Statistical modelling of sequences of no-rain days

Sirangelo B.[1], Caloiero T.[2], Coscarelli R.[3], and Ferrari E.[1*]

[1]*University of Calabria, Dept. of Environmental and Chemical Engineering, Rende (CS)*
[2]*National Research Council of Italy, Institute for Agriculture and Forest Systems in the Mediterranean (CNR-ISAFOM), Rende (CS)*
[3]*National Research Council of Italy, Research Institute for Geo-hydrological Protection (CNR-IRPI), Rende (CS)*
[*]*Corresponding author:* ennio.ferrari@unical.it

Abstract

The stochastic models, developed to simulate long-term hydrological data, can be subdivided in "driven data" models, which reproduce the principal characteristics of the available data series, and "physically based" models, which schematize the generating mechanism of atmospheric precipitation. The initial step of a "driven data" stochastic model, able to adequately simulate the sequences of wet and dry days, is the definition of the statistics of the model. In this paper, various statistical models for sequences of no-rain days are firstly presented: the models are based on an approach which considers the arrival of rainfall events as a Poisson process, homogenous or not. Moreover, the first results of an application of one of these models to the daily rainfall series registered at the Cosenza rain gauge (Calabria, Southern Italy) are also shown. In particular, the model applied is a non-homogeneous Poisson model which considers the rainfall as a pulse of random duration.

Keywords: *Sequences of no-rain days; Statistical models; Poisson distribution*

1. INTRODUCTION

Numerous stochastic models have been developed to simulate long-term hydrological data. All the models can be divided in two main categories: a) the "driven data" models, which reproduce the principal characteristics of the available data series; b) the "physically based" models, which schematize the generating mechanism of atmospheric precipitations.

In the first category the most popular models are the Autoregressive Moving Average (ARMA) ones [1, 2, 3, 4 ,5], used in literature to characterize the correlations within a time series [6]. An ARMA model captures the deterministic components (dominant linear trends) of a time series and leaves behind the stochastic component (residuals). The stochastic component of a time series is defined as persistent if the temporally adjacent values are positively correlated. The persistence may be short-term or long-term depending on the time range over which they are correlated [7].

The physically based models describe the rainfall occurrence (dry–wet) process and the distribution of rainfall amounts on wet days independently [8]. In this category rainfall occurrence can be represented in two ways: as a Markov process [9] or as an Alternating Renewal process for dry and wet sequences [10]. A major limitation of these models is that the adoption of short term memory neglects the existing dependence among different rainfall values separated by longer lags. In absence of such a representation of long term memory, annual sums of the daily simulations exhibit a lower variability than the one observed in reality [11]. Thyer and Kuczera [12] applied a Hidden Markov Model (HMM) for simulating long-term persistence in single site rainfall time series. Their results supported that the HMM provides a conceptually more adequate approach, for simulating long-term persistence in hydrological time series, than the ARMA-type processes. Successively, Thyer and Kuczera [13] presented a Bayesian approach for fully quantifying the parameter uncertainty of a HMM for simulating long-term rainfall time series at multiple sites. This extension showed several advantages over the single site HMM.

Generally, a good "driven data" stochastic model has to adequately simulate the sequences of wet and dry days, thus requiring the definition of the basic statistics of the data series of rainfall measured on ground through rain gauges. To this aim, the no-rain day sequences are strictly embedded to the arrival process of the rainfall events, that can be seen as instantaneous pulses (duration=0) or as pulses of random duration (duration>0).

Usually, the rainfall arrival can be seen as a cluster process, characterized by the variability of both intensity and duration in time and space, which are formulated as Poisson cluster processes [14]. Anyway, for time scale equal to or greater than one day, simple Poisson models characterized by a unique parameter λ (defined as process intensity) can be used. With reference to at-site rainfall measurements, if the λ parameter is constant in time, the model can be defined as temporally homogeneous. More generally, the Poisson model can be not homogeneous in time, thus requiring a time-varying parameter, $\lambda(t)$.

In the following, homogeneous and non-homogeneous Poisson models are presented for explaining no-rain day sequences, considering both instantaneous pulses and pulses of random duration.

Homogeneous Poisson model with instantaneous pulses

Let M be the random variable measuring the number of consecutive intervals of length Δt (equal to 1 day) with zero rainfall, followed by at least one rainy day (Fig. 1).

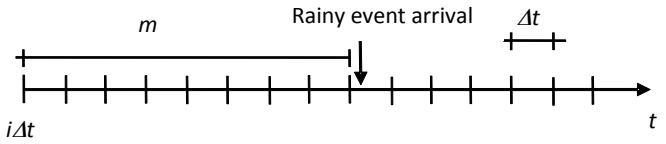

Fig. 1: Schematization of a no-rain day sequence with rainfall as instantaneous pulses.

Under the hypothesis that the arrival process of rainfall is a Poisson model with intensity λ, the probability that at the generic instant $i\Delta t$ a sequence of m no-rain intervals (Δt) will start, followed by the rainy interval $\left[(i+m)\Delta t,(i+m+1)\Delta t\right]$, is given by the following probability density function (pdf):

$$p_M(m)=\begin{cases}\theta^2(1-\theta)^m & m=1,2,\ldots \\ 1-\theta+\theta^2 & \textit{sequence does not start}\end{cases} \quad (1)$$

where:

$$\theta=1-exp(-\lambda\Delta t) \quad (2)$$

The conditional probability of the variable $M|M\geq m_*$, that is the number of m consecutive no-rain intervals equal to, or greater than, a fixed value $m_*>0$, is given by:

$$p_{M|M\geq m_*}(m)=\theta(1-\theta)^{m-m_*} \qquad m=m_*,m_*+1,... \tag{3}$$

By substituting eq. (2) in eq. (3) it can be obtained:

$$p_{M|M\geq m*}(m)=[1-exp(-\lambda\Delta t)]exp[-(m-m_*)\lambda\Delta t] \qquad m=m_*,m_*+1,... \tag{4}$$

The cumulative distribution function (cdf) of $M|M\geq m_*$ can be obtained as follows:

$$P_{M|M\geq m*}(m)=1-exp[-(m-m_*+1)\lambda\Delta t] \qquad m=m_*,m_*+1,... \tag{5}$$

The mean and the variance are equal to:

$$\mu_{M|M\geq m*}=m_*+\frac{exp(-\lambda\Delta t)}{1-exp(-\lambda\Delta t)} \qquad \sigma^2_{M|M\geq m*}=\frac{exp(-\lambda\Delta t)}{[1-exp(-\lambda\Delta t)]^2} \tag{6}$$

Homogeneous Poisson model with pulses of random duration

With reference to a generic instant $i\Delta t$ within a rainfall event, we define the random variable M' as the number of consecutive intervals of length Δt (equal to 1 day) without arrival of rainfall events, followed by at least one rainy day (Fig. 2).

The probability that, at the generic instant $i\Delta t$, a sequence of m' intervals of length Δt without arrival of rainfall events will start, followed by a rainy event is equal to eq. (1).

Within the sequence of m' intervals, it can be useful to define both the random variable L (l=0, 1, 2, …) as the number of rainy intervals Δt belonging to a rainfall event started before the instant $i\Delta t$, and the random variable K' (k'=0, 1, 2, …) as the number of consecutive intervals with no rainfall, where $K'=M'-L$.

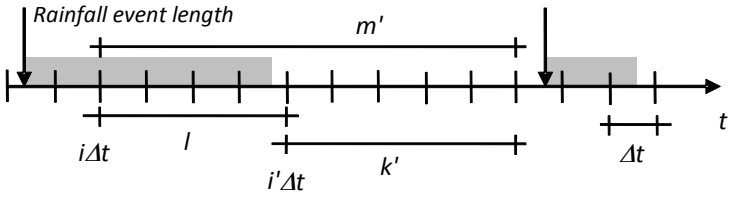

Fig. 2: Schematization of variables m' (sequence of days without arrival of rainfall events), l (sequence of rainy days within m') and k' (difference of the two variables m' and l) with reference to a generic instant $i\Delta t$.

The pdf that at the generic instant $i'\Delta t$ a sequence of k' intervals of length Δt without rainfall will start, followed by a rainy interval $\left[(i'+k')\Delta t,(i'+k'+1)\Delta t\right]$, can be expressed as:

$$p_{K'}(k')=\begin{cases} r\theta(1-\theta)^{k'} & k'=1,2,\ldots \\ 1-r(1-\theta) & \textit{sequence does not start} \end{cases} \tag{7}$$

where $r=\sum_{l=0}^{\infty} p_L(l)\theta^l$, being $p_L(l)$ the pdf of the variable L.

Given the number of consecutive intervals $K'|K'\geq k'_*$ of length Δt with no rainfall equal to, or greater than, $k'_*>0$, the probability that at the generic instant $i'\Delta t$ a sequence of k' intervals without rainfall followed by a rainy interval will start, is expressed by the pdf:

$$p_{K'|K'\geq k'_*}(k')=\theta(1-\theta)^{k'-k'_*} \qquad k'=k'_*,k'_*+1,\ldots \tag{8}$$

By substituting eq. (2) in eq. (8), it yields:

$$p_{K'|K'\geq k'_*}(k')=\left[1-exp(-\lambda\Delta t)\right]exp\left[-(k'-k'_*)\lambda\Delta t\right] \qquad k'=k'_*,k'_*+1,\ldots \tag{9}$$

The corresponding cdf is:

$$P_{K'|K'\geq k'_*}(k')=1-(1-\theta)^{k'-k'_*+1} \qquad k'=k'_*,k'_*+1,\ldots \tag{10}$$

By substituting eq. (2) in eq. (10), it yields:

$$P_{K'|K'\geq k'_*}(k')=1-exp\left[-(k'-k'_*+1)\lambda\Delta t\right] \qquad k'=k'_*,k'_*+1,\ldots \tag{11}$$

The mean and the variance are:

$$\mu_{K'|K'\geq k'_*} = k'_* + \frac{exp(-\lambda\Delta t)}{1-exp(-\lambda\Delta t)} \qquad \sigma^2_{K'|K'\geq k'_*} = \frac{exp(-\lambda\Delta t)}{[1-exp(-\lambda\Delta t)]^2} \tag{12}$$

Non-homogeneous Poisson model with instantaneous pulses

The variability of the meteorological conditions triggering rainfall are clearly time dependent [15, 16, 17]. Meteorological phenomena show statistically significant seasonal and daily features due to the revolution of the earth around itself and the sun. Particularly, the storm events in the Mediterranean area show non-stationary behavior with intensity depending on time, $\lambda(t)$, as shown by their marked reduction in summer. Due to this evident periodicity, occurrences of rainfalls can be considered reliably homogeneous only in short temporal intervals. In this case, a stochastic model based on a non-homogeneous Poisson process, characterized by a time-dependent intensity of rainfall occurrence $\lambda(t)$, can be well fitted to explain seasonal variation of no rainfall sequences. Since the probabilistic distribution of $\lambda(t)$ depends on the starting time of the no-rain day sequence, the temporal variation of the intensity parameter $\lambda(t)$ can be statistically developed through a truncated Fourier series. Thus the temporal variation of rainfall intensity $\lambda(t)$ can be expressed as a function of period D:

$$\lambda(t) = \frac{1}{2}a_0 + \sum_{j=1}^{n_h}\left[a_j cos\left(\frac{2\pi j}{D}t\right) + b_j sin\left(\frac{2\pi j}{D}t\right)\right] \tag{13}$$

where a_0, $a_j, b_j ; j = 1,2,\ldots,n_h$ are the coefficients of the truncated Fourier series, n_h is the number of harmonics and D is the period of the function ($D = 365.25$ for $\Delta t = 1$ day). The integral of the intensity function of non-homogeneous Poisson process is:

$$\Lambda(t) = \frac{1}{2}a_0 t + \frac{D}{2\pi}\sum_{j=1}^{n_h}\left[b_j cos\left(\frac{2\pi j}{D}t\right) - a_j sin\left(\frac{2\pi j}{D}t\right)\right] + c \tag{14}$$

where c is an integration constant. The mean value of the arrival number in the generic interval $(t_1, t_2]$ is given by:

$$\Lambda(t_2) - \Lambda(t_1) \tag{15}$$

Considering the random variables M and $M|M \geq m_*$ defined above, the pdf and cdf expressed by equations (4) and (5), respectively, assume the expressions:

$$p_{M|M\geq m_*}(m) = \\ = \left[1 - exp\left(-\Delta\Lambda_{i+m,\,i+m+1}\right)\right] exp\left[-\Delta\Lambda_{i+m_*,\,i+m+1}\right] \quad m = m_*, m_*+1, \ldots \tag{16}$$

where $\Delta\Lambda_{j_1,j_2} = \Lambda(j_2\Delta t) - \Lambda(j_1\Delta t)$, and:

$$P_{M|M\geq m_*}(m) = 1 - exp\left(-\Delta\Lambda_{i+m_*,\,i+m+1}\right) \quad m = m_*, m_*+1, \ldots \tag{17}$$

The mean and the variance of the random variable $M|M \geq m_*$ are:

$$\mu_{M|M\geq m_*} = m_* + \sum_{m=m_*}^{\infty} exp\left(-\Delta\Lambda_{i+m_*,\,i+m+1}\right) \tag{18}$$

$$\sigma^2_{M|M\geq m_*} = \\ = 2\sum_{m=m_*}^{\infty} m\, exp\left(-\Delta\Lambda_{i+m_*,\,i+m+1}\right) - \left(\mu_{M|M\geq m_*} - m_*\right)\left(1 + \mu_{M|M\geq m_*} - m_*\right) \tag{19}$$

Non-homogeneous Poisson model with pulses of random duration

With reference to the random variables M' and K' introduced above, in the case of non-homogeneous Poisson model, it can be verified that the pdf of the random variable K' assumes the following expression:

$$p_{K'|K'\geq k'_*}(k') = \\ = \left[1 - exp\left(-\Delta\Lambda_{i'+k',\,i'+k'+1}\right)\right] exp\left[-\Delta\Lambda_{i'+k'_*,\,i'+k'}\right] \quad k' = k'_*, k'_*+1, \ldots \tag{20}$$

and the cdf is:

$$P_{K'|K'\geq k'_*}(k') = 1 - exp\left(-\Delta\Lambda_{i'+k'_*,\,i'+k'+1}\right) \quad k' = k'_*, k'_*+1, \ldots \tag{21}$$

The mean and the variance are:

$$\mu_{K'|K'\geq k'_*} = k'_* + \sum_{k'=k'_*}^{\infty} exp\left(-\Delta\Lambda_{i'+k'_*,\, i'+k'+1}\right) \tag{22}$$

$$\sigma^2_{K'|K'\geq k'_*} = \\ = 2\sum_{k'=k'_*}^{\infty} k' exp\left(-\Delta\Lambda_{i'+k'_*,\, i'+k'+1}\right) - \left(\mu_{K'|K'\geq k'_*} - k'_*\right)\left(1+\mu_{K'|K'\geq k'_*} - k'_*\right) \tag{23}$$

2. RESULTS AND DISCUSSION

In this paper, in order to analyse no-rain day sequences, the non-homogeneous Poisson model with pulses of random duration has been applied to the Cosenza rain gauge, which has been chosen due to the high quality data. In fact, the observation period of daily rainfall data for this rain gauge spans from 1916 to 2010 and presents missing data only during the Second World War period. Thus, only data from 1951 to 2010 have been selected.

To ensure parsimony of the model, Fourier series has to be limited to the minimum number of harmonics. The estimation of the Fourier coefficients for one harmonic has been obtained through the maximum likelihood method, expressed as:

$$lnL(a_0,a_1,b_1) = -\sum_{n=1}^{N_d} \Delta\Lambda_{i'_n+k'_*,\, i'_n+k'_n}(a_0,a_1,b_1) + \\ + ln\prod_{n=1}^{N_d}\left[1-exp\left(-\Delta\Lambda_{i'_n+k'_n,\, i'_n+k'_n+1}(a_0,a_1,b_1)\right)\right] \tag{24}$$

where N_d is the number of no-rain day sequences. To guarantee stability in the statistical estimation, the sequences used in the application are the ones with duration equal to or greater than 5 days ($k'_* = 5$). Moreover, with the aim to analyse temporal change in the sequences, the whole series has been subdivided in two subseries, each composed by 30 years of observation: 1951-1980 and 1981-2010. The estimated Fourier parameters are reported in Tab. 1.

30-year subperiod	N_d	a_0	a_1	b_1
1951-1980	485	0.307	0.0613	0.0280
1981-2010	469	0.279	0.0626	0.0231

Table 1: Statistical features of the different subperiods and estimated values of Fourier coefficients. N_d=number of data of no-rain day sequences; a_0, a_1, b_1: Fourier coefficients.

In order to analyse the seasonality of the no-rain sequence duration, K', fig. 3 shows, for each day of the year, the mean value of this variable (in log scale) for both the 30-year periods. The distribution of K' presents maximum values between about the 150th and the 250th day of the year, a time span which roughly corresponds to the summer period, while minimum values are detected in winter. Moreover, by using the estimated Fourier parameters (Tab. 1), the expected values of K', obtained through a Monte Carlo simulation, have been compared with the corresponding observed values derived by the historical data series.

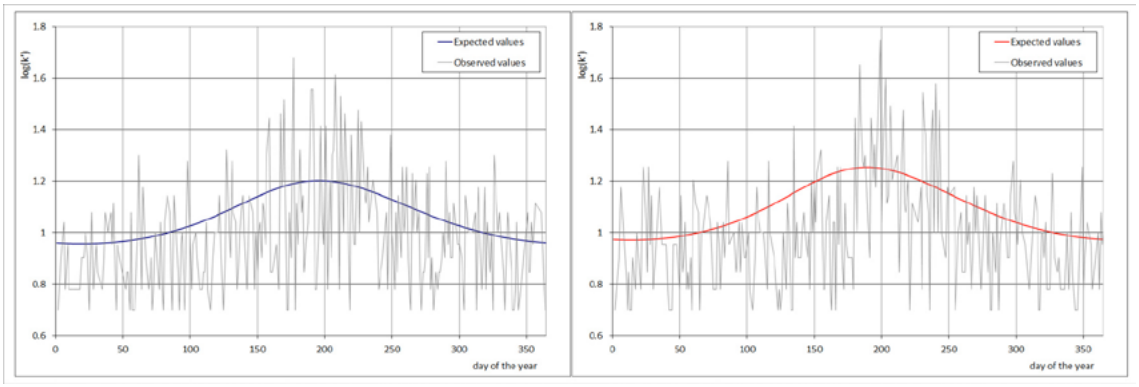

Fig. 3: Comparison between observed and expected values (maximum likelihood method) of K' for the two subperiods (left: 1951-1980; right: 1981-2010).

In Fig. 4 the difference between the daily distributions of the expected values of K', $E(K')$, estimated for both the subperiods, is shown. For each day of the year, $E(K')$ values estimated in the 1981-2010 subperiod are higher than the ones estimated in the 1951-1980 subperiod, and this difference is more evident during the summer period. Moreover, the maximum value of $E(K')$ for the 1981-2010 subperiod falls few days before the correspondent value of the 1951-1980 subperiod.

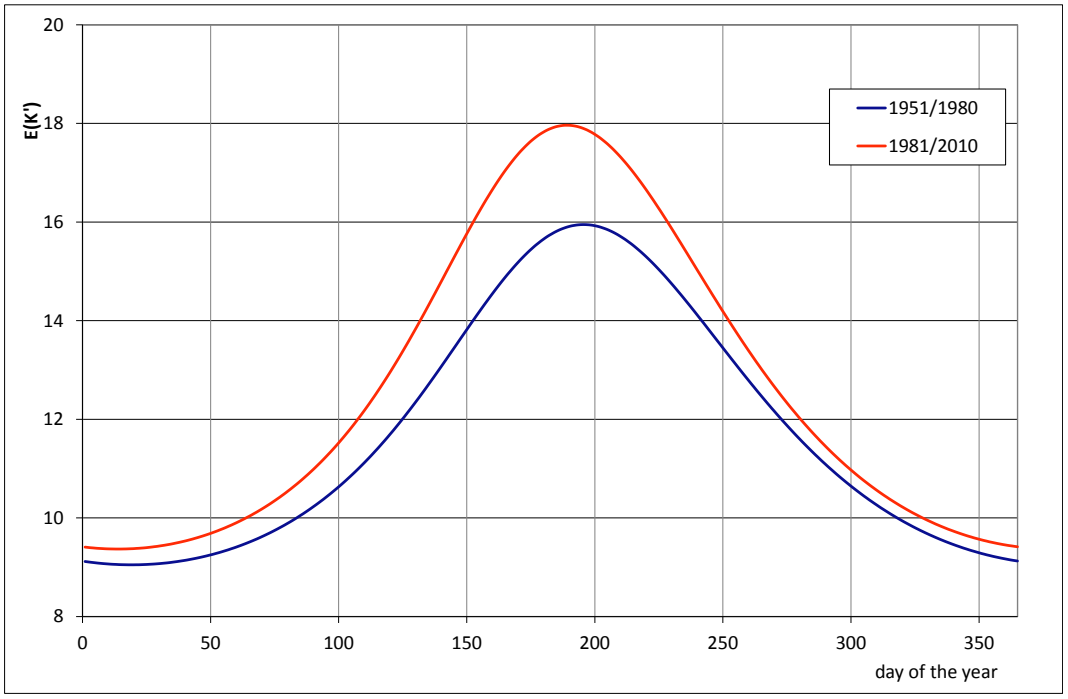

Fig. 4: Comparison between expected values of *K'* for the two subperiods.

These results show a change in rainfall distribution during the year, with a longer duration of the no-rain day sequences in the last 30-year period, even if the number of sequences (N_d) slowly decreases. The longer duration of the no-rain day sequences matches with the general tendency of increasing frequency of the drought periods, observed in various part of the world [18].

The research is the first step of a wider investigation regarding the probabilistic modelling of both occurrence and intensities of wet and dry rainfall periods, by considering, in particular, the cumulated rainfall lower than fixed thresholds. Anyway, the procedure presented in this work needs an improvement as regards parameters estimation and validation on a larger number of data series, that will be the object of a further application. Moreover, the at-site procedure can be replaced by adopting spatial analysis, taking into account the rainfall variability within a region. The tendency showed in this study, if confirmed by further analyses based on more detailed rainfall stochastic models applied to larger data base, could be very useful in water resources planning and management of specific drainage basins.

3. REFERENCES

Box G.E.P. & Jenkins G.M. (1970), *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco, USA. (1)

Salas J.D. & Smith R.A. (1981), *Physical basis of stochastic models of annual flows*, Water Resources Research, No. 17, pp. 428–430. (2)

Salas J.D. (1993), *Analysis and modeling of hydrologic time series*, In: Maidment, D. (Ed), Handbook of Hydrology, McGraw-Hill, New York, USA. (3)

Grayson R.B., Argent R.M., Nathan R.J., McMahon T.A. & Mein R.G. (1996), *Hydrological Recipes: Estimation Techniques in Australian Hydrology*, Cooperative Research Centre for Catchment Hydrology, Melbourne, Australia. (4)

Srikanthan R. & McMahon T.A. (2000), *Stochastic Generation of Climate Data: A Review, CRC for Catchment Hydrology*, Monash University, Clayton, Victoria, Australia. (5)

Malamud B.D. & Turcotte D.L. (1999), *Self affine time series: measures of weak and strong persistence*, Journal of Statistical Planning and Inference, No. 80, pp. 173–196. (6)

Koirala S., Gentry R.W., Mulholland P.J., Perfect E. & Schwartz J.S. (2010), *Time and frequency domain analysis of high-frequency hydrologic and chloride data in an east Tennessee watershed*, Journal of Hydrology, No. 387, pp. 256–264. (7)

Woolhiser D.A. (1992), *Modeling daily precipitation -progress and problems*, In: Walden, A.T. & Guttorp, P. (Eds), Statistics in the environmental and earth sciences. Edward Arnold, London, UK. (8)

Gabriel K.R. & Neumann J. (1962), *A Markov chain model for daily rainfall occurrences at Tel Aviv*. Quarterly Journal of the Royal Meteorological Society, No. 88, pp. 90–95. (9)

Buishand T.A. (1978), *Some remarks on the use of daily rainfall models*, Journal of Hydrology, No. 36, pp. 295–308. (10)

Srikanthan R. (2005), *Stochastic generation of daily rainfall using a nested transition probability model*. 29th Hydrology and Water Resources Symposium, Engineers Australia, Canberra, Australia. (11)

Thyer M. & Kuczera G. (2000), *Modelling long-term persistence in hydro-climatic time series using a hidden state Markov model*, Water Resources Research, No. 36, pp. 3301–3310. (12)

Thyer M. & Kuczera G. (2003), *A hidden Markov model for modelling long-term persistence in multi-site rainfall time series. 2. Real data analysis*, Journal of Hydrology, No. 275, pp. 27–48. (13)

Daley D.J. & Vere-Jones D. (1988), *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, USA. (14)

Waymire E. & Gupta V.K. (1981), *The mathematical structure of rainfall representations 1. A review of the stochastic rainfall models*. Water Resources Research, No. 17, pp. 1261–1272. (15)

Smith J.A. & Karr A.F. (1983), *A point process model of summer season rainfall occurrences*, Water Resources Research, No. 19, pp. 95–103. (16)

Chang T.J., Kavvas M.L. & Delleur J.W. (1984), Daily precipitation modeling by discrete autoregressive moving average processes, Water Resources Research, No. 20, pp. 565–580. (17)

IPCC (2007), *Climate change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, USA. (18)