Bayesian Time Series Modelling of the Italian Daily Rainfall Data using Mixed Distribution

Muhammad Safwan Bin Ibrahim, Muhammad Irfan Bin Abdul Jalal

Abstract: ***A combination of continuous and discrete elements is referred to as a mixed distribution. For example, daily rainfall data consist of zero and positive values. We aim to develop a Bayesian time series model that captures the evolution of the daily rainfall data in Italy, focussing on directly linking the amount and occurrence of rainfall. Two gamma (G1 and G2) distributions with different parameterisations and lognormal distribution were investigated to identify the ideal distribution representing the amount process. Truncated Fourier series was used to incorporate the seasonal effects which captures the variability in daily rainfall amounts throughout the year. A first-order Markov chain was used to model rainfall occurrence conditional on the presence or absence of rainfall on the previous day. We also built a hierarchical prior structure to represent our subjective beliefs and capture the initial uncertainties of the unknown model parameters for both amount and occurrence processes. The daily rainfall data from Urbino rain gauge station in Italy were then used to demonstrate the applicability of our proposed methods. Residual analysis and posterior predictive checking method were utilised to assess the adequacy of model fit. In conclusion, we clearly found that our proposed method satisfactorily and accurately fits the Italian daily rainfall data. The gamma distribution was found to be the ideal probability density function to represent the amount of daily rainfall.***

Keywords: ***Time series, Bayesian analysis, mixed distribution, first-order Markov chain, daily rainfall***

I. INTRODUCTION

Complex procedures are required in developing a model of daily rainfall data as the distribution consists of a mixture of discrete and continuous elements. The discrete element indicates rainfall occurrence and continuous element represents the amount of rainfall when rain occurs. This type of distribution is known as the mixed distribution. This distribution has been used for the past several decades in studies which implemented a two-stage strategy. Stern and Coe [1] demonstrated that the two-stage approach can be conveniently and directly used for developing a daily rainfall model and other hydrology applications. Tooze et al. [2] also implemented similar strategy to analyse the medical costs incurred in the United States. It was shown from both studies that the two-stage strategy is ideal for modelling mixed distributions.

The daily rainfall model could be broken down into two parts, namely the occurrence process and amount process. Specifically, through the occurrence process, the probability of rainfall could be modelled. Most research endeavours utilised the first-order Markov chain to model the occurrence process [1, 3, 4, 5]. On the other hand, conditional on rainfall occurrence, the amount of precipitation can be represented by the amount process. In previous literature, the gamma [1, 5, 6], lognormal [2, 7], power transformed truncated normal distributions [8], Weibull [2], exponential [9], and mixed exponential distributions [10] are the most frequently used distributions for the amount process.

Generalized Linear Models (GLMs) have also been used in previous studies to build a model for daily rainfall. Coe and Stern [3] and Stern and Coe [1] found that this approach possessed a direct function that can be exploited to model the daily rainfall pattern. Furthermore, a simple linear regression was applied to model both the occurrence and amount processes using a Fourier series as the model covariate. The same modelling strategy was used by Grunwald and Jones [5] for an Australian daily rainfall model through the involvement of more complex covariates in the GLM model. Besides, it was also possible for a multitude of other weather variables such as speed of wind, atmospheric circulation pattern, and temperature, to be included as additional covariates in the GLM [11,12,27,28]. However, the majority of this literature in rainfall modelling were conducted using frequentist approach.

In this study, we endeavoured to expand the methodological frameworks of Coe and Stern [3], Stern and Coe [1], and Grunwald and Jones [5] in which a Fourier series representing the covariates for occurrence and amount processes were used to model daily rainfall pattern within the Bayesian framework. This will be initially performed on a univariate setting prior to expanding the model to include multiple sites. Besides, we also investigated the link between these two processes and characterised the changes in the shape of the probability density function of the mixed distribution, especially when the parameters varied temporally. Finally, we also sought to identify and recommend the ideal probability density function to represent the amount process.

II. METHODOLOGY

A. General Structure

Let $W_t \in \mathbb{R}+$ denote a random variables for the amount of

Revised Manuscript Received on November 15, 2019
***** Correspondence Author
Muhammad Safwan Bin Ibrahim *, Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan. Email : msafwan@usim.edu.my
Abdul Jalal, School of Mathematics, Statistics and Physics, Newcastle University, Newcastle Upon Tyne, NE1 7RU United Kingdom. Email: muhammadirfan1504@gmail.com

the daily rainfall where w_t is an observed value at time t for $t = 1, \ldots, T$. Provided that every observation of the daily rainfall data is a realisation of a random process, the distribution of W_t conditional on p_t and μ_t (where $0 \leq p_t \leq 1$) can be represented by the following equation:

$$F_W(w_t|p_t,\mu_t) = \Pr(W_t \leq w_t|p_t,\mu_t) = \begin{cases} 0, & (w_t < 0) \\ 1 - p_t, & (w_t = 0) \\ p_t F_A(w_t|\mu_t), & (w_t > 0) \end{cases}$$

In this case, $F_A(w_t|\mu_t)$ represents the distribution function of the amount of rainfall, where $\Pr(W_t = 0)=1-p_t$. The conditional probability density function of W_t is denoted by $f(w_t|\mu_t)$, provided $W_t > 0$. The rainfall occurrence is represented by R_t which is also an indicator function for W_t. Hence,

$$R_t = \begin{cases} 0, & W_t = 0 \\ 1, & W_t > 0. \end{cases} \quad (1)$$

Thus, the random variable W_t can be re-expressed as

$$W_t = I(W_t > 0)Y_t = R_t Y_t. \quad (2)$$

Accordingly, $Y_t = g(Z_t)$ indicates a continuous random variable and $g(.)$ represents some monotonic function defining a suitable transformation (e.g. exponential) of Z_t. In this case, W_t represents the exact amount of rainfall which may take the value of zero, but it is always observable. Stern and Coe [1] and Grunwald and Jones [5] emphasised that Y_t could be considered as an intensity process which can also be viewed as the potential rainfall amount. Notably, although the value of Y_t was always positive, it was not constantly observed.

An example of a directed acyclic graph (DAG) for the univariate model of daily rainfall is presented in Fig. 1. Although $\ldots, Y_{t-1}, Y_t, Y_{t+1}, \ldots$ were considered independent in this model, their dependency on μ_t varied throughout time. This indicates that Y_t variables are not only conditionally independent of R_t, they are also independent of the model parameters.

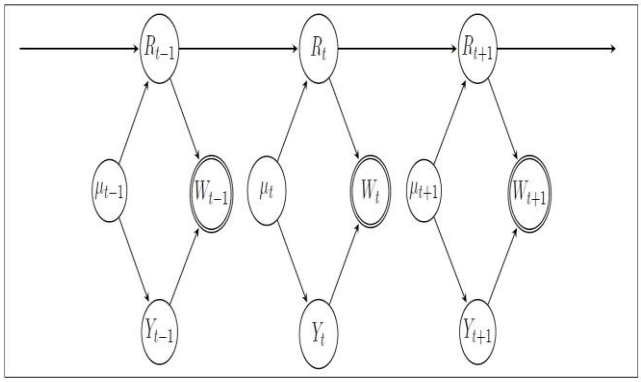


Fig. 1. A representation of the temporal dependence form of the model of Italian daily rainfall through a DAG

B. Seasonal Effects

A truncated Fourier series is a popular approach for the development of periodic time series in meteorological studies. Jones and Brelsford [13] and West and Harrison [14] emphasised that the model of the periodic form in time series could be developed through sinusoidal representations. This approach was subsequently implemented by Stern and Coe [1] in their assessment of daily rainfall amount and occurrence for agriculture planning. Furthermore, the seasonal pattern of the amount and occurrence processes of the Australian daily rainfall data was also elaborated through this method [5].

A Fourier series comprises of cosine and sine terms of harmonic frequencies and period ω. This is represented by

$$\hat{F}(t) = \sum_{f=1}^{F} \left[a_f \cos\left(\omega_f t\right) + b_f \sin\left(\omega_f t\right) \right]. \quad (3)$$

The number of sinusoids is denoted by F, while the Fourier coefficients are represented by a_f and b_f, with $\omega_f = 2\pi f/L$. As for the angular frequency, L represents the time. In respect of daily data which displayed an annual cyclic pattern, L equals to 365.25. It was highlighted in the previous research that the first three harmonics were usually sufficient for the development of the model that captures seasonal impacts on rainfall and to determine a Fourier function that fits the model well [15, 16, 17]. Therefore, the incorporation of the seasonal effects into the models used in this study is a cogent modelling strategy based on the aforementioned idea.

C. Amount Process

The development of the model for the amount process was based on the recommendations by Fernandes et al. [9] and Suhaila et al. [18]. Specifically, a number of distributions were selected for the purpose of identifying and evaluating the ideal distribution for this process. Several studies have been conducted to identify the best distribution for daily rainfall. However, the ideal probability density function to represent rainfall amount remains elusive [10, 19, 20].

In our study, both the lognormal and gamma distributions were chosen and compared to represent the amount process since both possess support from 0 to ∞. Such distributional support made them an ideal candidate to model the amount process. Previously, a similar comparison was conducted by Cho et al. [19] for rainfall data. They found that the gamma distribution was more appropriate to develop the wet region model, while the lognormal distribution was more suitable for dry regions. In this study, two different parameterisations were used and investigated for the gamma distribution. In the first parameterisation, a gamma distribution with variable scale (β) and fixed shape (α) parameters was used. This distribution was identified as "G1". As for the second parameterisation, a gamma distribution with fixed scale (β) and variable shape (α) parameters were utilised and this was considered as "G2" distribution. The full characteristic properties of each distribution are given as follows:

- Lognormal distribution:

  For lognormal distribution case, the rainfall amount is defined by

$$W_t = R_t Y_t$$

Accordingly, Y_t is presumed to follow a lognormal distribution. Since $Z_t = \log(Y_t)$, then Z_t is therefore normally distribution and the probability density function of Z_t is therefore given by

$$f(z_t|\vartheta_t, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}(z_t - \vartheta_t)^2\right\}. \quad (4)$$

Hence, $E(Z_t) = \vartheta_t$ and $Var(Z_t) = 1/\tau$. The mean and variance of Y_t are respectively given by $\mu_t = e^{\vartheta t + 1/2\tau}$ and $V_y = e^{2\vartheta t + 1/\tau}(e^{1/\tau} - 1)$.

- Gamma distribution:

Gamma distribution is another alternative that was ubiquitously used for representing rainfall amount. Furthermore, it is a notable approach in rainfall data due to the good fit it develops and its shape, which has a similarity to the shape of rainfall data histogram [21]. Following is the observation of rainfall amount:

$$W_t = R_t Y_t$$

where Y_t is based on the gamma distribution. The representation of the G1 distribution is as follows:

$$f(y_t|\alpha, \beta_t) = \frac{\beta_t^{\alpha} y_t^{\alpha-1} e^{-\beta_t y_t}}{\Gamma(\alpha)}. \quad (5)$$

Based on Eq. (5) above, $\beta_t = \alpha/\mu_t$ represents the variable scale parameter, while α (provide $\alpha > 0$) is the constant shape parameter. $E(Y_t) = \mu_t$ is the mean of this distribution, while $Var(Y_t) = \mu^2/\alpha$ represents its variance.

For G2 distribution, its probability density function is given by

$$f(y_t|\beta, \alpha_t) = \frac{\beta^{\alpha_t} y_t^{\alpha_t - 1} e^{-\beta y_t}}{\Gamma(\alpha_t)} \quad (6)$$

This distribution has a variable shape ($\alpha_t = \mu_t\beta$) parameter with a fixed scale parameter, $\beta > 0$. Although the expectations (means) of G1 and G2 distributions have identical mathematical forms, their variances are different. In G2 case, $Var(Yt) = \mu_t/\beta$.

An important consideration when selecting between a gamma and lognormal distribution is whether the density would go to zero when y closely approaches zero. The occurrence of changes in distribution shape and their factors should also be focused on provided if the mean is changed. Notably, this is the difference between the G1 and G2 models.

In the case of the rainfall amount, the mean of distribution is assumed to be varied throughout time. In determining a good fit, the information presented by Stern and Coe [1] were used in this study. Therefore, the truncated Fourier series was applied to develop the model of daily rainfall variation throughout the year. For the lognormal distribution, the mean of the log amount Z_t is represented as follows:

$$\vartheta_t = \hat{\eta}_t,$$

Meanwhile, following are the means for G1 and G2 distributions:

$$\mu_t = e^{\hat{\eta}_t}$$

where

$$\hat{\eta}_t = \eta_0 + \sum_{f=1}^{F}\left[a_f \cos\left(\frac{2\pi f t}{365.25}\right) + b_f \sin\left(\frac{2\pi f t}{365.25}\right)\right]. \quad (7)$$

The common level for the amount of rainfall is represented by the parameter η_0.

D. Occurrence Process

Let $R = \{R_1, R_2, \cdots, R_t\}$ represent the sequence of the daily rainfall occurrence where $t = 1, 2, \cdots, T$. A first-order Markov chain was used to develop a model of the probability of rainfall occurrence. This approach had been similarly used by Gabriel and Neumann [4] to elucidate the phases of daily rainfall in Tel Aviv, Israel. An extension of the Markov chain to higher-order models is also feasible. However, it was highlighted by Jimoh and Webster [22] that a first-order Markov chain model was adequate to predict the daily rainfall occurrence. Hence, utilising the first-order Markov chain is a germane strategy to model the occurrence process in our case.

The probabilities of the first-order Markov chain transitions are presented as follows:

$$p_{ij}(t) = \Pr\{r_t = j | r_{t-1} = i\}; \qquad i, j \in \{0, 1\}.$$

The transition matrix is therefore given by:

$$P_t = \begin{matrix} \\ r_{t-1} = 0 \\ r_{t-1} = 1 \end{matrix} \begin{matrix} r_t = 0 & r_t = 1 \\ \end{matrix} \begin{pmatrix} p_{00}(t) & p_{01}(t) \\ p_{10}(t) & p_{11}(t) \end{pmatrix}.$$

A logistic link function was then used to link the predictors with rainfall probability $p_{i1}(t)$. This can be succinctly represented by

$$\text{logit}\,[p_{i1}(t)] = \hat{\zeta}_{i,t}; \qquad i = 0, 1.$$

Accordingly, the conditional distribution of the occurrence process given $p_{i1}(t)$ has a Bernoulli form:

$$R_t \sim Bern\left[p_{i,t}(t)\right]; \qquad i = 0, 1.$$

If the distribution of Y_t is a lognormal distribution, then is parametrised as follows:

$$\hat{\zeta}_{i,t} = \zeta_0 + \sum_{f=1}^{F}\left[c_f \cos\left(\frac{2\pi f}{365.25}\right) + d_f \sin\left(\frac{2\pi f}{365.25}\right)\right] + \zeta_1(\vartheta_t - \zeta_0) + \zeta_2\left(i - \frac{1}{2}\right). \quad (8)$$

If Y_t follows a gamma distribution, then

$$\hat{\zeta}_{i,t} = \zeta_0 + \sum_{f=1}^{F}\left[c_f \cos\left(\frac{2\pi f}{365.25}\right) + d_f \sin\left(\frac{2\pi f}{365.25}\right)\right] + \zeta_1(\log(\mu_t) - \zeta_0) + \zeta_2\left(i - \frac{1}{2}\right). \quad (9)$$

If rain occurred on the previous day, the term i, which represents an indicator function, would take the value 1. If rain was absent, then i=0. Besides, the common intercept for logit $[p_{i1}(t)]$ is represented through the parameter ζ_0.

To develop a connection between the occurrence and amount processes, ϑ_t and $\log(\mu_t)$ were incorporated in Eqs. (8) and (9), respectively. Using this approach, the precision of the rainfall possibility could be considerably improved. To emphasise this, the rise of rainfall probability occurred in tandem with the rise of μ_t.

The conditional probabilities can be summarised as follows:

$$p_{11}(t) = \frac{\exp(\hat{\zeta}_{1,t})}{1+\exp(\hat{\zeta}_{1,t})}; \qquad p_{10}(t) = 1 - p_{11}(t),$$

$$p_{01}(t) = \frac{\exp(\hat{\zeta}_{0,t})}{1+\exp(\hat{\zeta}_{0,t})}; \qquad p_{00}(t) = 1 - p_{01}(t).$$

Meanwhile, the unconditional probability is given by

$$\begin{aligned}\hat{p}_t &= p_{11}(t)\hat{p}_{t-1} + p_{01}(t)(1-\hat{p}_{t-1}) \\ &= p_{01}(t) + [p_{11}(t) - p_{01}(t)]\hat{p}_{t-1} \\ \hat{p}_t + [p_{01}(t) - p_{11}(t)]\hat{p}_{t-1} &= p_{01}(t).\end{aligned}$$

The above equation could also be formulated as follows:

$$\ddot{A}\hat{\boldsymbol{P}} = \boldsymbol{p}$$

This will clearly lead to:

$$\hat{\boldsymbol{P}} = \ddot{A}^{-1}\boldsymbol{p} \qquad (10)$$

where

$$\ddot{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & \ddot{a}_1 \\ \ddot{a}_2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddot{a}_3 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & \ddot{a}_{365} & 1 \end{bmatrix}$$

and $\ddot{a}_t = p_{01}(t) - p_{11}(t)$.

E. Prior Specifications

Prior distributions were developed for the unknown parameters in our model. The prior structures will be elaborated in detail in this section. Furthermore, $\pi(\boldsymbol{\theta})$ represents the prior density of the unknown parameters. The development of these priors was based on the details obtained from past works of research and experts' subjective beliefs.

Suppose that the precision parameter, τ, in the lognormal distribution follows a gamma distribution:

$$\tau \sim Ga(g_\tau, h_\tau)$$

with constant g_τ and h_τ hyperparameters. Provided that the gamma distribution is semi-conjugate to the likelihood function of the lognormal form, this was a reasonable choice. Gamma priors were also assigned to the parameters α in the G1 distribution and β in the G2 distribution due to their strictly positive values. The priors are therefore represented by

$$\alpha \sim Ga(g_\alpha, h_\alpha); \qquad \beta \sim Ga(g_\beta, h_\beta).$$

The fixed hyperparameters are hence g_α, h_α, g_β and h_β.

In linking the linear predictor to the response variable, a GLM approach was implemented which involves a link function to relate the occurrence and amount processes. The set of linear coefficients for the amount process is indicated by $\eta = (\eta_0, a_1, a_2, a_3, b_1, b_2, b_3)$ while the parameter vector for the occurrence process was represented by $\zeta = (\zeta_0, \zeta_1, \zeta_2, c_1, c_2, c_3, d_1, d_2, d_3)$. A normal prior for each unknown parameter in the linear predictors is appropriate since their values may be between $-\infty$ to ∞. Since we assumed an independent prior structure between the parameters, the priors for all parameters are thus given by

- Amount process

$$\eta_0 \sim N(m_{A0}, v_{A0}) \qquad a_1 \sim N(m_{a_1}, v_{a_1})$$
$$a_2 \sim N(m_{a_2}, v_{a_2}) \qquad a_3 \sim N(m_{a_3}, v_{a_3})$$
$$b_1 \sim N(m_{b_1}, v_{b_1}) \qquad b_f \sim N(m_{b_2}, v_{b_2})$$
$$b_3 \sim N(m_{b_3}, v_{b_3})$$

- Occurrence process

$$\zeta_0 \sim N(m_{C0}, v_{C0}) \qquad \zeta_1 \sim N(m_{C1}, v_{C1})$$
$$\zeta_2 \sim N(m_{C2}, v_{C2}) \qquad c_1 \sim N(m_{c_1}, v_{c_1})$$
$$c_2 \sim N(m_{c_2}, v_{c_2}) \qquad c_3 \sim N(m_{c_3}, v_{c_3})$$
$$d_1 \sim N(m_{d_1}, v_{d_1}) \qquad d_2 \sim N(m_{d_2}, v_{d_2})$$
$$d_3 \sim N(m_{d_3}, v_{d_3}).$$

Thus, we can reasonably regard that the priors for $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ follow a multivariate normal distribution:

$$\boldsymbol{\eta} \sim N_7(\bar{\eta}, P_\eta^{-1})$$
$$\boldsymbol{\zeta} \sim N_9(\bar{\zeta}, P_\zeta^{-1}).$$

The mean vectors are indicated by $\acute{\eta}$ and $\acute{\zeta}$, with precisions P_η and P_ζ.

III. APPLICATION

A. Data

The daily rainfall data from the Urbino rain gauge station in Italy was used for modelling application. This dataset consists of daily rainfall observations, spanning the period 1981 – 2007 (27 years). In total, the dataset has 9861 observations. In this dataset, there was a low frequency of recorded rainfall, with inconsistent patterns of rainfall annually. On average, the most frequent rainfall occurred around November and January, while the least frequent rainfall took place in July and August.

B. Fitting the Model

The RJAGS package [23] was utilised to generate the posterior samples and this was applied in R software [24]. In our MCMC algorithm, the first 1000 iterations were discarded as a burn-in. The subsequent 20000 iterations were then obtained as the posterior samples. In total, 5 to 6 hours of computing time were required to gain 20000 posterior samples for all models using 3.40GHz Ergo Desktop AS4 All-in-One with Intel Core i7-3770 processor with 8 Gbytes of random-access memory (RAM).

It was found that 20000 iterations were adequate to obtain the realisations from the posterior distribution since the chain had satisfactorily converged after the initial burn-in period based on the analysis of trace plots. Tables I, II, III and IV illustrate the posterior means and standard deviations of the unknown parameters for both amount and occurrence processes.

Fig. 2 represents the posterior mean of the potential rainfall amount μ_t for the three proposed distributions for the amount process. It can be notably observed that the posterior means of the potential daily rainfall amounts are different for lognormal and gamma distributions. However, the results are very similar for the G1 and G2 distributions. Besides, the fitted values for the lognormal distribution were consistently larger than the fitted values for the G1 and G2 distributions. We also observed the presence of the seasonal variabilities throughout the year, based on the smooth plots of the fitted values. We found that the highest potential rainfall amounts in Urbino, Italy might occur between July and October based on the lognormal, G1 and G2 distributions. On the contrary, based on the findings using the lognormal and G2 distributions, the lowest degree of potential rainfall amount might occur between November and February. This is in contrast to the findings based on G1 distribution since the lowest degree of potential rainfall amounts might take place between January and March.

Table I: The Prior And Posterior Means And Standard Deviations (Sds) Of The Unknown Parameters' Using Lognormal Distribution For The Amount Process

Lognormal distribution (LN)				
Parameter	Prior mean	Prior SD	Posterior mean	Posterior SD
τ	0.54	0.27	0.45	0.01
η_0	1	0.74	1.04	0.03
a_1	0	0.80	-0.21	0.04
a_2	0	0.56	-0.07	0.04
a_3	0	0.46	0.05	0.04
b_1	0	0.80	-0.13	0.04
b_2	0	0.56	0.13	0.04
b_3	0	0.46	-0.08	0.04

In Fig. 3, the plot of posterior mean of the median potential rainfall amounts obtained using the lognormal, G1 and G2 distributions against time is presented. We could clearly observe that the posterior means of the median potential rainfall amounts based on the G1 and G2 distributions are higher than the ones obtained using the lognormal distribution. In this case, we postulate that this difference can be attributed to the different shapes of gamma and lognormal distributions used for our amount process model.

Table Ii: The Prior And Posterior Means And Standard Deviations (Sds) Of The Unknown Parameters' Using G1 Distribution For The Amount Process

Gamma 1 distribution (G1)				
Parameter	Prior mean	Prior SD	Posterior mean	Posterior SD
α	0.64	0.31	0.66	0.01
η_0	1.95	1.03	1.94	0.02
a_1	0	0.80	-0.12	0.03
a_2	0	0.56	-0.03	0.03
a_3	0	0.46	0.01	0.03
b_1	0	0.80	-0.16	0.03
b_2	0	0.56	0.05	0.03
b_3	0	0.46	-0.09	0.03

Table Iii: The Prior And Posterior Means And Standard Deviations (Sds) Of The Unknown Parameters' Using G2 Distribution For The Amount Process

Gamma 2 distribution (G2)				
Parameter	Prior mean	Prior SD	Posterior mean	Posterior SD
β	0.09	0.04	0.10	0.003
η_0	1.95	1.03	1.94	0.02
a_1	0	0.80	-0.10	0.02
a_2	0	0.56	-0.04	0.02
a_3	0	0.46	0.03	0.02
b_1	0	0.80	-0.06	0.02
b_2	0	0.56	0.07	0.02
b_3	0	0.46	-0.04	0.02

Table Iv: The Prior And Posterior Means, And Standard Deviations (Sds) Of The Unknown Parameters For The Occurrence Process

Occurrence process								
Parameter	Prior mean	Prior SD	Posterior mean			Posterior SD		
			LN	G1	G2	LN	G1	G2
ζ_0	-0.96	0.33	-0.57	-0.57	-0.57	0.02	0.02	0.02
ζ_1	0.82	0.35	0.52	0.55	0.54	0.30	0.30	0.30
ζ_2	1.76	0.88	1.45	1.45	1.45	0.05	0.04	0.05
c_1	0	0.80	0.39	0.35	0.34	0.07	0.05	0.05
c_2	0	0.56	-0.11	-0.13	-0.13	0.05	0.04	0.04
c_2	0	0.46	0.01	0.03	0.02	0.04	0.04	0.04
d_1	0	0.80	0.06	0.09	0.04	0.05	0.06	0.04
d_2	0	0.56	-0.31	-0.27	-0.28	0.06	0.04	0.04
d_3	0	0.46	0.06	0.06	0.04	0.05	0.05	0.04

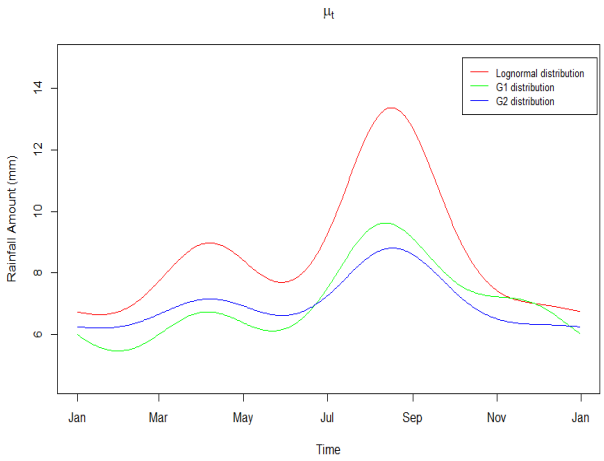


Fig. 2. The posterior mean of the mean potential rainfall amount based on three distinct (lognormal, G1 and G2) distributions

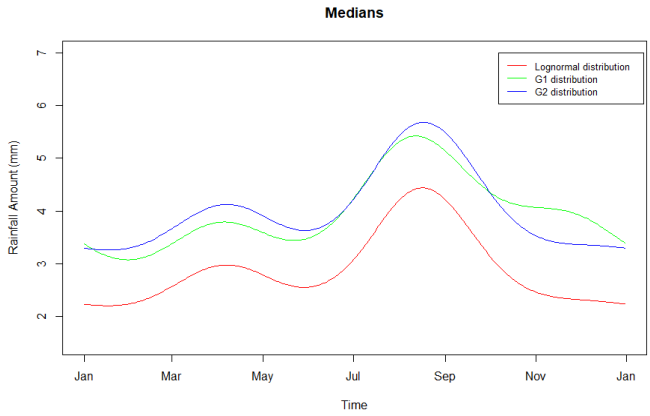


Fig. 3. The posterior mean of the median potential rainfall amounts based on the three distinct distributions

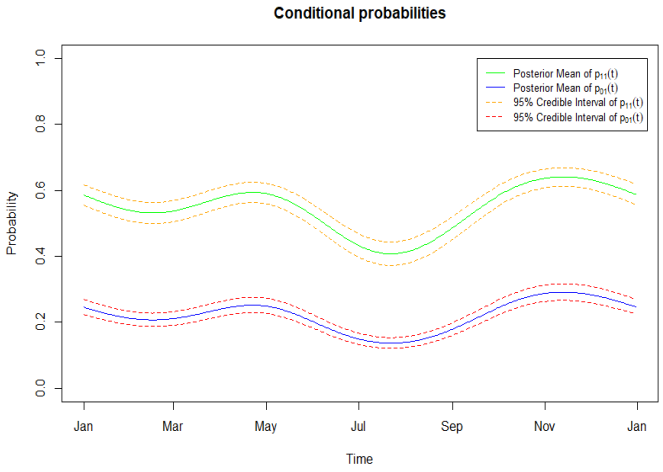


Fig. 4. The posterior mean of conditional possibilities, p_{01} and p_{11}

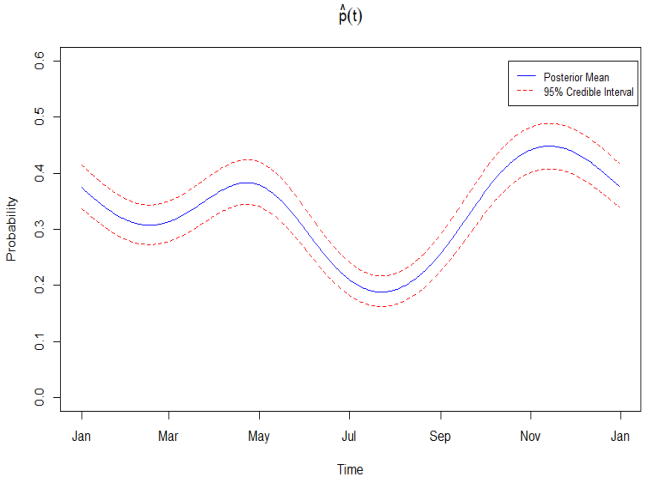


Fig. 5. The posterior mean of unconditional possibilities

Based on Fig. 4, a significant difference was also found in terms of the posterior means of the transition probabilities p_{01} and p_{11} for the occurrence process, regardless of the types of distributions used to model the amount process. To be specific, the probability of rainfall would be recorded in the present day is between 0.2 and 0.3, provided the rain was absent on the day before. Meanwhile, the probability of rainfall would exceed 0.5, provided the rain was present on the previous day. Furthermore, both plots were significant indicators of the rainfall possibility's relative dependence when it rained on the previous day. Based on Fig. 4, we could clearly observe that the conditional probabilities of rainfall were lowest from June to September. These findings are in contradiction to the findings obtained for the amount process since higher posterior means of rainfall amounts were observed between June and September (Fig. 2). For further analysis, we can also obtain the unconditional rainfall probabilities using Eq. (10) and these are presented in Fig. 5. On the whole, the probability of rainfall is lower than 0.5 throughout the year. This indicates the scarcity of rainfall in most time of the year in Urbino, Italy which explains the high number of zero observations in our dataset.

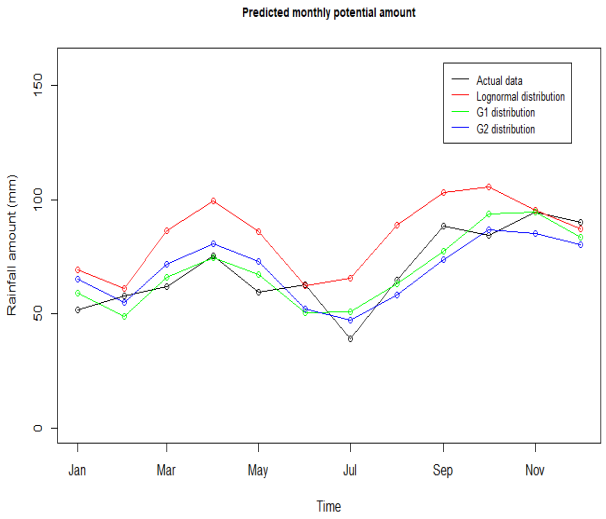


Fig. 6. The posterior mean for the predictive distribution of potential rainfall amounts every month

To investigate our model's validity, we can use posterior predictive checking method as recommended by Gelman et al. [25]. Based on this method, the model's fit is considered satisfactory if the observed data are similar the simulated values drawn from the replicated data's posterior predictive distributions [25]. Based on Fig. 6, similar patterns can be observed between the actual means of monthly rainfall and the posterior predictive means of rainfall throughout the year, especially for the G1 and G2 distributions. In the case of the lognormal distribution, however, more significant deviations were seen between the posterior predictive means of rainfall amount and actual means of monthly rainfall compared to the G1 and G2 distributions. To be more specific, for lognormal distribution, the maximum recorded amount of rainfall was found to occur in April based on the its posterior predictive means whilst this occurs in November for G1 and G2 distributions which is clearly in agreement with the actual data.

The rationale behind our findings can be attributed to the fact that the shape parameter of our gamma distribution was less than 1. Consequently, the gamma density did not correspond to zero when the actual rainfall amount is zero. On the other hand, the lognormal distribution density always goes to zero when the actual rainfall amount is zero. Hence, the shape of gamma distribution is always different from lognormal distribution especially when the values are close to the origin. This resulted in the challenges for the lognormal distribution to develop a good fit to such component of the distribution. This will in turn definitely affect the posterior values of the model parameters' summary statistics.

1) Residuals

The assumption that Y_t's are independent given the model parameters had been made for our model and this will be investigated by analysing the residuals. To achieve this, several transformations on y_t were performed. Let $u_t = G_t(y_t)$, where $G_t(.)$ represents the cumulative distribution function (cdf) of Y_t. In this case, $G_t(.)$ is the standard normal cdf $\Phi^{-1}(.)$. We then use the following relationship, $\hat{d}_t = \Phi^{-1}(u_t)$, to obtain $\hat{d}_t$.

We subsequently developed partial autocorrelation function (PACF) and the autocorrelation function (ACF) for these residuals and these are presented in the ACF and PACF plots in Fig. 7. It is clearly evident that the assumption of conditional independence between Y_t and Y_s ($s \neq t$) for this model, given μ_t and μ_s respectively, is not substantial. Therefore, it is desirable that further research is required in the future to develop new models that take into consideration the temporal correlation between the observations in the amount process.

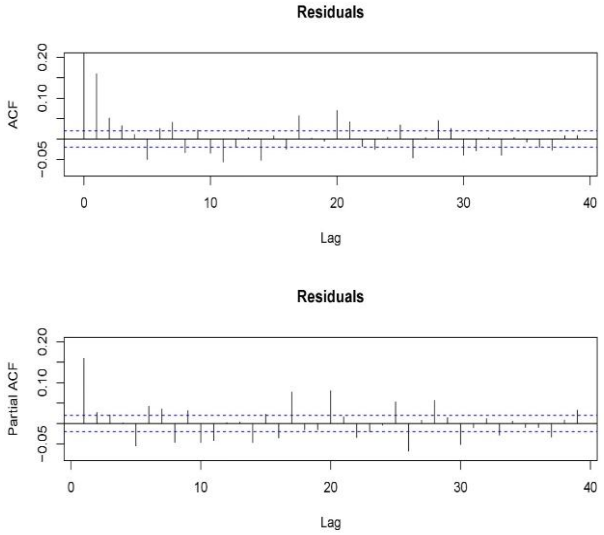
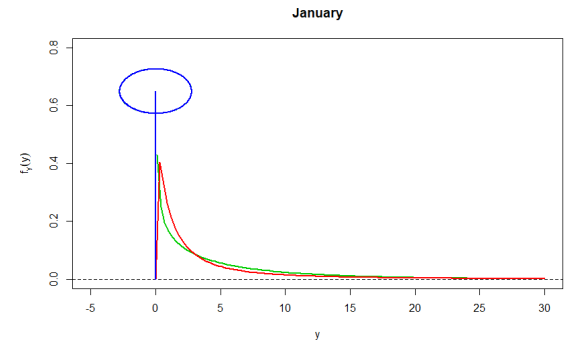


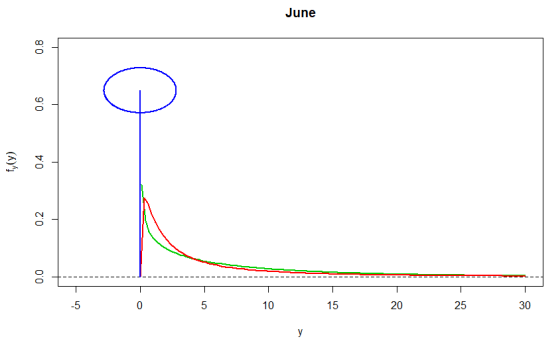
Fig. 7. The ACF and PACF plots for residuals

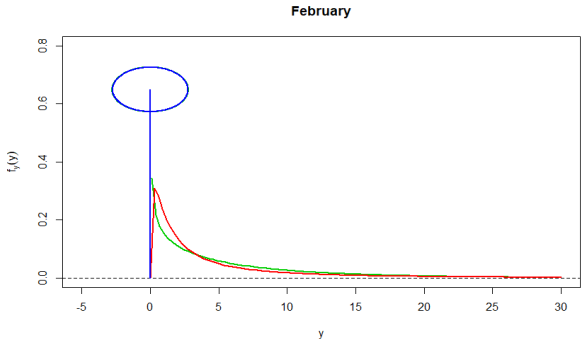
2) Zero/positive Distribution

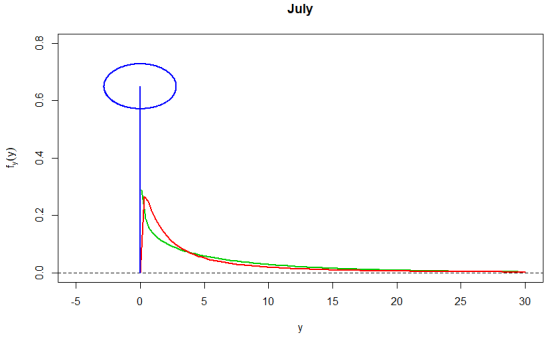
A mixed distribution could be illustrated with the use of the "lollipop" component at zero for the zero-rainfall probability and a density curve for the probability density function of the non-zero rainfall amount. These two components are then linked together, and the ellipse area represents to the probability of zero rainfall. Zero-positive plots are presented in Fig. 8 to illustrate the changes in the probability density function's shape and scale representing the amount of rainfall from month to month and the variability in the monthly rainfall pattern.

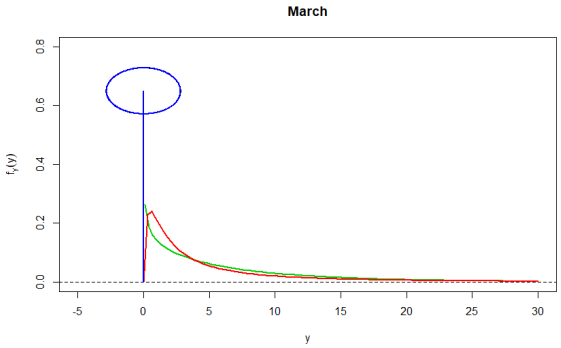
As a comparison, we used a single gamma distribution to represent the findings from G1 and G2 distributions due to the nearly identical posterior summary statistics were observed previously (Tables II and III, Figs. 2 and 3). Such findings would be subsequently compared with the ones obtained using the lognormal distribution. With respect to the gamma distribution, the distribution become more "*L*" shaped when the shape parameter is less than 1 as $f_y(y)$ approaches ∞. In contrast, $f_y(y)$ approaches 0 when $y = 0$ for the lognormal distribution. Overall, the distribution shape did not change significantly for the amount process from month to month. However, only scale parameter went through significant changes due to a relative reduction in the amount in August and September. Larger sizes of ellipses were observed in July and September due to infrequent rainfall.

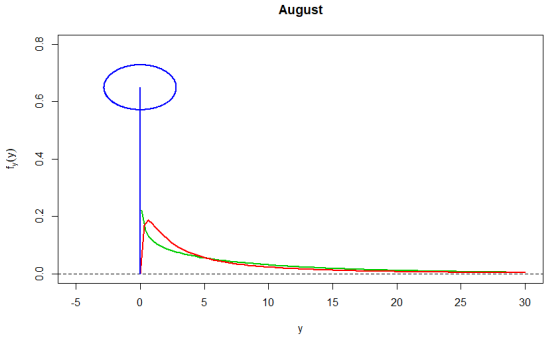

January
$f_Y(y)$
y

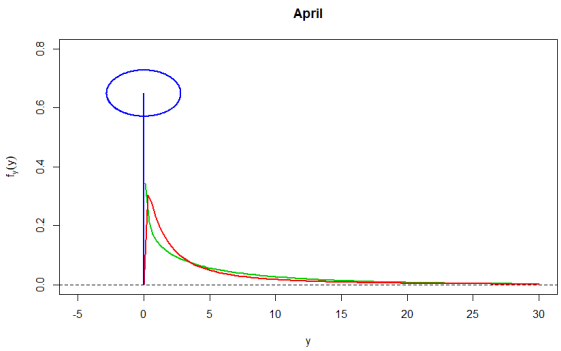

June
$f_Y(y)$
y

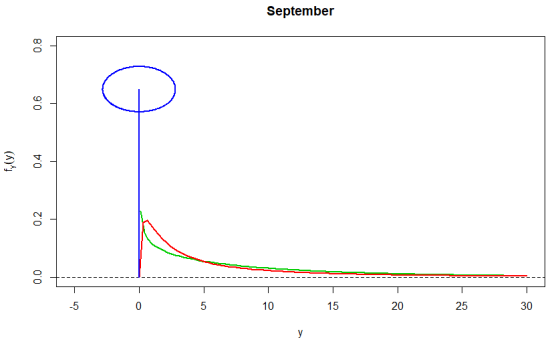

February
$f_Y(y)$
y

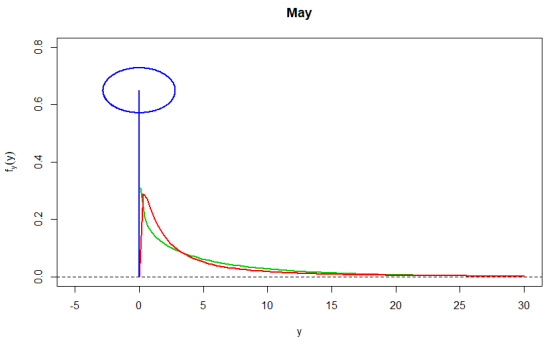

July
$f_Y(y)$
y

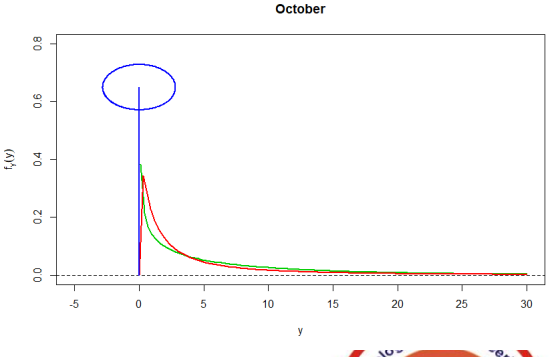

March
$f_Y(y)$
y


August
$f_Y(y)$
y


April
$f_Y(y)$
y


September
$f_Y(y)$
y


May
$f_Y(y)$
y


October
$f_Y(y)$
y

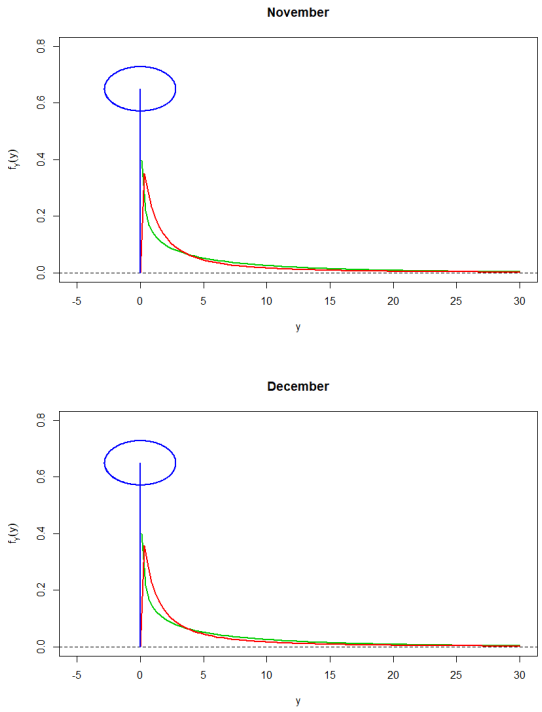


Fig. 8. The zero-positive plots for lognormal (red) and gamma (green) distributions representing the amount of rainfall from January to December

IV. CONCLUSION

In conclusion, it could be inferred from the findings that the gamma distribution is the best candidate for representing the amount process, as evident from its remarkable fit of the Italian rainfall data as presented in Fig. 6. To corroborate these findings, it had been verified in previous studies that gamma distribution is more superior to other distributions in terms of modelling daily rainfall amount. Fernandes et al. [9] found that the gamma distribution is more superior to either the lognormal or exponential distributions with respect to predictive properties. Furthermore, Bruno et al. [26] found that the gamma distribution is better than the lognormal distribution when it comes depicting the pattern of rainfall pattern in San Pietro Capofiume, Brazil spatiotemporally. It is also evidently demonstrated in our study that there was no remarkable change in the parameter shape over the year. In contrast, shifts could be clearly seen for the scale parameter when comparisons of rainfall amount were conducted on monthly basis. This may indicate that the parameterisation used in the G1 distribution is more accurate than the parameterisation utilised for the G2 distribution in the context of developing a daily rainfall model.

This study mainly aims to determine which model is the most compatible with the amount and occurrence processes in daily rainfall setting. For the amount process, three distinct distributions were investigated to ascertain the distribution that best fit this study's data. It was found that the G1 and G2 distributions are more accurate than the lognormal distribution in terms of representing the amount of daily rainfall which is evident from their more accurate posterior mean and posterior predictive values of the rainfall amount.

It was hoped that an extension of this model could be developed in future studies by integrating the atmospheric circulation patterns such as the Lamb weather types (LWTs) in the model. This is theoretically feasible since the LWTs may be directly incorporated into the models for both the amount and occurrence processes via the mean of amount distribution and rainfall probabilities.

REFERENCES

1. R. D. Stern and R. Coe, "A Model Fitting Analysis of Daily Rainfall Data," Journal of the Royal Statistical Society. Series A (General), vol. 147, p. 1, 1984.
2. J. A. Tooze, G. K. Grunwald and R. H. Jones, "Analysis of repeated measures data with clumping at zero," Statistical Methods in Medical Research, vol. 11, pp. 341-355, 2002.
3. R. Coe and R. D. Stern, "Fitting Models to Daily Rainfall Data," Journal of Applied Meteorology, vol. 21, pp. 1024-1031, 1982.
4. K. R. Gabriel and J. Neumann, "A Markov chain model for daily rainfall occurrence at Tel Aviv," Quarterly Journal of the Royal Meteorological Society, vol. 88, pp. 90-95, 1962.
5. G. K. Grunwald and R. H. Jones, "Markov models for time series with mixed distribution," Environmetrics, vol. 11, pp. 327-339, 2000.
6. R. W. Katz, "Precipitation as a Chain-Dependent Process," Journal of Applied Meteorology, vol. 16, pp. 671-676, 1977.
7. S. E. Heaps, R. J. Boys and M. Farrow, "Bayesian modelling of rainfall data by using non-homogeneous hidden Markov models and latent Gaussian variables," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 71, pp. 543-568, 2015.
8. B. Sanso and L. Guenni, "A stochastic model for tropical rainfall at a single location," Journal of Hydrology, vol. 214, pp. 64-73, 1999.
9. M. V. Fernandes, A. M. Schmidt and H. S. Migon, "Modelling zero-inflated spatio-temporal processes," Statistical Modelling, vol. 9, pp. 3-25, 2009.
10. D. S. Wilks, "Interannual variability and extreme-value characteristics of several stochastic daily precipitation models," Agricultural and Forest Meteorology, vol. 93, pp. 153-169, 1999.
11. R. E. Chandler and H. S. Wheater, "Analysis of rainfall variability using generalized linear models: A case study from the West of Ireland," Water Resources Research, vol. 38, pp. 10-1-10-11, 2002.
12. E. M. Furrer and R. W. Katz, "Generalized linear modeling approach to stochastic weather generators," Climate Research, vol. 34, pp. 129-144, 2007.
13. R. H. Jones and W. M. Brelsford, "Time series with periodic structure," Biometrika, vol. 54, pp. 403-408, 1967.
14. M. West and J. Harrison, Bayesian forecasting and dynamic methods, Springer-Verlag, 1997.
15. Y. Liu, W. Zhang, Y. Shao and K. Zhang, "A comparison of four precipitation distribution models used in daily stochastic models," Advances in Atmospheric Sciences, vol. 28, pp. 809-820, 2011.
16. C. W. Richardson, "Stochastic simulation of daily precipitation, temperature, and solar radiation," Water Resources Research, vol. 17, pp. 182-190, 1981.
17. J. Roldán and D. A. Woolhiser, "Stochastic daily precipitation models: 1. A comparison of occurrence processes," Water Resources Research, vol. 18, pp. 1451-1459, 1982.
18. J. Suhaila, K. Ching-Yee, Y. Fadhilah and F. Hui-Mean, "Introducing the Mixed Distribution in Fitting Rainfall Data," Open Journal of Modern Hydrology, vol. 01, pp. 11-22, 2011.
19. H.-K. Cho, K. P. Bowman and G. R. North, "A Comparison of Gamma and Lognormal Distributions for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission," Journal of Applied Meteorology, vol. 43, pp. 1586-1597, 2004.
20. B. Kedem, L. S. Chiu and G. R. North, "Estimation of mean rain rate: Application to satellite observations," Journal of Geophysical Research: Atmospheres, vol. 95, pp. 1965-1972, 1990.
21. T. Ben-Gai, A. Bitan, A. Manes, P. Alpert and S. Rubin, "Spatial and Temporal Changes in Rainfall Frequency Distribution Patterns in Israel," Theoretical and Applied Climatology, vol. 61, pp. 177-190, 01 12 1998.
22. O. D. Jimoh and P. Webster, "The optimum order of a Markov chain model for daily rainfall in Nigeria," Journal of Hydrology, vol. 185, pp. 45-69, 1996.
23. M. Plummer, Bayesian Graphical Models using MCMC, 3-7 ed., CRAN, 2012.

24. R. D. C. Team, "R: A Language and Environment for Statistical Computing," Vienna, 2008.
25. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, Bayesian Data Analysis, Third Edition, Taylor & Francis, 2013.
26. F. Bruno, D. Cocchi, F. Greco and E. Scardovi, "Spatial reconstruction of rainfall fields from rain gauge and radar data," Stochastic Environmental Research and Risk Assessment, vol. 28, pp. 1235-1245, 2014.

AUTHORS PROFILE

Muhammad Safwan Bin Ibrahim obtained his B.Sc. in Industrial Mathematics and M.Sc. in Mathematics from Universiti Teknologi Malaysia in 2010 and 2012, respectively. Then he obtained his PhD in Statistics from Newcastle University in 2018. Currently, he is lecturer in Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM). His current research interests are time series and statistical modelling.

Muhammad Irfan bin Abdul Jalal graduated with a Bachelor's Degree in Medicine and Surgery (MBChB, BAO) from Queen's University of Belfast in 2006. He subsequently obtained his Master's degree in Medical Statistics (MSc Medical Statistics) from Universiti Sains Malaysia (USM) in 2011. He recently successfully defended his PhD viva and is anticipated to be awarded with a PhD in Statistics from Newcastle University in December 2019. His main research interests are Bayesian statistics with applications in medical research, specifically in clinical trial design and survival analysis.