Bivariate copula for flood frequency analysis in Johor river basin

N A Jafry [1], J Suhaila [1,2*], F Yusof [1], S R M Nor [1] and N E Alias [3]

[1] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia.

[2] UTM Centre for Industrial and Applied Mathematics (UTM-CIAM), Ibnu Sina Institute for Scientific and Industrial Research, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia.

[3] Department of Water and Environmental Engineering, School of Civil Engineering, 81310, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia.

*Corresponding E-mail: suhailasj@utm.my

Abstract. Flooding is a multi-attribute event that is described by many factors such as peak flow and flood volume. It is extremely vital to consider both the flood volume and the flood peak while studying the flood frequency analysis as the univariate analysis cannot accurately portray the flood issue and suffers from an underestimation and an overestimation problem. Traditional univariate and multivariate modeling techniques have several mathematical shortcomings including the inability to distinguish between the marginal and joint behavior of the variables under study. Therefore, the copula function was introduced to tackle the above restriction. Six copula models will be applied in this study to find the best bivariate distribution between the flood variables in Johor River Basin, Malaysia, including Gaussian, Student-t, Clayton, Gumbel, Frank, and Joe. Before that, several marginal distributions were fitted to select the most appropriate distribution for flood variables. It was found that the Pearson Type-III fits both the flood peak flow and the flood volume best. The evaluation of the best univariate marginal distribution and the copula model will be based on Akaike Information Criterion (AIC). Our findings suggested that Frank Copula is more suited to represent the relationship between peak flow and flood volume as it portrays the lowest AIC values of -69.41 and highest log-likelihood values of 35.7, where both values outperform the other proposed copula models. However, future research which considers all three flood variables which are peak flow, volume, and duration should be conducted to attain a more reliable result.

Keywords: Johor River Basin, Bivariate Copula, Flood

1. Introduction

Flood is also a natural hazard that poses a substantial risk to human lives, infrastructures, and the environment since it can cause major destructiveness to the parties involved within a split second. Hence, decision-makers and practitioners need to provide the best flood recurrence estimation and an early warning for the flood risk analysis. Flood assessments are also essential since they are used to determine the design and functioning of hydraulic infrastructure like dams and reservoirs [1]. A multivariate model that describes the flood characteristics needs to be considered in flood frequency analysis due to the inability of a univariate analysis to capture the whole situation thus the model will face an underestimation problem [2]. Numerous studies have emphasized its lack of credibility, contending that univariate frequency analysis methods cannot effectively describe inflow hydrographs or reduce the uncertainty in flood analysis [3]. Moreover, univariate analysis was found to be inadequate for flood frequency analysis since the event is characterized by multiple strongly correlated characteristics which play an important part in the flood frequency analysis.

Traditional univariate and multivariate modeling techniques, however, have several mathematical shortcomings that restrict their application, including the inability to distinguish between the marginal and joint behavior of the variables under study. Moreover, traditional multivariate modeling also has the major drawback of requiring the same parametric family of univariate distributions to characterize each flood's entities, which is not the case in real-world applications where each variable may have various distributions. Therefore, the traditional multivariate distributions might not provide decision-makers with the best outcomes, demanding the need for a more precise model [4]. Therefore, to get over the aforementioned restriction, the copula function, which is a technique that can display the structure of dependence between two or more random variables, was proposed [5].

Because they are adaptable tools that allow the marginals to be represented using any type of distribution without affecting the dependencies between them, copulas are becoming more popular and gaining a lot of attention in the financial industry as well as hydrology. This is because the limiting assumptions of normality and linear dependency can be avoided. [6]. [7] looked into the effects of climate change in the Azarshahr chay watershed, using the Gumbel–Hougaard copula function for future bivariate flood peak and volume variables, and discovered that using bivariate flood frequency analysis with a copula function instead of observational values results in a much more reasonable analysis and improved risk assessment. [8] developed a bivariate joint relationship between flood attributes for the Kelantan River Basin in Malaysia since multivariate probabilistic analyses of flood features and associated return times might be a holistic technique to make defensive risk-based judgements in numerous basin water-related concerns as compared to univariate flood frequency analysis. In a study conducted by [9], they discovered that the joint design value for peak and volume to represent flood events in the Qinhuai River basin is higher than the univariate design value of each variable assessed separately.

2. Materials and Methods

2.1 Study area

There are 189 river basins in Malaysia, including 89 in Peninsular Malaysia, and 85 out of 189 basins in Malaysia are subjected to recurring floods. These basins all have major rivers that immediately flow into the South China Sea. However, the Johor states will be the main focus of this study, particularly the Johor River Basin (JRB), which is located in Peninsular Malaysia's south and flows from Mount Gemuruh into the Johor Straits. JRB is prone to flooding again as a result of the influence of a brief period of rainfall with high intensity or a lengthy period of lower intensity rainfall which occurs over a few weeks [10].

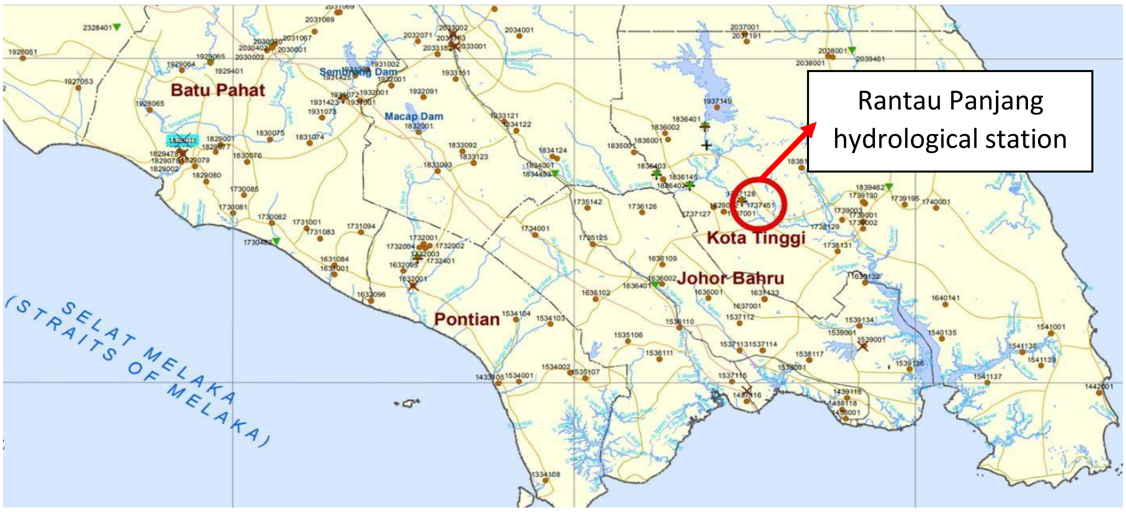
Rantau Panjang hydrological station
Batu Pahat
Semberong Dam
Macap Dam
Kota Tinggi
Johor Bahru
Pontian
SELAT MELAKA (STRAITS OF MELAKA)

Figure 1. Johor River Basin and the location of Rantau Panjang hydrological station (1737451).

Water year or hydrologic year will be divided into two categories which are Northern Hemisphere and Southern Hemisphere. The yearly cycle is related to the natural movement of the hydrologic seasons; in general, 1 October to 30 September in the Northern Hemisphere, 1 July to 30 June in the Southern Hemisphere. It begins with the start of the soil moisture recharge season, includes the maximum runoff season (or maximum groundwater recharge season, if any), and ends with the end of the maximum evapotranspiration season (or season of maximum soil moisture utilization). For this study, the water year for Southern Hemisphere will be used to calculate the Annual Maximum instead of the calendar year which normally commences from 1 January until 31 December. The data pre-processing in this study will be conducted using a seasonal basis (northeast monsoon and southwest monsoon). Daily discharge data from a gauging station in the Johor River Basin (JRB) which was obtained from Rantau Panjang hydrological station with station number 1737451, as marked red in Figure 1 will be used for this study. These datasets, which span 49 years beginning in 1972 and ending in 2021, were gathered from the Department of Irrigation and Drainage Malaysia.

2.2 Bivariate Copula

Let (X, Y) be bivariate random variables with continuous marginal distributions, such as $u_1 = F_X(x)$ and $u_2 = F_Y(y)$, it is possible to describe them individually by using the Copula function or C, which is defined on the unit square:

$$H_{X,Y}(x, y) = C[F_X(x), F_Y(y)] = C(u_1, u_2) \qquad (1)$$

where C is the bivariate copula, $F_X(x)$ *and* $F_Y(y)$ is the cumulative distribution function (CDF) of the observations. $H_{X,Y}(x,\ y)$ will be the bivariate joint probability density function (PDF). Additionally, if $F_X(x)$ and $F_Y(y)$ are continuous, the copula C must be distinct or unique.

Archimedean and meta-elliptical copulas are the most often used copula families for hydrologic applications [11]. For this study, the Archimedean family such as Clayton, Gumbel, Frank, and, Joe, as well as the Elliptical family such as Gaussian and Student-t copula will be used. In addition to the normal Archimedean copula family, the rotating versions of the Clayton were also added, Gumbel, and Joe families to allow for more flexibility since the normal copula cannot model negative dependence. The related surviving copulas may be obtained by rotating them by 180 degrees, while rotations of 90 and 270 degrees enable the modeling of negative dependency, which is not feasible with the original, unrotated forms. Table 1 is the summary of the copula families and their parameter range [12].

Table 1. Copula families and its parameter range.

Copulas	Copula	Parameter range
Gaussian	$= \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho)$ where the distribution function of a standard normal is $\Phi(\cdot)$ with *N (0,1),* while the bivariate normal distribution function with unit variances, zero means, and correlation ρ is called $\Phi_2(\cdot\ ,\cdot\ ;\ \theta_1)$. While $x_1 := \Phi^{-1}(u_1)$ and $x_2 := \Phi^{-1}(u_2)$. The copula density is as follows: $= \frac{1}{\phi(x_1)\phi(x_2)} \frac{1}{\sqrt{1-\rho^2}} exp\{-\frac{\rho^2(x_1^2+x_2^2)-2\rho x_1 x_2}{2(1-\rho^2)}\}$	$\rho \in (-1,1)$
Student-t	$= \int_0^{u_1} \int_0^{u_2} \frac{t_v(T_v^{-1}(v_1), T_v^{-1}(v_2);\ v, \rho)}{t_v(T_v^{-1}(v_1))\, t_v(T_v^{-1}(v_2))} dv_1 dv_2$ $= \int_{-\infty}^{b_1} \int_{-\infty}^{b_2} t(x_1, x_2; v, \rho) dx_1 dx_2$ where the variable transformation $b_1 := T_v^{-1}(u_1)$ and $b_2 := T_v^{-1}(u_2)$ is used, and *v* is the degree of freedom. $= \frac{t_v(T_v^{-1}(v_1), T_v^{-1}(v_2);\ v, \rho)}{t_v(T_v^{-1}(v_1))\, t_v(T_v^{-1}(v_2))}$	$\rho \in (-1,1)$
Clayton	$= (u_1^{-\theta_1} + u_2^{-\theta_1} - 1)^{-\frac{1}{\theta_1}}$	$\theta_1 \in (0, +\infty)$
Gumbel	$= exp\,(-((-ln\,u_1)^{\theta_1} + (-ln\,u_2)^{\theta_1})^{\frac{1}{\theta_1}})$	$\theta_1 \in [1, +\infty)$
Frank	$= \frac{1}{\theta_1} log\,[1 + \frac{(e^{-\theta_1 u_1}-1)(e^{-\theta_1 u_2}-1)}{(e^{-\theta_1}-1)}]$	$\theta_1 \in R \setminus \{0\}$
Joe	$= 1 - ((1-u_1)^{\theta_1} + (1-u_2)^{\theta_1} - (1-u_1)^{\theta_1}(1-u_2)^{\theta_1})^{\frac{1}{\theta_1}}$	$\theta_1 \in (1, +\infty)$

2.3 Evaluation of the best model

AIC will be the indicator used to measure the performance of various marginal distributions fitted to the flood variables as well as the accuracy of the best-fitted copula model [13]. The definition of the AIC of a bivariate copula family, *c* with parameter(s) $\boldsymbol{\theta}$ can be seen below:

$$\text{AIC} = -2 \sum_{i=1}^{N} \ln[c(u_{i,1}, u_{i,2}|\boldsymbol{\theta})] + 2k \tag{2}$$

where *k* = 1 for one parameter copula, *k* = 2 for two parameter copulas.

3. Results and discussion

This section will be discussing our findings and results.

3.1 Bivariate copula model

In this part, the interdependence between peak flow and flood volume by fitting a bivariate copula model will be modeled. Numerous multivariate copulas, including the well-known Archimedean copulas and elliptical copulas, will be assessed to model the flood events. Nonetheless, the first step is to choose the most appropriate marginal distribution for each flood variable, which will be determined by the maximum likelihood estimation approach. The computed AIC will be used to determine the best marginal distribution of each variable and the parameters will be estimated using the maximum likelihood method. Referring to Table 2, Pearson Type III distribution is the best-fitted distribution for both flood peak flow and the flood volume as it depicts the lowest AIC value compared to Gamma, Weibull, Gumbel, Generalized Extreme Value (GEV), and Generalized Pareto distributions. The next phase involved selecting the bivariate copula and estimating the appropriate parameters. Various copulas were selected as the candidate copulas which includes the widely used meta-elliptical copulas (Gaussian and Student t), as well as the Archimedean copulas (Clayton, Gumbel, Frank, and Joe), and their corresponding rotated forms with 90°,180°, and 270°.

Table 2. Performance of various probability models for fitting marginal distributions for flood variables.

Variables	Distributions	Estimated parameters	AIC
Peak Flow	Gamma	$\hat{\alpha}$; 3.002386, $\hat{\beta}$; 72.956530	556.4182
	Weibull	ζ; -55.993904, β; 172.474001, δ;1.173483	552.6702
	Gumbel	$\hat{x}$; 160.25394, $\hat{\alpha}$; 90.33617	557.8984
	GEV	$\hat{x}$; 146.4641109, $\hat{\alpha}$; 76.047642, $\hat{\kappa}$; 76.0476425	553.6768
	Generalized Pareto	$\hat{x}$; 56.867736, $\hat{\alpha}$; 183.5680776 $\hat{\kappa}$; 0.1353049	553.2137
	Pearson Type III	$\hat{\mu}$; **219.067249** $\hat{\sigma}$; **139.479456,** $\hat{\gamma}$; **1.696318**	**552.4969**
Volume	Gamma	$\hat{\alpha}$; 1.929435, $\hat{\beta}$; 697.198057	732.8629
	Weibull	ζ; -179.638291, β; 1186.330905, δ; 1.051536	731.2997
	Gumbel	$\hat{x}$; 922.3846, $\hat{\alpha}$; 682.5995	737.0723
	GEV	$\hat{x}$; 851.1630955, $\hat{\alpha}$; 618.0908464, $\hat{\kappa}$; -0.1944975	736.3083
	Generalized Pareto	$\hat{x}$; 181.5998933, $\hat{\alpha}$; 1504.879475; $\hat{\kappa}$; 0.1524635	731.8816
	Pearson Type III*	$\hat{\mu}$;**1326.825508,** $\hat{\sigma}$;**1163.254342,** $\hat{\gamma}$; **2.031485**	**729.5271**

[Note: Bold letter with an asterisk * shows that the performance of distribution is the most satisfactory]

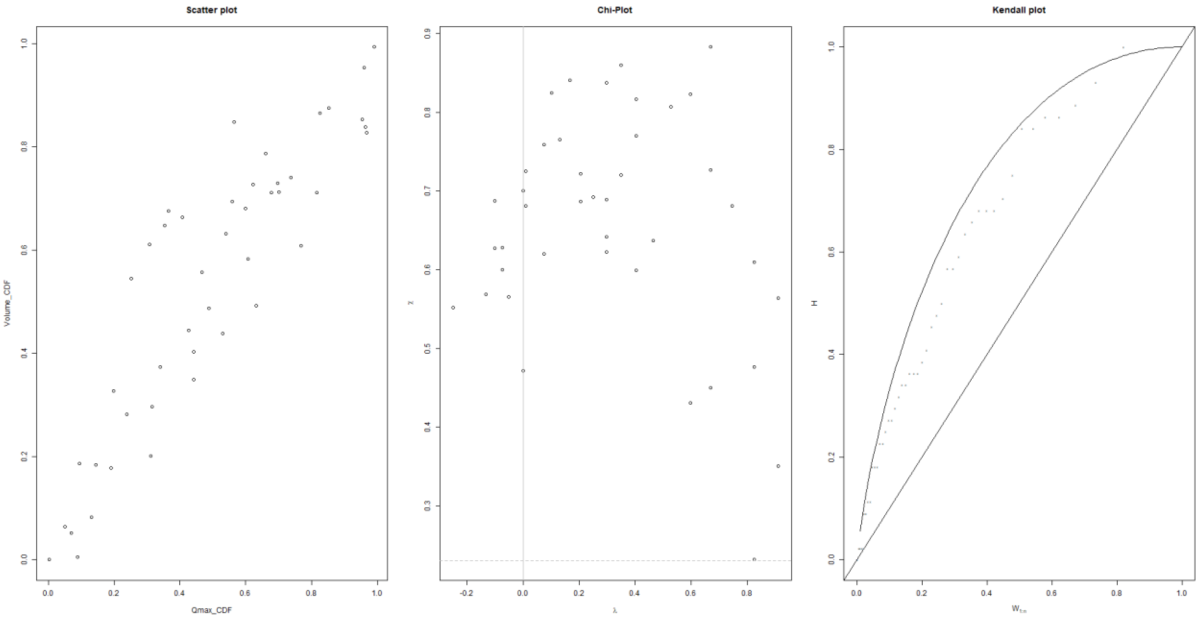

Figure 2. Graphical representation of the strength of dependence of flood peak flow and flood volume using scatter plot, Chi-plot, and Kendall plot.

Numerous graphical representations such as scatter plot, Chi-plot, and Kendall plot will be used to measure the strength dependency of flood variables. Based on the graphical representation, it can be seen that the data are positively correlated. The scatter plot shows a moderately strong, positive linear association between the two variables. On the other hand, the Kendall plot on the right column in **Figure 2** reveals that there is a divergence from the diagonal line points that demonstrates the strong positive relationship between peak flow and flood volume. Additionally, since the Kendall plot is above the diagonal line, it has come to our conclusion that these two variables are displaying positive dependence. To further support our assumption, we will be looking at the chi-plot which was positioned in the middle column in **Figure 2**. Given that the dots are situated on the top side of the confidence bands, the dependency structure between the pairs is positive. Thus, it can be confirmed that the flood peak flow and the flood volume exhibited a strong positive dependency.

Table 3 illustrates the performance of various bivariate copula models including Gaussian, Student-t, Clayton, Gumbel, Frank, and Joe, and their best-rotated form. After that, the best-fitted bivariate copula model was chosen using the standard measurement error criteria, such as log-likelihood and AIC values, as shown in **Table 3**. Based on **Table 3**, it has concluded that the Frank copula was found to be the most appropriate copula as it outperforms the other proposed copula models to be fitted to the bivariate flood peak and volume data since it shows the lowest AIC values and the highest log-likelihood values.

Table 3. Performance of various bivariate copula models of flood variables.

Copula	Parameters	Log-likelihood	AIC
Flood peak- Volume			
Gaussian	0.72	23.57	-45.15
Student-t	0.87, 5.69	28.62	-53.24
Clayton	0.19	5.930	-9.870

Gumbel[a]	1.55	10.92	-19.84
Frank*	**12.78**	**35.7**	**-69.41**
Joe[a]	1.31	5.38	-8.77

[Note: Bold letter with an asterisk * shows that the performance of copula is the most satisfactory, while [a] denotes the rotated 180^{o} copula]

4. Conclusion

Copula has been receiving so much attention and has been used in many fields, specifically hydrology, as it managed to model the dependency between two or more variables that have different univariate distributions, thus tackling the limitation of traditional multivariate modeling. Moreover, researchers have been utilizing this approach and highlighted its ability to distinguish between the marginal and joint behavior of the variables under study. Several copulas such as Gaussian, Student-t, Clayton, Gumbel, Frank, Joe, and their rotated version were fitted to model the dependency of flood peak and volume for the Johor River Basin. According to the log-likelihood and the AIC values that have been calculated, it was found that Frank copula is the best bivariate copula for the flood variables in Johor River Basin. However, more comprehensive insights can be obtained by considering all three flood characteristics or variables such as peak flow, volume, and duration for future research.

References

[1] Sedghi H, Telvari A and Babazadeha H 2017 Flood Analysis in Karkheh River Basin using Stochastic Model. Civil *Engineer. J.* **3** (9)

[2] Latif S and Mustafa F 2020 Copula-based multivariate flood probability construction: a review. Arabian Journal of Geosciences. **13** (3)

[3] Daneshkhah A *et al* 2020 Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model. *J. Hydrol.* **54** 469-487

[4] Tosunoglu F, Gürbüz F and İspirli M 2020 Multivariate modeling of flood characteristics using Vine copulas. *Environ. Earth Sci.* **79**(19)

[5] Sklar M 1959 Fonctions de repartition an dimensions et leurs marges. Publ. inst. statist. univ. Paris, **8** 229-231

[6] Mahfoud M and Michael M 2012 Bivariate Archimedean copulas: an application to two stock market indices. BMI Paper, 2012. **1517333**.

[7] Goodarzi, M R, A Fatehifar, and A Moradi, Predicting future flood frequency under climate change using Copula function. *Water Environment J.* **34** 710-727

[8] Latif S and Mustafa F 2021 Bivariate joint distribution analysis of the flood characteristics under semiparametric copula distribution framework for the Kelantan River basin in Malaysia. *J. Ocean Engineer. Sci.* **6**(2) 128-145

[9] Gao Y *et al* 2018 Multivariate flood risk analysis at a watershed scale considering climatic factors. *Water* **10** 1821-1817.

[10] Saudi A *et al* 2015 Flood risk index assessment in Johor River Basin. Malaysian *J Analy. Sci.* **19** 991-1000

[11] Chen L and Guo S 2019 Copulas and its application in hydrology and water resources. Springer.

[12] Brechmann E C and Schepsmeier U 2013 Modeling Dependence with C- and D-Vine Copulas: TheRPackageCDVine. *J. Stat. Soft.* **52**

[13] Ni L *et al* 2020 Vine copula selection using mutual information for hydrological dependence modeling. *Environ. Res.* **186** 109604

Acknowledgments

This study was funded by the University Fundamental Research Grant from Universiti Teknologi Malaysia, with a vote no. Q.J130000.2554.21H65. Furthermore, we would like to greatly thank the Department of Irrigation and Drainage, Malaysia, who has provided the authors with daily discharge data for the Johor River Basin.