CHAPTER 1

An Introduction to Copula Modelling

Fadhilah Yusof and Shariffah Suhaila Syed Jamaludin

1.1 INTRODUCTION

Hydrological events such as floods and droughts are multivariate because they are characterized by more than one correlated random variable. Therefore, a single variable analysis would only provide limited assessments of these events (Yue et al., 2001). Many studies have been done to derive multivariate distributions of random variables from flood or drought characteristics. The derivations were either from random variables of similar distributions, assumed as a joint normal distribution, or assumed independent (Zhang & Singh, 2006). However, in the actual situation, the correlated random variables are generally dependent, do not follow the normal distribution, and do not have the same type of marginal distribution. Therefore, deriving multivariate distribution that reflects the actual hydrological process is mathematically complicated. The method of copula was then introduced to overcome the mentioned problem.

In the copula theory, the complexity of obtaining multivariate distributions by separating the analysis of marginals and the dependence structure of multivariate distribution is reduced. Furthermore, copula can provide a flexible approach that allows more choices of marginal distributions and dependence structures in the multivariate analysis (Nelson, 2006; Kao & Govindaraju, 2008).

1.2 WHAT IS COPULA?

Let us start by defining copula. Copulas are functions developed by Sklar (1959) that link the univariate distribution functions to form a multivariate distribution function. Frechet, in the late fifties, queried how to determine the relationship between the multivariate distribution functions and its lower-dimensional margin. Sklar (1959) then answered this query by discovering that at least one function that always exists, which he named "copula", that links a joint distribution to its marginal via the following expression:

$$H(x_1, x_2, \dots, x_n) = C\big(F_1(x_1), F_2(x_2), \dots, F_n(x_n)\big),$$

where $x_1, x_2, \dots, x_n$ represent the continuous random variables and $F_1(x_1), F_2(x_2), \dots, F_n(x_n)$ are the marginal distributions. This is then reflected in the term copula, which originates from the Latin verb *copulare,* which means "to join together". It usually consists of a combination of two or more uniform marginal distributions. The copula theory can overcome the limitations of the traditional approach that only focus on the same distributions for example bivariate gamma or bivariate weibull. Copula allows any distribution function as the marginal distributions to be joined in a distribution where the dependence structure of the variables is constructed. Zhang and Singh (2007a) have proven that the copula method can derive bivariate joint distributions of rainfall variables with different marginal distributions without assuming the variables are normal or independent. Hence, the copula method has been widely used in many fields of application due to its ability to incorporate dependency elements in the distribution.

Let us now define the copula based on the theorem from Sklar (1959) and some definitions from Nelson (2006). For n-dimensional continuous random variables $\{X_1, X_2, \dots, X_n\}$ with marginal distributions, $F_1(x_1), F_2(x_2), \dots, F_n(x_n)$ there exists a unique copula C such that

$$H(x_1, x_2, \dots, x_n) = C\big(F_1(x_1), F_2(x_2), \dots, F_n(x_n)\big) \\ = C(u_1, u_2, \dots, u_n) \quad (1.1)$$

The joint distribution function of an n-copula is defined as a multivariate uniform distribution with the cumulative distribution function C mapping as shown below:

$$C: [0,1]^n \to [0,1]$$

where $F_k(x_k) = u_k$ for $k = 1, 2, \dots, n$ with $U_k \sim U(0,1)$ and H is a joint distribution function with margins $F_1(x_1), F_2(x_2), \dots, F_n(x_n)$.

Based on Sklar (1959), copula has the following properties:

(1) Let $u = [u_1, u_2, \dots, u_n]$ where $u_i = F_i(x_i) \in [0,1]$, if $\mu_i = 0$ for any $i \leq n$ (at least one coordinate of u equals 0), then $C(u_1, u_2, \dots, u_d) = 0$.

(2) $C(u) = u_i$ if all the coordinates are equal to 1 except u_i, i.e., $C(1,1, \dots, u_i, \dots, 1,1) = u_i$, for every $i \in \{1,2, \dots, n\}, u_i \in [0,1]$.

(3) $C(u_1, u_2, \dots, u_n)$ is bounded i.e., $0 \leq C(u_1, u_2, \dots, u_n) \leq 1$. This property represents the limit of the cumulative joint distribution, i.e., in the range $[0,1]$.