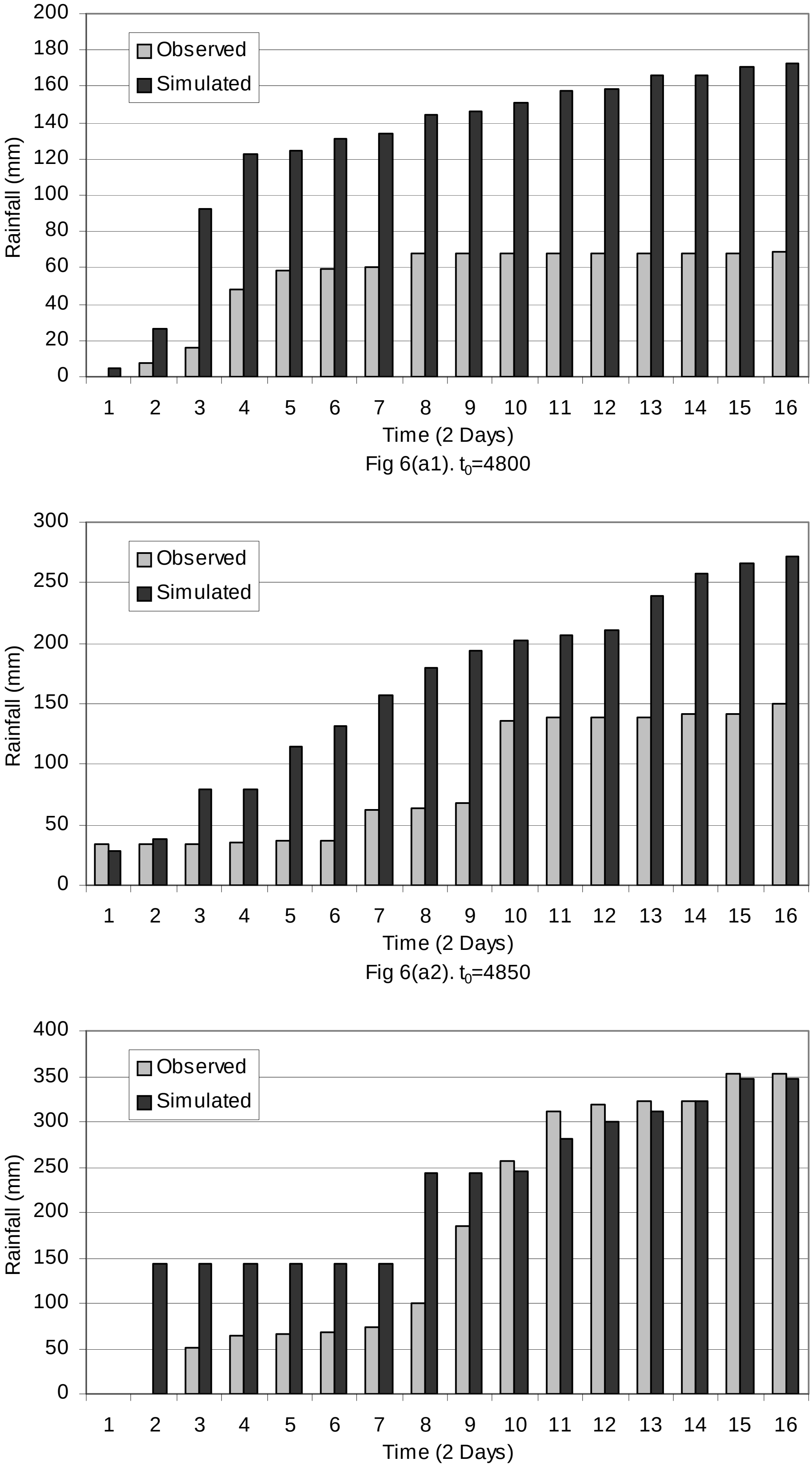


Fig 6(a1). t_0=4800

Fig 6(a2). t_0=4850

Fig 6(a3). t_0=4900

Figure 6(a).

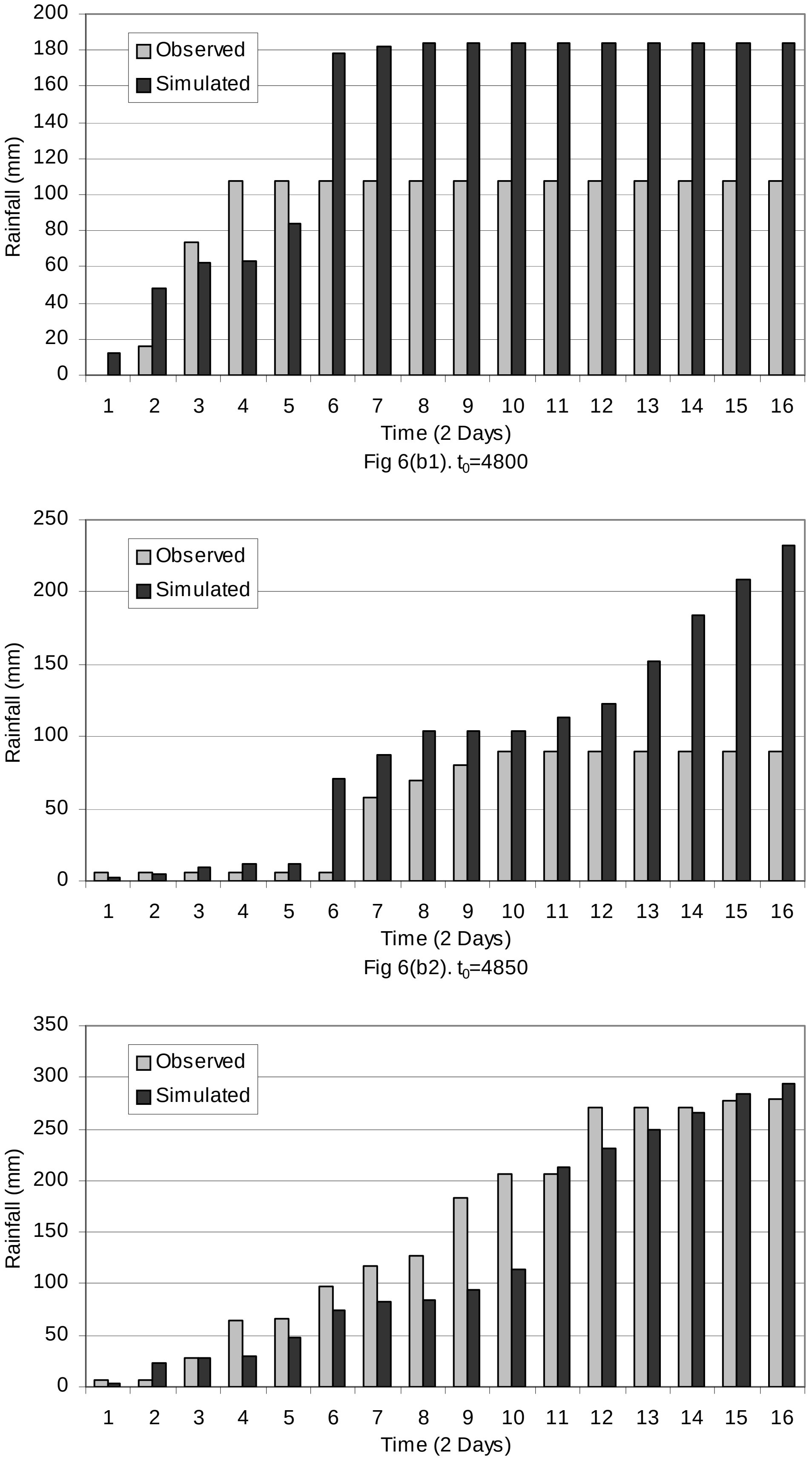


Fig 6(b1). t_0=4800

Fig 6(b2). t_0=4850

Fig 6(b3). t_0=4900

Figure 6(b).

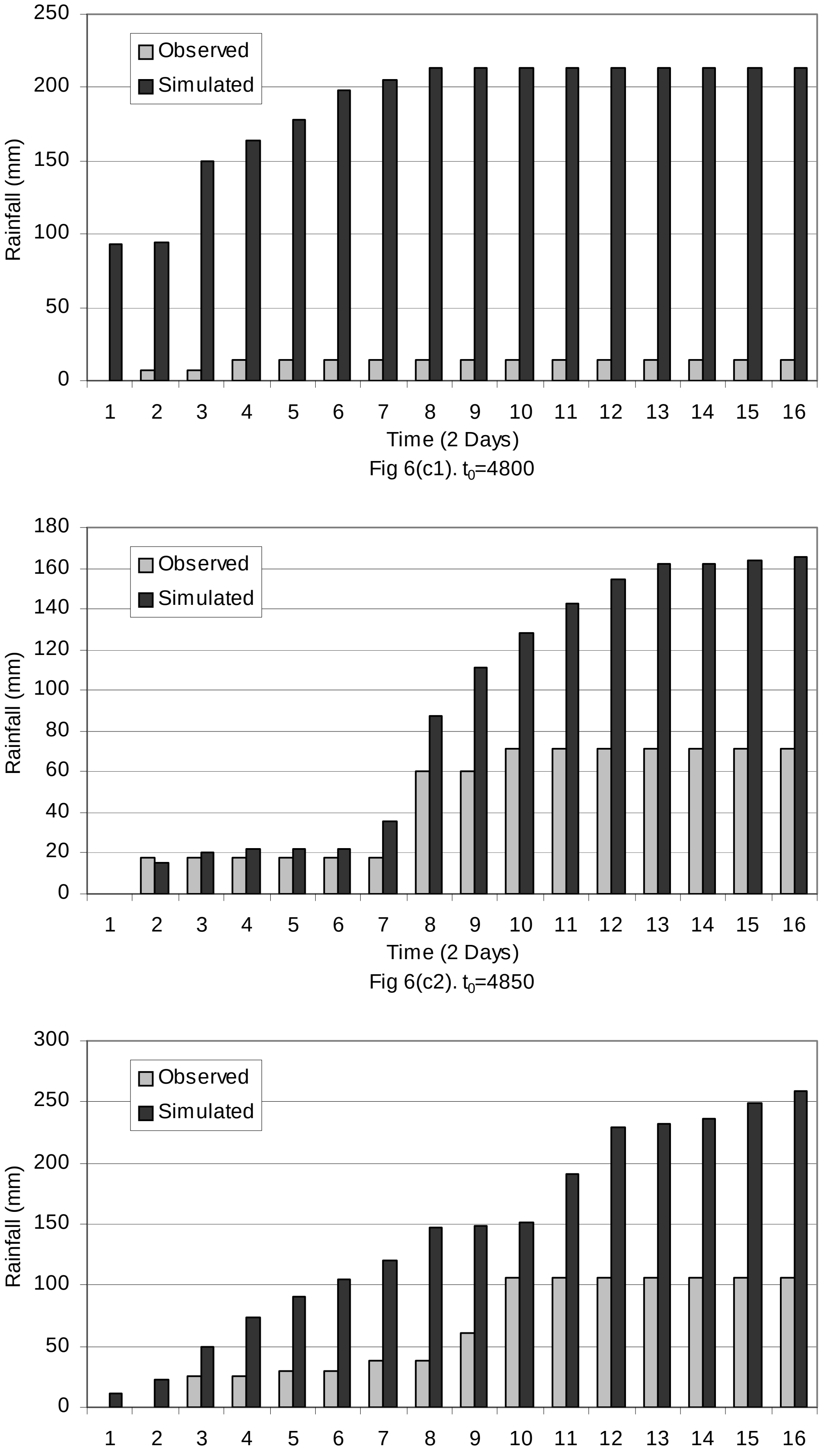


Fig 6(c1). t_0=4800

Fig 6(c2). t_0=4850

Fig 6(c3). t_0=4900

Figure 6(c).

Figure 7.

[redacted]

Figure 7.