MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Lokoman *et al*. | Malaysian Journal of Fundamental and Applied Sciences, Vol. 17 (2021) 306-320

RESEARCH ARTICLE

Construction of Dependence Structure for Rainfall Stations by Joining Time Series Models with Copula Method

Rahmah Mohd Lokoman[a], Fadhilah Yusof[a,*], Nor Eliza Alias[b], Zulkifli Yusop[c]

[a] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; [b] Department of Hydraulics and Hydrology, Faculty of Civil Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia; [c] Centre for Environmental Sustainability and Water Security, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

Abstract Copula model has applied in various hydrologic studies, however, most analyses conducted does not considering the non-stationary conditions that may exist in the time series. To investigate the dependence structure between two rainfall stations at Johor Bahru, two methods have been applied. The first method considers the non-stationary condition that exists in the data, while the second method assumes stationarity in the time series data. Through goodness-off-fit (GOF) and simulation tests, performance of both methods are compared in this study. The results obtained in this study highlight the importance of considering non-stationarity conditions in the hydrological data.

Keywords: Bivariate copula, time series models, dependence modelling, non-stationary, rainfall.

Introduction

Hydrologic phenomena, such as drought, flood and rainfall, are natural phenomena that often provide dependent multivariate observations. For example, we can observe drought duration, intensity, and magnitude from a drought phenomenon. According to Salvadori and De Michele [1], those random variables each play an important role, where such an analysis of a joint distribution between the variables can identify the characteristics of drought. Therefore, it is necessary to find the joint distribution and estimate the dependence between the variables.

***For correspondence:** fadhilahy@utm.my

Received: 27 Jan 2019
Accepted: 2 August 2021

Recently, determining a joint distribution of variables using the copula method and constructing its dependence structure has become an interesting research in many fields, including finance, management, and hydrology. Consequently, through the copula method, the dependency impact of the variables in those fields can be analyzed. Sklar [2] has introduced the copula method or copula function as a function of combined distribution of two or more uniform marginal distribution. The copula method outlines the limitations of the common traditional joint distribution method as it allows us to determine the distribution function for the marginal variables. It also allows us to find the fitted copula from any different copula families to build the dependency structure of the marginal variables. The proof that this method can obtain a joint distribution of bivariate rainfall variables with different distribution functions of marginal is shown in the works of Zhang and Singh [3]. Their works were completed without any assumption that the variables were independent or normally distributed.

There are several copula families that can accommodate the scope of the broad dependency structure

that have been proposed and developed, for example, the Student's *t* family, the Gaussian family, and the Archimedean family. In the hydrological field, the copula families that have been most used for analyses are the Archimedean copula families. According to Zhang and Singh [3] and Nelsen [4], the Archimedean copula family, which usually have closed form, is very popular and desirable in constructing the dependence structure of the hydrologic variables. It is because of the ease in constructing the joint distribution functions, and they can be used when the correlation of the variables is either negative or positive.

Other than the copula method, statistical model called as time series model has also been widely applied to the analyses of hydrological variables. The time series model is used mainly to forecast future hydrological data. To build the time series model, the data must be in a stationary condition. However, hydrological data is a continuous time series where the observations are measured at every time. The conditional mean or variance of the time series can be varying or volatile over time. To counter the non-stationary problem, the time series models: Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models are often used in many areas of time series analysis. According to Yusof et al. [5], the ARIMA model is used to see the behavior of the mean from time series data, while the GARCH model is used to model the heteroscedasticity (variance behavior) of the residuals from the time series model of ARIMA.

In the finance field, volatile data are very common. Thus, the time series model is one of the statistical models that has been mostly applied. In addition, a combination model between time series model and copula model also has been developed and widely applied. Finance studies, such as Oliveria et. al. [6] and Oliveria et. al. [7], stated that the accuracy of the forecasting process can be significantly improved by combining the time series forecasting models. The combined time series models have better performance than the univariate time series model. Therefore, to combine the time series models, copula method has been applied. This combination model has brought benefit to the finance field as forecasting are the most important subject in finance. Furthermore, another advantage of this combination is that the time series models can be applied to the marginal variables to handle the non-stationary condition that may exist.

While, for hydrological field, although the time series and the copula models have been widely applied in statistical analysis of hydrological data, the combination model is not widely applied yet. Many hydrologic studies, such as Salvadori and De Michele [1], Nelsen [4], Ariff et. al. [8], Yusof et. al. [9], and Yee et. al. [10], have applied the copula method. However, the analyses only captured the statistical distribution of the marginal variable without considering the non-stationary condition that may exist in the hydrological time series. Therefore, to fill this gap, the time series model will be applied to model the hydrologic variables in this study.

The aim of this study is to apply copula models by considering the non-stationary conditions such as seasonal variation and volatility that may exist in the hydrologic variables. This application of combining time series models with copula method are widely used in financial econometrics and risk management, but not in hydrologic field. Most of the time, copulas are applied to hydrologic time series data, however, they often are treated as stationary over time.

Materials and methods

Data

Malaysia is located at the Southeast Asia and that is composed of two major regions, Peninsular Malaysia and East Malaysia. Located near the equator, Malaysia is hot and humid throughout the year. The climate in Malaysia is quite uniform, but it can change during the two different monsoon seasons that occur yearly. The first is the Northeast Monsoon (NEM), which takes place between November and February. It is characterized by the wind from northeast and heavy rain. The second monsoon is the Southwest Monsoon (SWM). Malaysia faces drier period with less rainfall during the SWM. In between these two monsoon seasons, there are also two inter-monsoon seasons. The inter-monsoons happen

between March and April and from September to October. They bring heavy rainfall and commonly happens in the convective form.

Johor is a state in Malaysia that is in the south of Peninsular Malaysia. Johor has a climate with consistent temperature. The temperature ranges from 23.8 °C to 30.5 °C on a cool day and from 25.0 °C to 32.2 °C on a hot day. It also has a high amount of rainfall, mostly from November to February. In 2017, according to local news [11], [12], [13], the NEM phenomenon which lasted from the end of November 2017 until April 2018 carried an enormous rainfall, causing a flood disaster. This is a common phenomenon that occurs every year, and consequently increases the average rainfall compared with prior months.

For this research, monthly rainfall data for 31 years (1981–2011) were taken from two rain gauge stations located in Johor: Station Ldg. Lim Lim Bhd., Masai, and Station Ldg. Sg. Tiram, Johor Bahru. The data are provided by the Department of Irrigation and Drainage Johor. The distance between the two rainfall stations is 1.4 km and the location of these two stations are shown in Figure 1. Station Ldg. Lim Lim Bhd., Masai is renamed as Station M and Station Ldg. Sg. Tiram, Johor Bahru is renamed as Station T.

Figure 1. Location of Station M and T in Johor.

Copula Model

Based on Nelsen [3], for random variables Y_1 and Y_2, the equation of joint cumulative distribution function (CDF) in the form of the function of copula can be written as below in Equation 1:

$$C(v_1, v_2;\ \beta_1, \beta_2,\ \theta_c) = C_\theta\left[F_{Y_1}(y_1\,;\beta_1), F_{Y_2}(y_2\,;\beta_2)\ ;\ \theta_c\right] = H(y_1, y_2\,;\ \beta_1, \beta_2,\ \theta_c) \tag{1}$$

where $F_{Y_1}(y_1\,;\beta_1) = v_1$ and $F_{Y_2}(y_2\,;\beta_2) = v_2$ are the CDF of Y_1 and Y_2 respectively and β_1 and β_2 as the marginal parameters for Y_1 and Y_2 respectively. Meanwhile, θ_c is the copula dependence parameter.

The equation of joint probability density function (PDF) of copula is written as in Equation 2:

$$\begin{aligned} &h(v_1, v_2;\ \beta_1, \beta_2,\ \theta_c) \\ &\quad = \frac{\partial^2}{\partial y_1 \partial y_2} H(y_1, y_2\,;\ \beta_1, \beta_2,\ \theta_c) \\ &\quad = c\left[F_{Y_1}(y_1\,;\beta_1), F_{Y_2}(y_2\,;\beta_2); \theta\right] \cdot f_{Y_1}(y_1\,;\beta_1)\ \cdot f_{Y_2}(y_2\,;\beta_2) \end{aligned} \tag{2}$$

Where

$$c\left[F_{Y_1}(y_1\,;\beta_1), F_{Y_2}(y_2\,;\beta_2)\ ;\theta\right] = c(v_1, v_2;\ \theta_c) = \frac{\partial^2}{\partial v_1 \partial v_1} C(v_1, v_2;\ \theta_c) \tag{3}$$

Equation 3 is the PDF of the copula function and $f_{Y_1}(y_1\,;\beta_1)$ and $f_{Y_2}(y_2\,;\beta_2)$ are the PDF of random variables of Y_1 and Y_2 respectively.

Time series modeling

ARIMA Model

A time series of autoregressive moving average (ARMA) model, ARMA (*p, q*) is formulated as in Equation 4:

$$y_t = \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} \quad (4)$$

where $\varphi_p \neq 0$, $\theta_q \neq 0$ and ε_t with $N(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 > 0$. The parameter *p* indicates the autoregressive [AR(*p*)] order and *q* indicates the moving average [MA(*q*)] order.

There are two conditions that exist in a time series data, non-stationary and stationary. A differencing process is applied to make a non-stationary time series to become a stationary time series. If no differencing order is needed, the ARMA model is applied to the stationary time series. Meanwhile, if the differencing order is needed, a time series called Autoregressive Integrated Moving Average (ARIMA) is applied. The ARIMA model is written as ARIMA (*p, d, q*). The *d* in ARIMA (*p, d, q*) indicates the differencing order.

To detect the *p* and *q* orders, the process of inspecting both partial autocorrelation function (PACF) and autocorrelation function (ACF) is very helpful. In addition, to find the best time series model, an information criterion called the Akaike Information Criterion (AIC) is used in this research. The best time series model is chosen based on the lowest value of AIC.

Ljung-Box Series Test

After either the ARMA or ARIMA model has been chosen, the residuals from the model need to be inspected whether it shows an autocorrelation. The examination can be done by using Ljung-Box test. Essentially, the result from test will show whether the chosen model is fitted and adequate for the time series data. After the adequate model is found, Autoregressive Conditional Heteroscedastic (ARCH) effect is examined before ARCH or GARCH model construction.

Box-Jenkins Steps

The methodology that have been explained before for ARMA/ARIMA time series modelling can be summarized as a Box-Jenkins steps. The steps are summarized as follow.

Step 1: Check the stationarity of the time series by using Augmented Dickey Fuller (ADF) test.
Step 2: Identify the order of *p* and *q* by observing the plots of ACF and PACF
Step 3: Estimate the parameters for the ARIMA models and calculate the AIC value. Choose the best model based on the smallest value of AIC.
Step 4: Check the adequacy of the selected ARIMA model by using the Ljung-Box test.

ARCH Effect Inspection

Before constructing the time series model ARMA-GARCH, or ARIMA-GARCH, the squared residuals of ARMA/ARIMA time series model need to be checked first to determine whether the squared residuals show autocorrelation. This checking process is simply called as the ARCH effect inspection. The significance of ARCH effect inspection is done by using McLeod-Li and Ljung-Box Series test. If the test shows that the squared residuals are autocorrelated, then the ARCH effect is present.

GARCH Modeling

Bollerslev [14] has generalized ARCH to become Generalized Autoregressive Conditional Heteroscedastic (GARCH) model (Kane and Yusof [15]) and (Chen et al. [16]). The variance equation of the GARCH (p,q) model can be referred to Equation 5 as follows:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 = \omega + \alpha(B)\varepsilon_t^2 + \beta(B)\sigma_t^2 \quad (5)$$

where $\varepsilon_t = z_t\sigma_t \ , z_t \sim \psi_\tau(0,1)$

is the residuals of ARMA model and $\psi_\tau(0,1)$ is the PDF of the residuals with 0 mean and variance 1. τ is the parameter that define the distribution shape. While, σ_t^2 the estimated conditional variance, and $\alpha(B)$ and $\beta(B)$ are polynomials in the backshift operator *B*.

Information Criteria and Statistical Goodness of Fit Test

The AIC is used as the information criteria to select the best time series forecasting model and as the statistical goodness of fit (GOF) test to select the best fitted marginal and copula distribution models. The time series forecasting model, the marginal and copula distribution models for the rainfall variables that has the smallest value of AIC will be chosen as the best fitted model.

Equation 6 shows the equation for AIC use in time series modelling:

$$AIC = 2m - 2\ln(\hat{L}) \tag{6}$$

where m is the number of parameters in the statistical model, *n* is the observations number and $\hat{L}$ is the maximum value of the likelihood function for the model.

As for the copula model, the AIC values can be obtained by calculating the maximum likelihood of the copula log-likelihood model. Therefore, the equation for AIC use in copula modelling is as in Equation 7:

$$AIC_C = 2m - 2\sum_{i=1}^{n} \ln c[\hat{F}_{Y_1}(y_{1i}), \hat{F}_{Y_2}(y_{2i})\,;\, \hat{\theta}_c] \tag{7}$$

where $\hat{\theta}_c$ is the estimated copula dependence parameter, $\hat{F}_{Y_1}(y_{1i})$ and $\hat{F}_{Y_2}(y_{2i})$ are the values of the estimated cumulative distribution at y_{1i} and y_{2i} respectively, and m is the number of parameters in the copula model.

Empirical Study

In this study, two methods were used to find the fitted copula. The first method is the combination of time series and copula models. The time series modelling which is applied first to the rainfall data. After that, the residuals of the time series models are taken as the marginal variables for the Copula function. Two types of marginal distributions which are the Normal and Student's *t* distributions are considered in fitting the residuals data. The second method is only the copula model was applied and the rainfall data are taken as the marginal variables for the Copula function. Three marginal distributions: Normal, Gamma, and Weibull distributions are considered in fitting the rainfall data. The frameworks for both methods are is illustrated in Figure 2. The bivariate copulas that have been used in this study are listed in Table 1.

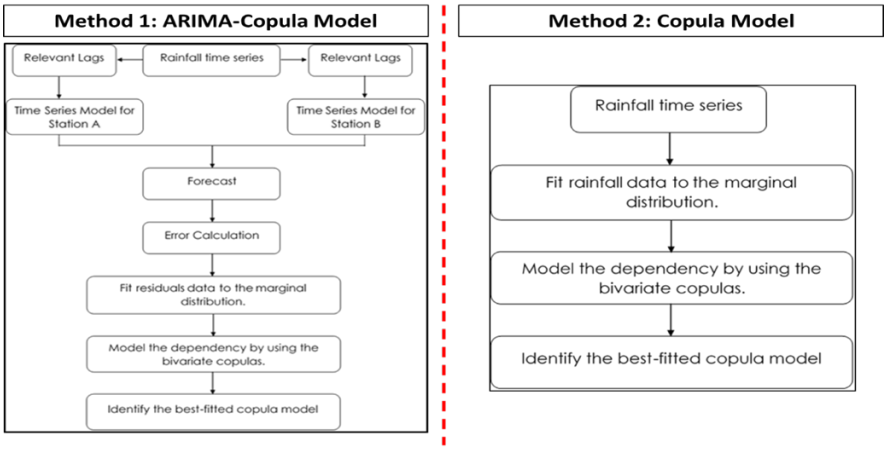

Figure 2. The framework for Method 1 and 2.

Table 1. Properties of elliptical and Archimedean copula.

Copula	Distribution functions of Copula, $C(v_1, v_2; \theta_c)$	Parameter, θ_c
Gaussian	$\int_{-\infty}^{\phi^{-1}(v_1)} \int_{-\infty}^{\phi^{-1}(v_2)} \frac{1}{2\pi(1-{\theta_c}^2)^{0.5}} \exp\left\{-\frac{{y_1}^2 - 2y_1y_2\theta_c + {y_2}^2}{2(1-\theta_c)^2}\right\} dy_1\, dy_2$	$\theta_c \in [-1,1]$
Student's *t*	$\int_{-\infty}^{t^{-1}(v_1)} \int_{-\infty}^{t^{-1}(v_2)} \frac{1}{2\pi(1-{\theta_c}^2)^{0.5}} \left\{1 + \frac{{y_1}^2 - 2y_1y_2\theta_c + {y_2}^2}{(1-\theta_c)^2}\right\}^{-(r+1)/2} dy_1\, dy_2$	$\theta_c \in [-1,1]$
Clayton	$({v_1}^{-\theta_c} + {v_2}^{-\theta_c} - 1)^{-1/\theta_c}$	$\theta_c \geq -1$
Frank	$-\frac{1}{\theta_c} \ln\left[1 + \frac{(e^{-\theta_c v_1} - 1)(e^{-\theta_c v_2} - 1)}{e^{-\theta_c} - 1}\right]$	$\theta_c \neq 0$
Gumbel-Hougaard	$\exp -\left[(-\ln v_1)^{\theta_c} + (-\ln v_2)^{\theta_c}\right]^{(1/\theta_c)}$	$\theta_c \geq 1$

Simulation

After the fitted copulas have been identified in both methods, simulation of random data is conducted. The simulation is repeated for 500 times. The simulation of randomly generated data will aid in finding the accuracy of the fitted distribution. The accuracy is calculated by using a performance measure, such as root mean squared error (RMSE). Furthermore, from the simulation process, the justification is also done based on the visualization of comparing real data and generated data.

Results and discussion

Patterns and rainfall distribution between neighboring stations are often found to be the same and have significant correlations. An individual rainfall station analysis may not show the correlation which may exist. Therefore, bivariate or multivariate statistical analysis is needed to identify the dependence correlation. The analysis can be done to by modeling together the neighboring stations together with a joint distribution. Identification of the joint distribution is very important in some hydrological works. For example, from the fitted joint distribution, simulation of a runoff in a river basin, where both the stations belong to can be done.

The rainfall data used in this study is from two rain gauge stations in Johor. The rainfall stations are Station Ldg. Lim Lim Bhd., Masai and Station Ldg. Sg. Tiram, Johor Bahru. Their descriptive statistics are presented in Table 2.

Table 2. Descriptive statistics of the monthly rainfall for Station M and Station T.

Descriptive statistics	Station M	Station T
Minimum (mm)	0.00	0.00
Maximum (mm)	940.00	634.80
Mean (mm)	9.38	7.91
Standard Deviation (mm)	116.52	116.52

Table 2 shows the descriptive statistics measures for the monthly rainfall at Station M and Station T. The statistics measures include the minimum, maximum, mean, and standard deviation of the rainfall data. The minimum monthly rainfall for both stations is 0.00 mm. Zero value shows that there was no rainfall has happened in that month. Meanwhile, the highest monthly rainfall collected at Station M is 940.00 mm. Although the distance between Station M and T is 1.4km, the maximum rainfall amount at Station M is higher compared with the rainfall amount at Station T. The maximum rainfall at Station T is only about 634.80 mm. As for average rainfall, the average monthly rainfall at Station M is also higher than Station T, with 9.38 mm and 7.91 mm respectively. However, the standard deviation of the monthly rainfall for each station is same, 116.52 mm. This means that the variability of the monthly rainfall is same for both stations although the average monthly rainfall at Station M are slightly higher than Station T.

Correlation Between the Rainfall Data

The correlation between the rainfall data from Station M and Station T is shown in the scatter plot in Figure 3 below.

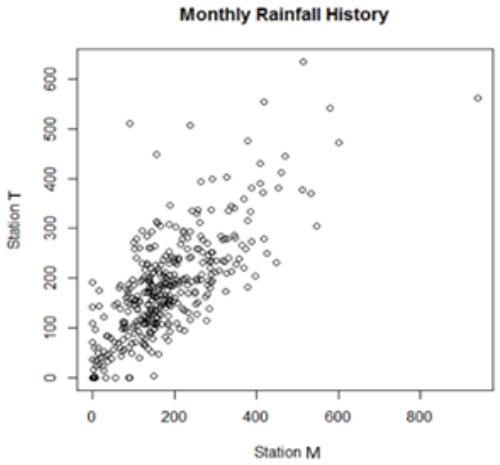

Figure 3. Scatter plot of the monthly rainfall at Station M and T in millimeter unit (mm).

Based on the observation in Figure 3, the rainfall at Station M is positively correlated with the rainfall at Station T. Kendall's tau value was used to measure the correlation between the rainfall data of both stations. The correlation level or the value of Kendall's tau of the two rainfall stations is 0.5179 with *p*-values equal to 0.000 at the significance level of $\alpha = 0.05$. Thus, the correlation between these two rainfall stations is statistically significant because the *p*-value is less than 0.05.

Method 1: ARIMA-Copula Model

In the first method, the non-stationary condition that exist in the data has been considered. To consider non-stationary condition, time series modelling is applied first to the rainfall data. After that, the residuals of the time series models are taken as the marginal variables for the Copula function. Two types of marginal distributions: Normal and Student's *t* distributions were considered in fitting the residuals data.

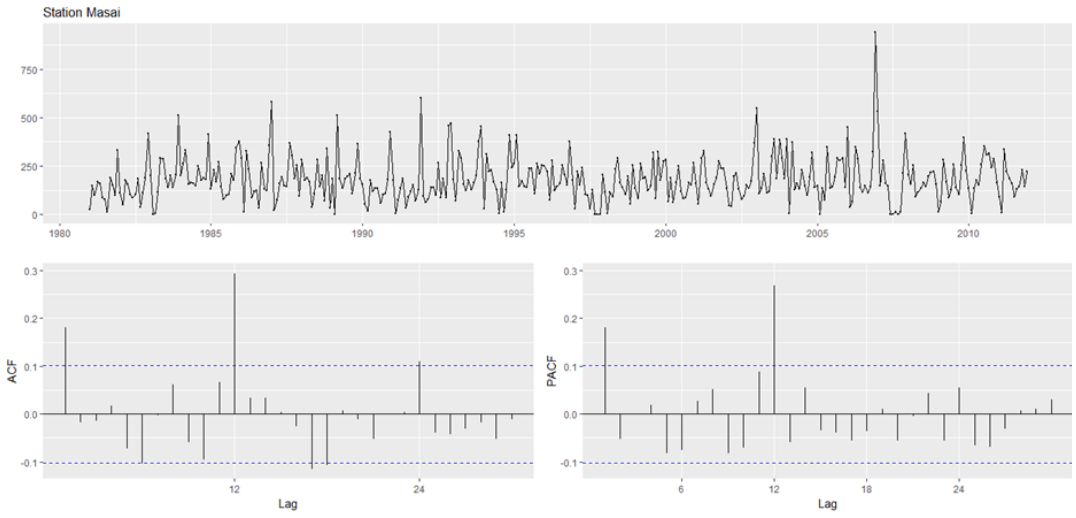

Figure 4. The plot of time series, ACF and PACF for rainfall at Station M.

Time series models for rainfall at Station M and T

To identify the tentative model, we need to do visual inspection for both ACF and PACF. For Box-Jenkins ARIMA method, the required model is ARIMA (*p,d,q*). Therefore, we can obtain autoregressive order, AR (*p*) and moving average order, MA (*q*) by using PACF and ACF plots respectively. Figure 4 shows the plot of time series, ACF and PACF for rainfall at Station M and Figure 5 for Station T.

Based on Figure 4, there are spikes at lag 12 and 24 in the ACF plot and a spike at lag 12 in the PACF plot. This results shows that the suitable time series model is seasonal ARIMA$(p,d,q)(P,D,Q)_{12}$ model (SARIMA). The ACF and PACF have a cut off after the 1st lag which indicating no differencing is needed. Thus, the order(s) of (*p,d,q*) are: $p = 0$ and 1, $d = 0$, and $q = 0$ and 1 and the order(s) of (*P,D,Q*) are: $P = 0$ and 1, $D = 0$, and $Q = 0$, 1 and 2.

There are 18 possible candidates of SARIMA model that can be considered for Station M. After the model and parameters have been identified and estimated, diagnostic checking is then applied to check the adequacy of both models. Table 3 shows the significance of T-test for model parameters, the adequacy and the AIC value of 18 possible candidates of SARIMA model.

Table 3. Results of time series modelling for Station M

Possible candidate SARIMA model	T-test for parameters	Adequacy	AIC
(0,0,0)(0,0,1)[12]	Significant	No	4570.35
(0,0,0)(0,0,2)[12]	Not significant	No	4569.64
(0,0,0)(1,0,0)[12]	Significant	No	4566.87
(0,0,0)(1,0,1)[12]	Not significant	No	4568.20
(0,0,0)(1,0,2)[12]	Significant	No	4549.15
(0,0,1)(0,0,1)[12]	Significant	Yes	4560.25
(0,0,1)(0,0,2)[12]	Not significant	Yes	4558.33
(0,0,1)(1,0,0)[12]	Significant	Yes	4555.89
(0,0,1)(1,0,1)[12]	Not significant	Yes	4556.81
(0,0,1)(1,0,2)[12]	Significant	Yes	4542.82
(1,0,0)(0,0,1)[12]	Significant	Yes	4560.72
(1,0,0)(0,0,2)[12]	Not significant	Yes	4558.91
(1,0,0)(1,0,0)[12]	Significant	Yes	4556.39
(1,0,0)(1,0,1)[12]	Not significant	Yes	4557.27
(1,0,0)(1,0,2)[12]	**Significant**	**Yes**	**4542.41**
(1,0,1)(0,0,1)[12]	Not significant	Yes	4562.25
(1,0,1)(0,0,2)[12]	Not significant	Yes	4560.33
(1,0,1)(1,0,0)[12]	Not significant	Yes	4556.41

As shown in Table 3, ARIMA(1,0,0)(1,0,2)[12] is selected as the best model to forecast Station M as the parameters of the model are significant, adequate, and has the smallest AIC value.

For Station T, Figure 5 shows there are spikes at lag 12 for both ACF and PACF plots. This results also shows that the suitable time series model is seasonal ARIMA$(p,d,q)(P,D,Q)_{12}$ model (SARIMA). The ACF and PACF have a cut off after the 1st lag, indicating no differencing is needed. Thus, the order(s) of (*p,d,q*) are: $p = 0$ and 1, $d = 0$, and $q = 0$ and 1 and the order(s) of (*P,D,Q*) are: $P = 0$ and 1, $D = 0$, and $Q = 0$, and 1.

There are 12 possible candidates of SARIMA model that can be considered for Station T. After the model and parameters have been identified and estimated, diagnostic checking is then applied to check the adequacy of both models. Table 4 shows the significance of T-test for model parameters, the adequacy and the AIC value of the 12 possible candidates of SARIMA model that can be considered for Station T.

From Table 4, ARIMA(0,0,1)(1,0,0)[12] is selected as the best model to forecast Station T as the parameters of the model are significant, adequate and has the smallest AIC value. Both time series models are then tested for ARCH effect by using McLeod Li and Ljung-Box tests. Figures 6 and 7 display

the results of McLeod Li and Ljung-Box tests, respectively, for Station M, while Figures 8 and 9 display the results of McLeod Li and Ljung-Box tests, respectively, for Station T.

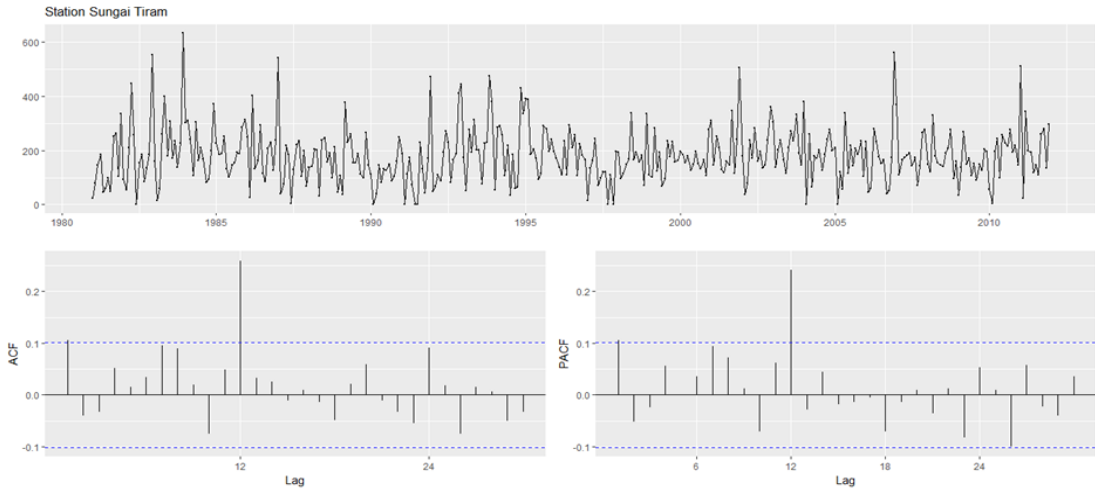

Figure 5. The plot of time series, ACF and PACF for rainfall at Station T.

Table 4. Results of time series modelling for Station T.

Possible candidate SARIMA model	T-test for parameters	Adequacy	AIC
(0,0,0)(0,0,1)[12]	Significant	No	4485.01
(0,0,0)(1,0,0)[12]	Significant	No	4482.12
(0,0,0)(1,0,1)[12]	Not significant	No	4483.85
(0,0,1)(0,0,1)[12]	Not significant	Yes	4483.00
(0,0,1)(1,0,0)[12]	**Significant**	**Yes**	**4479.94**
(0,0,1)(1,0,1)[12]	Not significant	Yes	4481.53
(1,0,0)(0,0,1)[12]	Not significant	Yes	4483.34
(1,0,0)(1,0,0)[12]	Not significant	Yes	4480.26
(1,0,0)(1,0,1)[12]	Not significant	Yes	4481.82
(1,0,1)(0,0,1)[12]	Not significant	Yes	4484.82
(1,0,1)(1,0,0)[12]	Not significant	Yes	4481.80
(1,0,1)(1,0,1)[12]	Not significant	Yes	4483.43

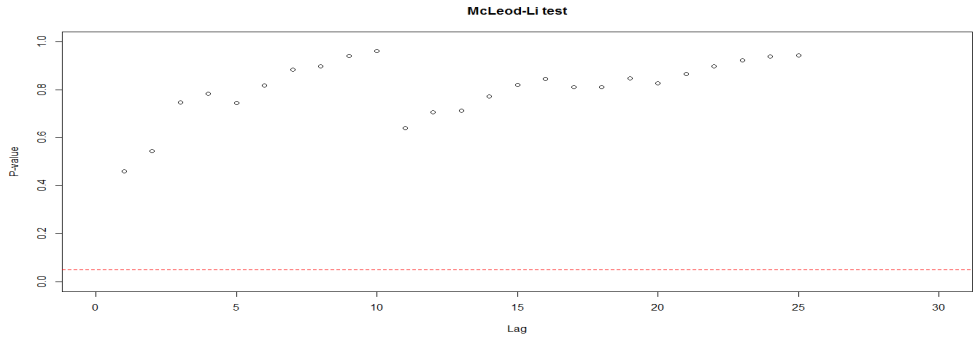

Figure 6. P-value plot of McLeod Li test for squared residuals of ARIMA(1,0,0)(1,0,2)[12] model.

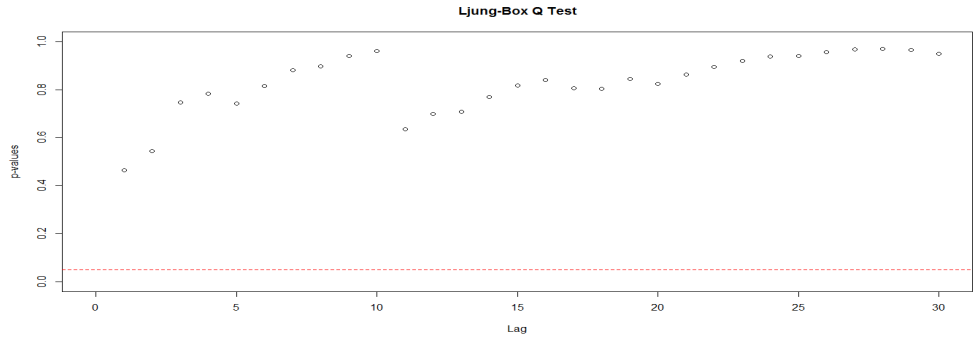

Figure 7. P-value plot of Ljung-Box test for squared residuals of ARIMA(1,0,0)(1,0,2)[12] model.

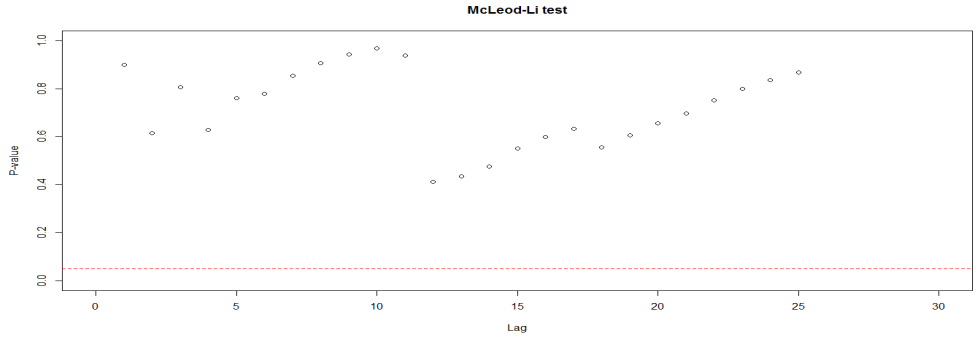

Figure 8. P-value plot of McLeod Li test for squared residuals of ARIMA(0,0,1)(1,0,0)[12] model.

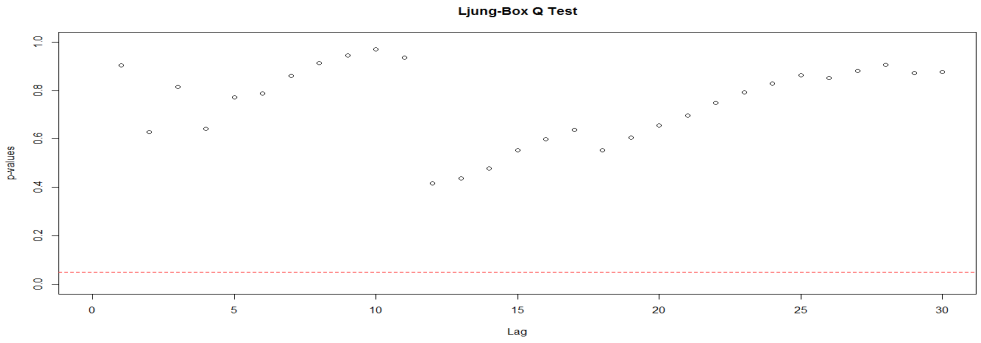

Figure 9. P-value plot of Ljung-Box test for squared residuals of ARIMA(0,0,1)(1,0,0)[12] model.

Figure 6 - 9 show that SARIMA models ARIMA(1,0,0)(1,0,2)[12] and ARIMA(0,0,1)(1,0,0)[12] have no ARCH effect since all the p-value of the lags are bigger that $\alpha = 0.05$. Therefore, GARCH modelling is not needed. Table 5 summarizes the best time series model for rainfall data of Station M and T.

Table 5. Time series models for Station M and T

Rainfall Stations	Model
M	ARIMA(1,0,0)(1,0,2)[12]
T	ARIMA(0,0,1)(1,0,0)[12]

Marginal Distributions for Residuals

Two types of marginal distributions, Normal and Student's *t* distributions were considered in fitting the residuals data. The results of AIC values from the GOF test is displayed in Table 6. The results show that the residuals data of time series models of both stations are normally distributed as for both rain gauge stations, the AIC values for Normal distribution are the smallest.

Table 6. The AIC result of marginal distributions for residuals data

Marginal Distribution	Station M AIC	Station T AIC
Normal	4520.393	4474.989
Student's *t*	5371.810	5410.148

Copula Distribution

GOF test based on the AIC measurement has also been applied to choose the fitted copula model. The AIC of every copula model that have been applied in this research are listed in Table 7.

Table 7. The AIC result of copula distributions for residuals data.

Copula	Theta, θ	AIC
Gumbel	1.8527	-177.864
Clayton	1.7054	-221.040
Frank	5.0568	-209.412
Gaussian	0.6616	-222.332
Student's *t*	**0.6616**	**-232.374**

Based on the results shown on Table 7, it is identified that Student's *t* copula is the best fitted joint distribution for the residuals data. This is because the AIC value of Student's *t* copula is smaller relative to the AIC value of the other copula models considered.

Method 2: Copula Model

In the second method, the time series data is assumed to be stationary, which the mean and the variance do not change over time. The common copula models that has been used in hydrological analysis to fit the marginal rainfall data has been applied. Three marginal distributions: Normal, Gamma, and Weibull distributions were considered in fitting the rainfall data.

Marginal Distributions for Rainfall Data

The results of AIC values from the GOF test is displayed in Table 8. Based on the lowest AIC values, it shows that the best-fitted model for Station M is Weibull distribution and Normal distribution is the best-fitted model for Station T.

Table 8. The AIC result of marginal distributions for rainfall data.

Marginal Distribution	Station M AIC	Station T AIC
Normal	5513.529	**4507.607**
Weibull	**4588.460**	5634.098
Gamma	5592.913	4906.643

Copula Distribution

The AIC of each copula models for Station M and T are listed in Table 9.

Table 9. The AIC result of copula distributions for rainfall data.

Copula	Theta, θ	AIC
Gumbel	1.8970	-218.200
Clayton	2.1970	-239.000
Frank	**7.7780**	**-291.400**
Gaussian	0.5421	-173.320
Student's *t*	0.6811	-196.800

Based on the results shown on Table 9, it is identified that Frank copula is the best fitted joint distribution for the original rainfall data. This is because the AIC value of Frank copula is the smallest compared to the AIC value of the other copulas.

Comparison Between Method 1 and Method 2

Based on Tables 7 and 9, both methods get different results for the fitted copula. In Method 1, the best fitted copula is the Student's *t* Copula. On the other hand, in Method 2, the best fitted copula is the Frank Copula. The AIC value for Student's *t* Copula is -232.374 and the AIC value for Frank Copula is -291.400. However, to compare the performance of copula model fitted from Method 1 and 2, the AIC value cannot be used as the performance measure. This is because the marginal data for copula model for both methods are from different data sets. In Method 1, the original rainfall data have been used as the marginal variables, while in the Method 2, residuals data from time series model have been used as the marginal variables. Therefore, to compare the performance of copula model fitted from Method 1 and 2, simulation process has been done.

The rainfall data for Method 1 and 2 are simulated based on the chosen fitted marginal and copula distribution as shown in the framework Figure 10. The simulation is repeated for 500 times.

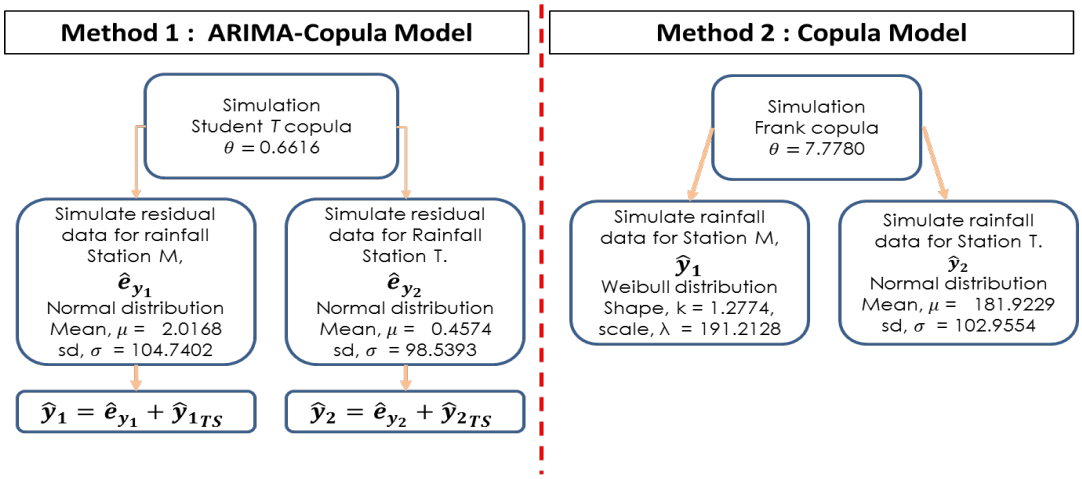

Figure 10. The framework for the data simulation for Method 1 and 2.

The plot of the observed (black colored) and simulated rainfall data are (red colored) shown in Figure 11. For Method 1, the plot is shown in Figure 11(a) and Figure 11(b) for Method 2.

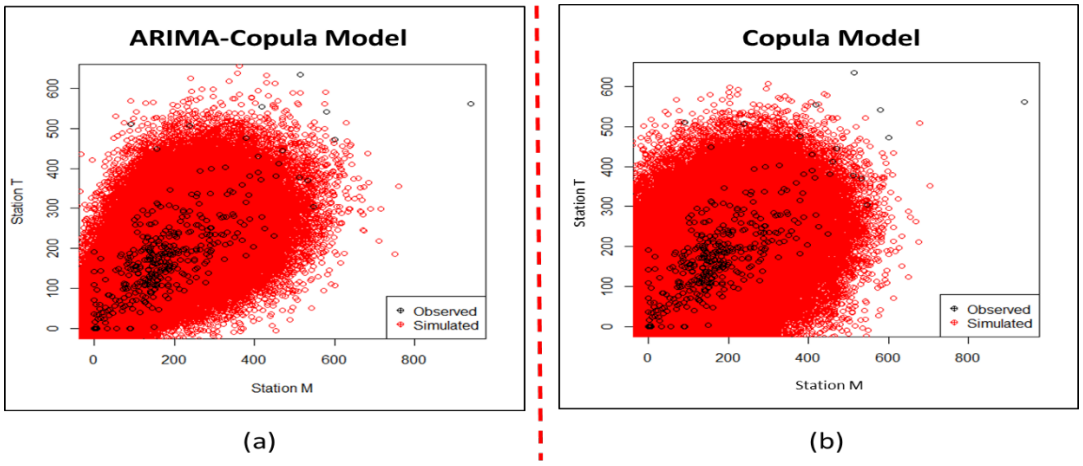

Figure 11. Plot of the observed and generated rainfall data for Method 1 and 2.

From the visualization of Figure 11, the spreading size of the 500 simulations of rainfall data from Method 1 is smaller than Method 2. This means that the simulated data in Method 1 is much closer to observed data compared to Method 2. Consequently, when the simulated data is closer to observed data, the error will be small. Thus, Method 1 has better performance in generating data compared to Method 2.

To assist and to find the accuracy from the visualization results of Figure 11, a performance measure is needed. In this study, root mean squared error (RMSE) based on equation (8) and (9) has been applied to be the performance measure. The equation (8) and (9) is for RMSE Method 1 and Method 2 respectively. The results of RMSE for Method 1 and 2 are shown in Table 10.

$$RMSE(\hat{\mathbf{y}}_1) = \sqrt{\sum_{i=1}^{500} \frac{(\hat{y}_{1i} - y_{1i})^2}{500}} \tag{8}$$

$$RMSE(\hat{\mathbf{y}}_2) = \sqrt{\sum_{i=1}^{500} \frac{(\hat{y}_{2i} - y_{2i})^2}{500}} \tag{9}$$

Table 10. RMSE results for simulated data from Method 1 and 2.

Method	Variables	RMSE
1	Station M, y_1	2858.41
	Station T, y_2	2687.98
2	Station M, y_1	3167.25
	Station T, y_2	2809.66

For Method 1, the RMSE for simulated rainfall data in Station M is 2858.41 and Station T is 2687.98. Meanwhile, for Method 2, the RMSE for simulated rainfall data in Station M is 3167.25 and Station T is 2809.66. Based on these results, RMSE of simulated rainfall data for both Station M and T in Method 1 are smaller compared to Method 2.

Consequently, the results from Figure 11 and Table 10 shows that Method 1 has better performance than Method 2. It is shown that considering the non-stationary condition either because of seasonal variability or volatility (mean and variance can change over time) in the time series marginal variables has a significant effect in finding the best fitted copula distribution. It because the best fitted copula should has better performance in generating random data and produce small error.

Conclusions

To examine the dependence structure of hydrological time series data between two rainfall gauge stations, a non-stationary condition is considered when fitting the marginal distributions. Copula model is presented in this paper to model the variable dependence structure between two rainfall series from Station Ldg. Lim Lim Bhd., Masai and Station Ldg. Sg. Tiram which are both located in Johor Bahru. There are two methods have been applied to identify the fitted copula model. First method is by considering the non-stationary condition that exist in the rainfall data, time series forecasting models: Seasonal ARIMA (SARIMA) model was applied and GARCH was not applied as the SARIMA models for Station M and T have no ARCH effect. The residuals data of the time series models were taken as the marginal variables. Second method is that the rainfall data is assumed as stationary and the observed rainfall data are directly taken as the marginal variables. In Method 1, the best fitted copula is Student's *t* Copula with AIC = -232.374. While, in Method 2, the best fitted copula in Frank Copula with AIC = is -291.400. To justify the results for Method 1 and 2, a simulation work is performed for both copula models and it is shown that Method 1 has better performance compared to Method 2.

The results from this study show that combination of time series model and copula does give different results than the static copula. This shows that the non-stationarity conditions should be considered. The developed copula model can be used to measure the conditional probability and return period of rainfall occurrences. It is also essentials in predicting the rainfall occurrences and mitigating the rainfall impacts such as drought and flood. In general, the findings have emphasized the significance of considering the non-stationary condition in the hydrological time series as the hydrological data are also collected under a changing environment over time.

Other than the combination of time series model and copula model to cater to the time varying data, there is another type of copula approach that has been proposed in recent years, called the time-varying copula, where existing copula models allowing for time varying dependencies in the marginal and copula parameters. This time-varying copula method is developed because it is possible that dependencies can change over time. There are three situations for the time-varying copula model can be constructed. First situation is that the copula parameter is constant and at least one marginal parameter is time varying. Second, the copula parameter is time varying and all the marginal parameters are constant. Third, at least the copula parameter and one marginal parameter are both time varying. Hence, the future study should apply time varying copula in developing the dependence modelling for the hydrological time series data.

Data availability

Data are available on request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding statement

This study was supported by Research University Grant (RUG) for High Impact Research Grant (HIR) [Q.J130000.2426.04G34] provided by the Malaysia Ministry of Higher Education (MOHE).

Acknowledgments

This research is done with the support from Universiti Teknologi Malaysia and the Malaysia Ministry of Higher Education (MOHE). The authors would like to thank the University and the grant for providing the sponsorship. The authors also would like to thank the Department of Irrigation and Drainage Johor for providing the rainfall data.

References

[1] G. Salvadori and C. De Michele, "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events", Water Resources Research, vol. 40, no. 12, W12511, 2004.

[2] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges", Publications de l'Institut de Statistique de L'Université de Paris, vol. 8, pp. 229-231, 1959.

[3] L. Zhang and V. P. Singh, "Bivariate rainfall frequency distributions using Archimedean copulas", Journal of Hydrology, vol. 332, no. 1-2, pp. 93-109, 2007.

[4] R.B. Nelsen, "An Introduction to Copulas", Springer, New York, 2006.

[5] F. Yusof, I. L. Kane and Z. Yusop, "Hybrid of ARIMA-GARCH Modeling in Rainfall Time Series", Jurnal Teknologi, vol. 63, no. 2, pp. 27–34, 2013.

[6] R. T. A. de Oliveira, T. F. Oliveira, P. R. A. Firmino and T. A. E. Ferreira, "Combining Time Series Forecasting Models via Gumbel-Hougaard Copulas", BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, pp. 568-573, 2013.

[7] R.T.A. de Oliveira, T.F.O. de Assis, P. R. A. Firmino and T. A. E. Ferreira, "Copulas-based time series combined forecasters", Information Sciences, vol. 376, pp. 110–124, 2017.

[8] N.M. Ariff, A.A. Jemain, K. Ibrahim and W.Z. Wan Zin, "IDF relationships using bivariate copula for storm events in Peninsular Malaysia", Journal of Hydrology, vol. 470-471, pp. 158-171, 2012.

[9] F. Yusof, F. H. Mean, S. Jamaludin and Z. Yusop, "Characterization of Drought Properties with Bivariate Copula Analysis", Water Resource Management, vol. 27, pp. 4183-4207, 2013.

[10] K. C. Yee, S. Jamaludin., F. Yusof and F. H. Mean, "Bivariate copula in fitting rainfall data", AIP Conference Proceedings, vol. 1605, no. 1, pp. 986-990, 2014.

[11] K. A. Kili, (2017, Nov 09), "Several areas in JB hit by flash floods", The Star Online. Retrieved from https://www.thestar.com.my/news/nation/2017/11/09/several-areas-in-jb-hit-by-flash-floods/

[12] A. F. Othman, (2017, Nov 14), "Flash floods hit several parts of Johor; bad weather predicted across 8 other states", New Straits Times. Retrieved from https://www.nst.com.my/news/nation/2017/11/303153/flash-floods-hit-several-parts-johor-bad-weather-predicted-across-8-other

[13] B. Tan, (2017, Nov 09), "Five areas in Johor Baru hit by flash floods", Malay Mail. Retrieved from https://www.malaymail.com/news/malaysia/2017/11/09/five-areas-in-johor-baru-hit-by-flash-floods/1506765

[14] T. Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity", Journal of Econometrics, vol. 31, pp. 307-327, 1986.

[15] I.L. Kane and F. Yusof, "Assessment of Risk of Rainfall Events with a Hybrid of ARFIMA-GARCH", Modern Applied Science, vol. 7, no. 12, pp.78-89, 2013.

[16] H. Chen, Q. Wan, F. Li and Y. Wang, "GARCH in mean type models for wind power forecasting", IEEE Power and Energy Society General Meeting, pp. 1-5, 2013.