Modelling the statistical dependence of rainfall variables using copula: A case study of Terengganu state

M F Fauzi¹, U F Abdul Rauf²*, R Khan³ and N A A Jamaludin⁴

¹ Department of Defence Science, Faculty of Technology and Defence Science, National Defence University of Malaysia, 57000 Sg. Besi, Kuala Lumpur, Malaysia
² Department of Mathematics, Centre for Defence Foundation Studies, National Defence University of Malaysia, 57000 Sg. Besi, Kuala Lumpur, Malaysia
³ Sypaq System Pty Ltd, 441 St Kilda Rd, Melbourne, Victoria 3004 Australia
⁴ Department of Computer Science, Faculty of Technology and Defence Science, National Defence University of Malaysia, 57000 Sg. Besi, Kuala Lumpur, Malaysia

ummul@upnm.edu.my

Abstract. Extreme rainfall can have severe negative impacts possibly resulting in loss of life and property, hence the need for means of predicting the occurrence of extreme rainfall. One method that addresses extreme phenomena is the extreme value theory. In bivariate data, extreme value analysis can be approximated by the distribution of maximum likelihood estimation. This paper discusses the role of the copula in modeling the structure of dependencies between two rainfall variables, namely duration, and severity. Goodness-of-fit approaches are used to resolve whether or not to reject a parametric copula which then determines the best copula to utilise for the data set. In this paper the Akaike information criterion (AIC) is investigated for its ability to choose an appropriate copula model from Archimedean copula models. The analysis shows that the Frank copula is the best model to explain the dependency structure of the two variables discussed.

1. Introduction

Analysis of dependency between two or more variables is an important study in many disciplines such as marketing, clinical trials, economics, and business to name a few. The standard of analysis is usually displayed in the form of a correlation value, which then provides an indicator to researchers about the analyses to be conducted.

This paper presents a method that has received a lot of attention in the last few years in hydrology, namely copula modeling, which is a statistical method intended to model the dependence structure between two or more variables. It is capable of providing more information about dependencies than the Pearson, Kendall or Spearman correlation, as highlighted by several authors, e.g., [1], [2], and [3]. In practice, the precursors often ignore or do not even know the marginal distribution of the variable being analyzed. Pearson correlation is often the easiest and simplest choice for measuring dependencies however it assumes that the variable has a normal distribution.
The role of copula becomes important when one or both variables have an abnormal marginal distribution or have tail dependencies. Many bivariate or multivariate distributions have been developed as alternatives to overcome cases of abnormalities in marginal distributions such as bivariate gamma [4], bivariate exponential distribution [5], bivariate exponential distribution [6], etc. However, these distributions are limited to the same marginal and have a complex structure in its probability density function (including its dependency structure) which is certainly not preferred in practice.

Copula comes with flexibility, where the marginal distribution of variables can be different or even unknown. Copula has been well applied in many fields such as hydrology [7], [8], [9], [10], finance and insurance [11], [12] and many more.

The research herein is a continuation of work by [13] where the authors presented forecasting of shortest and longest-term timescale of the wet event using monthly precipitation index data. In this paper we discuss the application of copula in modeling the dependence structure between two climate variables, namely rainfall duration and rainfall severity. The paper is outlined as follows. Section 2 introduces the methodology used throughout this paper. Sections 3 presents the results of a case study where copula is successfully applied to determine dependencies of these two variables. Finally, Section 4 presents the conclusions.

2. Methodology
It is important to understand how marginal distributions and Akaike information criterion work to fit copula models before conducting any analysis. This is because each distribution function can influence the result differently. This section will describe this methodology in detail.

2.1 Marginal Distributions and Akaike Information Criterion
To fit a copula model to the data the characteristics of precipitation exploited in this study to determine the marginal distribution are rainfall severity and rainfall duration. Four distribution functions have been tested the Gamma, Log-normal, Weibull and Exponential distributions. Maximum likelihood estimation (MLE) method is the standard method used to estimate the parameters of these marginal distributions. The equations of the probability density functions (pdfs) of the four distributions are listed below:

| Distribution    | Probability Density Function |
|-----------------|------------------------------|
| Gamma           | \( f (x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} e^{-\frac{x}{\beta}} x^{\alpha-1} \) for \( x > 0 \) | (1) |
| Log-normal      | \( f (x) = \frac{1}{\sqrt{2\pi} \sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} \) for \( x > 0 \) | (2) |
| Weibull         | \( f (x) = \left( \frac{a}{b} \right) \left( \frac{x}{b} \right)^{(a-1)} e^{-\left(\frac{x}{b}\right)^a} \) | (3) |
| Exponential     | \( f (x) = \lambda e^{-\lambda x} \) | (4) |

The Akaike information criterion (AIC) was introduced by [14] based on information theory. The goodness of fit of the copulas was evaluated by AIC. AIC provides a relative estimate of the information lost when a given model is used to represent the process that generates the data. The Akaike information criterion is defined in (5):

\[
AIC = 2k - 2 \ln(L_{max})
\]  

Here \( L_{max} \) is the maximum value of the likelihood function for the model, \( k \) is the number of estimated parameters in the model dependent on the type of univariate marginal distribution and copula parameters.
2.2 Copula

Copula is a joint distribution function of some marginal distribution function hence is used to analyse the dependence of random variables in structures described by the combined function [15]. Generation of the multivariate distribution function from the joint univariate distribution function are generally the function of the copula. Copulas are great tools for modelling and simulating correlated random variables that were created by [17]. In this paper, severity (S) and duration (D) are the two rainfall variables with marginal distribution functions represented as $F_S(s)$ and $F_D(d)$. For this work the copula function is modelled as a joint bivariate cumulative distribution function with these two characteristics as is given in Equation (6) below;

$$F_{S,D}(s, d) = C(F_S(s), F_D(d))$$

and become unique if both $F_S(s)$ and $F_D(d)$ are continuous.

2.3 Archimedean Copula

Archimedean copulas are among the most used in practice and have been used in a wide range of applications because they are easily generated and can capture a wide range of dependence. The Archimedean copula are of the form:

$$C(u, v) = \phi^{-1} (\phi(u) + \phi(v))$$

where $\phi$ is generator function over the interval $[0,1]$, continuous, convex, decreasing, with $\phi(1) = 0$. There are three Archimedean copulas commonly used in hydrology research namely, the Clayton, Frank, and Gumbel-Hougaard copulas. These copulas and their generators are displayed below.

**Table 2.** Families of bivariate Archimedean copulas.

| Archimedean Copula     | $C(u, v)$                                                                 |
|------------------------|---------------------------------------------------------------------------|
| Clayton                | $C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$   |
| Frank                  | $C_\theta(u, v) = -\theta^{-1}\log \left( 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right)$ |
| Gumbel-Hougaard        | $C_\theta(u, v) = \exp \left[ -\left(\log u\right)^{\theta} + \left(-\left(\log v\right)^{\theta}\right)^{\frac{1}{\theta}} \right]$ |

The following section will apply the method described above to a real hydrology data set to determine the dependence between rainfall severity and duration to help predict rainfall behavior.

3. **Case Study Area: Terengganu State**

Terengganu is a state in the East Coast which has the effects of Northeast Monsoon, from November to March every year, and the level of floods that occur is dependent on the intensity and rainfall value that falls at a time during the season. The precipitation time series used in this research consists of three rain-gauge stations across the states of Pahang consists of 45 years of rainfall observations from 1970 to 2014, collected from the Department of Irrigation and Drainage (DID) Ampang, Malaysia, retrieved to monthly time series data for tail dependence analysis. Table 3 presents the geographic locations of the selected rain-gauge stations.
Table 3. Summary of selected station in Terengganu States.

| Station | Station Number | Station Name       | Latitude   | Longitude   |
|---------|----------------|--------------------|------------|-------------|
| ET1     | 5230041        | SK Kuala Telemong  | 05°12'N    | 103°01'E    |
| ET2     | 4334094        | SK Kijal           | 04°19'N    | 103°29'E    |
| ET3     | 4434093        | SK Kemasek         | 04°25'N    | 103°27'E    |

3.1 Marginal Distributions and Akaike Information Criterion

As previously mentioned in this study, rainfall severity and rainfall duration are used to find the marginal distribution before copula parametric estimation methods can be applied to real hydrological data to avoid the misspecification of the marginal distributions. The four types of distributions described in section 2.1 were fitted to monthly rain-gauge station data and the aggregates were examined to identify the most suitable distribution for rainfall duration and severity. The parameters found for each of these marginal distributions are given in table 4.

Table 4. Marginal distributions of parameter estimate for severity and duration.

| Characteristics | Station | Estimates | Gamma | Log Normal | Weibull | Exponential |
|-----------------|---------|-----------|-------|------------|---------|-------------|
|                 |         |           | $\alpha$ 1/$\beta$ | $\mu$ $\sigma$ | $\alpha$ $\beta$ | $\Lambda$ |
| Severity        | ET1     | 1.85 0.46 | 1.10 0.75 | 1.33 4.44 | 0.25    |
|                 | ET2     | 2.15 0.57 | 1.08 0.72 | 1.50 4.22 | 0.26    |
|                 | ET3     | 1.23 0.25 | 1.03 0.92 | 1.06 5.02 | 0.20    |
| Duration        | ET1     | 2.08 0.75 | 0.76 0.71 | 1.43 3.07 | 0.36    |
|                 | ET2     | 2.83 1.05 | 0.80 0.62 | 1.76 3.04 | 0.37    |
|                 | ET3     | 1.60 0.48 | 0.86 0.81 | 1.24 3.59 | 0.30    |

The most suitable marginal distribution is selected based on the AIC goodness of fit test where the lower the AIC value the better the fit. The AIC results for each distribution model are presented in table 5 for all 3 stations (ET1, ET2 and ET3) for both variables (severity and duration). The results show the Log-normal distribution as having the lowest AIC value hence is the best marginal distribution for the given data set.

Table 5. Goodness of fits test for marginal distributions-AIC.

| Characteristics | Station | Marginal Distributions | Gamma | Lognormal | Weibull | Exponential |
|-----------------|---------|------------------------|-------|-----------|---------|-------------|
|                 |         |                        | 142.18 | **138.46** | 143.96 | 145.84 |
| Severity        | ET1     |                        | 118.10 | **117.16** | 119.22 | 123.24 |
|                 | ET2     |                        | 117.20 | **112.44** | 117.63 | 115.77 |
|                 | ET3     |                        | 117.42 | **114.19** | 119.29 | 123.06 |
| Duration        | ET1     |                        | 95.57  | **94.96**  | 96.72  | 105.50 |
|                 | ET2     |                        | 98.21  | **94.90**  | 99.14  | 99.77 |

3.2 Copula

The following method is used to estimate the dependence parameter for copula. The calculation of parameters, $\theta$ (section 2.2), and the value of the maximum log-likelihood is calculated to determine the best dependency structure model on copula. Copula functions and the parameter space of the copulas selected in the study are presented in table 6. The entries outlined in bold in table 6 show the best model for the given station where best is defined as the largest log-likelihood value. The
Frank copula model was identified as the best copula fit for two out of the three stations. Therefore, it can be concluded that the most appropriate model for describing the dependency structure between the two variables, duration and the severity of the rainfall is from the copula Archimedean family model, Frank.

| Station | COPULA | MLE | Estimate |
|---------|--------|-----|----------|
| ET1     | Clayton| 38.14| 4.49     |
|         | Frank  | 31.95| 19.41    |
|         | Gumbel | 35.36| 5.61     |
| ET2     | Clayton| 19.05| 3.96     |
|         | Frank  | 26.94| 17.96    |
|         | Gumbel | 22.99| 4.01     |
| ET3     | Clayton| 24.78| 7.74     |
|         | Frank  | 27.76| 24.37    |
|         | Gumbel | 25.94| 5.70     |

For future work the behavior of the upper part of the distribution should will be analysed by comparing the upper tail dependence of the observed data with the upper tail dependence provided by each copula model to select the copula that best fits the data.

4. Conclusion
To estimate the best fitted copula dependence parameter, the empirical study was analyzed to provide statistical evidence in choosing which parameter estimation methods that are more accurate and efficient. Log-normal distribution was fitted to each characteristic using the AIC goodness-of-fit method and the Frank copula was found to be the best copula estimate by maximum log-likelihood.

Acknowledgement
The funding for this research is fully supported by Ministry of Higher Education Malaysia under Fundamental Research Grant Schemes (FRGS /1/2015/SG04/UPNM/02/2).

References
[1] Pearson K 1895 *Proc. of the Royal Society of London* 59 69–71
[2] Kendall M G 1970 *Rank Correlation Methods* (London: Griffin) p 202
[3] Zhang A 2006 *Advanced Analysis of Gene Expression Microarray Data* World Scientific Publishing Company
[4] Moran P A P 1969 *Statistical Inference with Bivariate Gamma Distributions* (Oxford University Press) p 627
[5] Conway D A 1979 Farlie-Gumbel-Morgenstern *Distributions Encyclopedia of Statistical Sciences* 3 28-31
[6] Gumbel E J 1960 *J. of the American Statistical Association* 55 698-707
[7] Abdul Rauf U F and Zeephongsekul P 2014 *Theor Appl Climatol* 115 153-166
[8] Masina M, Lamberti A and Archetti R 2015 *Coastal Engineering* 97 37-52
[9] Balistrocchi M & Orlandini S, Ranzi R and Bacchi B 2017 *Water Resources Research* 53 9883-9900
[10] Stamatatou N, Vasiliades L and Loukas A 2018 *Proc. 3rd EWaS Int. Conf.* vol 2 (Greece: Lefkada Island) p 635
[11] Mejdoub H and Ben Arab M 2017 *Asian Development Policy Review* 5 100-119
[12] Tamraz M 2018 *Mixture copulas and insurance applications* (Cambridge University Press) p 391
[13] Fauzi M F et al. 2017 Proc. 1st ICoAIMS Int. Conf. 890 012159
[14] Akaike H 1974 IEEE 19 716-723
[15] Nelsen R 1998 An Introduction to Copulas 139 (New York: Springer Verlag)
[16] Sklar A 1973 J. of Information Theory and Automation of the American (Kybernetika) 9 449–460