# Part One

## Theory

# 1
# Introduction

**ABSTRACT**

This chapter briefly reviews the development of the copula theory and its applications in the field of water resources engineering (flood, drought, rainfall, groundwater, etc.). It points out the need for applying the copula theory in hydrology and engineering. The chapter is concluded with an outline of the structure of the book.

## 1.1 Need for Copulas

Complex hydrological processes, such as floods, droughts, winds, rainstorms, and snowfall, are characterized by more than one correlated random variable. Hydrologic events emanating from these processes are multivariate and their treatment requires multivariate analysis. Yue (1999, 2000a, 2000b, 2000c), Yue et al. (2001), and Yue and Rasmussen (2002) reviewed some applications of multivariate hydrological analyses using traditional frequency analysis methods with multivariate distributions.

Multivariate frequency distributions have usually been derived using one of three fundamental assumptions (Zhang and Singh, 2006): (1) the random variables each have the same type of marginal probability distribution; (2) the variables are assumed to have a joint normal distribution or are transformed to have a joint normal distribution; or (3) the variables are assumed independent – a trivial case. In reality, the correlated random variables are generally dependent, do not follow the normal distribution, and/or do not have the same type of marginal distributions. In general, multivariate hydrological analyses are mathematically complicated, and the resulting joint distributions may be valid only in a limited solution space.

When deriving multivariate distributions, it has been demonstrated in the last two decades that the aforementioned difficulties can be overcome with the use of copulas because: (1) they separate the dependence function from the marginal distributions of random variables; (2) the dependence function represented by the copula function is the cumulative joint distribution of correlated random variables; and (3) the mutual information (bivariate/multivariate) may be expressed as the negative copula entropy that avoids the complexity of evaluating the uncertainty with the use of entropy theory (information theory). In what follows, we briefly summarize copulas and their applications.

## 1.2 Introduction of Copulas and Their Application

Copula was first introduced by Sklar (1959). Later on, Joe (1997) and Nelsen (2006) further discussed the dependence structure of multivariate random variables using the copula theory. The copula theory was first developed in the fields of statistics and finance (more specifically econometrics). In this section, we will first briefly introduce the history of development of copulas, followed by a brief introduction of copula properties, parameter estimation, and applications to the field of water resources engineering.

### *1.2.1 Development and Applications of Copulas in Statistics and Finance*

Copula theory has been developed and applied in the fields of statistics and finance. Ali et al. (1978) proposed a bivariate distribution family, i.e., the bivariate logistic distribution by considering the survival odds ratio. They also studied the properties of the bivariate distribution. Now it is named the Ali–Mikhail–Haq (AMH) copula family. It is worth noting that this copula family may not be applicable, unless Kendall's tau rank correlation coefficient falls in the range of (–1/3 to 1/3).

Cook and Johnson (1981) proposed a simple bivariate distribution family to represent nonelliptical symmetric bivariate random variables. The proposed copula, however, may only be applied to the positively correlated random variables. They also proved that multivariate Pareto, Burr, and logistic distributions were special cases, and that copula is now named the Cook–Johnson (Clayton) Archimedean copula family.

Genest and McKay (1986) described bivariate distributions with uniform marginals on a unit interval. They discussed how bivariate distributions (copula) may be applied for singular components and the geometric interpretation of Kendall's tau. Genest (1987) studied the Frank family of bivariate distributions and concluded that it was appropriate to apply the Frank family to construct the bivariate distribution with any given marginals and cover all possible dependence structures. He then introduced three nonparametric estimators and one parametric estimator, i.e., the maximum likelihood estimation (MLE) method. Genest and Rivest (1993) studied the Archimedean one-parameter copula. They applied Kendall's tau for parameter estimation and found that Kendall's tau may also be applied for selecting the appropriate copula for certain multivariate random variables, and analyzed uranium exploration data to explain how to apply the estimation procedure.

Genest et al. (1995) investigated the properties of another semi-parametric estimation method to estimate copula parameters. This semi-parametric estimation method can be considered as a pseudo-likelihood method that is found to be consistent and asymptotically normal. The performance of the pseudo-likelihood method was investigated by analyzing the bivariate Clayton (Cook–Johnson) copula. Later, Caperaa et al. (1997) proposed a new nonparametric method and examined its asymptotic properties and small sample behavior compared to the estimation method through Kendall's tau statistic and maximum likelihood method. They found that the proposed method was strongly convergent and asymptotically unbiased.

Genest and Boies (2003) discussed the Kendall plot as a measure of dependence. Similar to chi-plot, the Kendall plot is invariant with respect to the monotone transformation of marginal distributions. They also found that the Kendall plot is easier to interpret than the chi-plot, which may also be extended to multivariate analysis (dimension $\geq 3$). Genest et al. (2006, 2007a) investigated the formal goodness-of-fit statistical tests for copulas. Chakak and Koehler (1995) presented a procedure to construct families of multivariate distributions through specified univariate and bivariate margins. Their procedure constructs multivariate distributions through conditional distributions.

Zheng and Klein (1995) proposed a copula-graphic estimator, which is a maximum likelihood estimator. The copula-graphic estimator was applied for the estimation of marginal distributions from the given copula for survival analysis. Simulation was performed using the Monte Carlo method, and the robustness of the method showed that the assumption of completely specifying the copula allowed for estimating the complete joint survival function based only on the competing risk data.

Quesada-Molina and Rodriguez-Lallena (1995a, b) investigated bivariate copulas with quadratic and cubic sections, which were derived from simple univariate real-valued functions on the interval [0, 1]. They applied various positive dependence structures (i.e., quadrant dependence and total positivity), measures of association (i.e., Kendall's $\tau$ and Spearman's $\rho$), stochastic ordering, and various notions of symmetry, which were shown to be equivalent to certain simple properties of univariate functions used for constructing bivariate copulas. They applied several examples to illustrate how these copulas can be constructed.

Müller and Scarsini (2001) considered two random vectors $X$ and $Y$ with the component of $X$ dominated in the convex order by the corresponding components of $Y$. They found that the positive linear combination of the components of $X$ dominated in the convex order by the same positive linear combination of the components of $Y$ had the properties as the two random vectors having the common copula and conditionally increasing.

Frees and Valdez (1997) applied copulas, i.e., the Archimedean copula in an actuarial study, and estimated their parameters by both nonparametric and parametric methods. It was concluded that the Archimedean copula could be used to represent the bivariate distribution in the actuarial study fairly well.

Sancetta and Satchell (2001) analyzed financial multivariate data whose marginals were not normally distributed. Based on the nice Bernstein properties, they applied the Bernstein polynomial approximation to copulas and then investigated the multivariate convergence properties. The portfolio data were applied to investigate statistical properties and applications of Bernstein copulas. Chen and Fan (2002) investigated the issue related to the density forecast by applying a copula. They proposed a parametric test for the correct density forecasts by nesting a series of independently identically distributed random variables from stationary Markov processes. By applying the copula, they found that this test exhibited a large variety of marginal properties. Coupling the same marginals with different copula functions, they found that the test again exhibited numerous dependence properties.

Fang et al. (2002) investigated the joint probability density function of continuous random variables with given marginals by analyzing elliptically contoured distributions, e.g., normal distribution. They named this joint density function as meta-elliptical distribution. The analytical formulation, conditional distribution, and dependence properties of this meta-elliptical density function were discussed. They found that meta-elliptical joint distribution held the same Kendall tau as did the meta-Gaussian joint distribution belonging to the meta-elliptical joint distribution. Brakekers and Veraverbeke (2005) extended the estimator proposed by Rivest and Wells (2001) to the fixed design regression application. In survival analysis, the variables were generally assumed independent, which may be invalid in certain practical applications.

### *1.2.2 Construction and Parameter Estimation of Copulas*

With the development of copula theories in statistics, Nelsen (2006) summarized the four most efficient methods to construct the copulas: (1) inversion method, (2) geometric method, (3) algebraic method, and (4) with specified properties. A detailed discussion of the construction of copulas and their properties will be provided in Chapter 3.

For any given copulas, their parameters may be estimated non-parametrically, parametrically, or semi-parametrically. The nonparametric method estimates the parameters with the rank correlation coefficient, i.e., Kendall's $\tau$ or Spearman's $\rho$. This method yields the analytical solution if there is a closed-form solution between rank correlation coefficient and copula parameters (e.g., certain Archimedean copulas that will be discussed in Chapter 4).

The copula parameters may be estimated parametrically with the use of one of the following three methods:

- Full MLE, by which the parameters of marginal distributions and copulas are estimated simultaneously.
- Two-stage MLE, by which the parameters of marginal distributions and the parameters of copula function are estimated separately using MLE. In this case, the fitted parametric marginal distributions will be applied to estimate the copula parameters through MLE.
- The semi-parametric method (also called pseudo-MLE: PMLE), which applies the empirical distribution (computed using probability plotting-position formula or kernel density) to estimate the copula parameters using MLE. Unlike the parametric approach, the semi-parametric method is marginal free.

Details of the estimation methods will be discussed in Chapter 3 and the following chapters.

To assess the goodness-of-fit of the fitted or proposed copula functions, Genest and Boies (2003), Genest et al. (2006), and Genest et al. (2007a) proposed the graphical and numerical assessment tools. These goodness-of-fit measures will be further introduced and applied in the chapters that follow.

### 1.2.3 Application of Copulas in Water Resources Engineering

With the theoretical development of copula theory and its advancement in statistics and econometrics, copulas have been adopted and applied in the fields of hydrology, water resources, and environmental engineering. These applications are briefly reviewed in the following section.

*Copula Applications in Flood Frequency Analysis*

Salvadori and De Michele (2004) provided a general theoretical framework exploiting copulas to determine return periods of bivariate hydrological events. They concluded the following: (1) copula may greatly simply the calculations of return period and may even yield an analytical solution; (2) copula may be associated with the return period of specific events; (3) with the use of copula, one may define sub-, super-, and critical events as well as those of primary and secondary return periods; and (4) the copula approach may be easily generalized to multivariate cases. The proposed methodology was further illustrated using flood peak and flood volume in a river basin in southern Taiwan, the spillway design flood of an existing Italian dam, and the annual maximum peak flow at Chute-des-Passes. Using flood variables (i.e., peak discharge, flood volume, and flood duration) observed at Kanawa River as an example, Grimaldi and Serinaldi (2006a) showed that (1) the flood variables were correlated; and (2) the dependence may not be symmetric among the flood variables, depending on the threshold used to identify the flood event. Employing the asymmetric Frank copula, the symmetric Frank copula, and the logistic Gumbel distribution through case studies, they presented the following: (1) the possible improvement obtained using the asymmetric copula and (2) the advantages in using the asymmetric copula.

Zhang and Singh (2006) applied the copula method to derive bivariate distributions of flood peak and volume, and flood volume and duration, such that the mariginals may follow different probability distributions. The conditional return periods for hydrologic design were tested using flood data from Amite River at Denham Springs, Louisiana, and the Ashuapmushuan River at Saguenay, Quebec, Canada. Comparing the derived distributions with the Gumbel mixed distribution and the bivariate Box–Cox transformed normal distribution, the copula-based distributions were found to result in the best agreement with plotting position-based frequency estimates. Genest et al. (2007b) presented how meta-elliptical copulas could be used to model the dependence structure of random vectors when observed differences between their bivariate margins precluded the use of exchangeable copula families, e.g., the Archimedean copula family. A case of peak, volume, and duration of the annual spring flood for the Romaine River was employed to illustrate rank-based estimation and goodness-of-fit techniques for this broad extension of the multivariate normal distribution. Analysis of annual spring flood for the Romaine River suggested that in view of the short length of the series, any of the eight meta-elliptical copula models considered in their studies could be used for prediction purposes. Only with additional evidence could one hope to distinguish between these dependence structures.

Simonovic and Karmakar (2007) focused on the selection of marginal distribution functions for flood characteristics by parametric and nonparametric estimation procedures, and demonstrated how the concept of copula may be used for establishing a joint distribution function with mixed marginal distributions for 70 years of streamflow data of Red River at Grand Forks in North Dakota, United States. Zhang and Singh (2007b) employed the Gumbel–Hougaard copula to model trivariate distributions of flood peak, volume, and duration, and then obtained conditional return periods. The derived distributions were tested using flood data from the Amite River basin in Louisiana. A major advantage of the copula method is that marginal distributions of individual variables can be of any form and the variables can be correlated.

Grimaldi and Serinaldi (2006a) described the fully nested (asymmetric) Archimedean copula properties and the inference procedure, and applied the copulas to multivariate flood frequency analysis of the Kanawha River (Kanawha Falls, West Virginia, drainage area 21,681 $km^2$) recorded from 1877 to 2003, and multivariate sea wave frequency analysis of Rete Ondametrica Nazionale (RON) network off the La Spezia (Liguria region, Italy). They found the following: (1) the inference procedure via copulas was quite easy to perform; and (2) asymmetric Archimedean copulas were useful to describe trivariate structures of dependence of nonexchangeable variables with different mutual degrees of correlation fulfilling the conditions described in Section 5.2.1; and finally, (3) comparison between observed and synthetic samples generated by estimated trivariate distributions confirmed the satisfactory performance of the Chen–Fan–Patton (CFP) test in order to choose the best-fitting copula. But asymmetric Archimedean copulas were not able to describe all mutually different structures of dependence. In addition, since the CFP test is based on Rosenblatt's transformation, its application becomes difficult when the number of variables increases. Consequently, further studies are needed to find both families of copulas that are capable of describing more complex structures of dependence and goodness-of-fit tests suitable for application to every copula class and high dimensions.

Wang et al. (2009) used a copula-based flood frequency (COFF) approach to estimate the risk of floods at confluence points. The four often-used Archimedean copulas (Ali–Mikhail-Haq, Clayton, Frank, and Gumbel–Hougaard) were applied in a river basin for the joint probability estimation. The Frank copula and Gumbel–Hougaard copula performed the best for the discharge data collected at two United States Geological Survey (USGS) gauge stations located on the Des Moines River at Fort Dodge, Iowa (USGS 05480500; Station A) and the Boone River near Webster City, Iowa (USGS 05471000; Station B), upstream of Des Moines River basin near Stratford, Iowa. It was shown that the copula method for specifying the multivariate distribution function was powerful, because it avoided the requirement that the marginal distributions be of the same type, which is assumed in most studies of empirical multivariate distributions. They also explained that it avoided the complex formulas that arise for many multivariate distribution functions. Zhang and Singh (2014) studied the trivariate flood frequency analysis by allowing different lengths of the records for maximum daily discharge at different locations.

*Copula Application to Precipitation and Storm Characteristics Analysis*

Salvadori and De Michele (2006) presented a statistical procedure to estimate probability distributions of storm characteristics. They discussed a method to describe the temporal dynamics of rainfall via a reward alternating renewal process that describes wet and dry phases of storms. The dependence among the three variables of interest (I for average rainfall intensity, W for the wet phase, and D for the dry one) was given via a Frank 3-copula. Based on real data collected by the Italian Sea Wave Measurement Network, De Michele et al. (2007) focused on how copulas can be used for the multidimensional frequency analysis of sea storm significant wave height (H), storm duration (D), storm direction (A), and storm interarrival time (I) (i.e., the calm period separating two successive storms). These included the following analyses:

- The construction of a bivariate model for the pair (H, D). In turn, this yielded the statistics of the sea storm magnitude M.
- Calculation of the return period of multivariate events. This gives the possibility to calculate the probability of occurrence of supercritical events and yielded an estimate of the minimum energetic content of sea storms having an assigned (multivariate) return period.
- Construction of a trivariate model for a triplet (H, D, A). This provided useful indications about the relation between sea storm magnitude and direction.
- Extension to storm interarrival duration I. This yielded a trivariate model for the triple (D, I, A) that cast new light on the relation between sea storm timing and direction.
- The construction of a global model for the vector (H, D, I, A). The overall structure was that of a reward alternating renewal process, whose dynamics develops along a random direction. In turn, this gave the possibility to simulate a sequence of sea storm events, accounting for all the variables of interest and their mutual relations.

These statistical analyses are very important when dealing with coastal dynamics, marine structure reliability, or the planning of operations at sea.

Zhang and Singh (2007a) derived trivariate rainfall frequency distributions using the Gumbel–Hougaard copula, which does not assume the rainfall variables to be independent or normal or have the same type of marginal distributions. The trivariate distribution was then employed to determine joint conditional return periods and was tested using rainfall data from the Amite River basin in Louisiana. Zhang and Singh (2007c) derived bivariate rainfall frequency distributions using the copula method in which four Archimedean copulas (Gumbel–Hougaard, Ali–Mikhail–Haq, Frank, and Cook–Johnson) were examined and compared. Results indicated that the advantage of the copula method is that no assumption is needed for the rainfall variables to be independent or normal or have the same type of marginal distributions. They also used the aforementioned Archimedean copulas to determine joint and conditional return periods, and tested using rainfall data from the Amite River basin in Louisiana, United States. Salvadori and De Michele (2007) summarized a general theoretical framework for studying the return period of hydrological events and presented a trivariate Frank copula model for the temporal structure of the

sequence of storms at the Scoffera station, located in the Bisagno River basin (Thyrrhenian Liguria, northwestern Italy). The model includes, simplifies, and generalizes many of the approaches already present in the literature. They also gave an explicit derivation of the storm volume statistics for any suitable copula and marginals and a copula-based procedure for estimating the probability law of antecedent moisture conditions. Results indicated that the copula may have important applications in many fields of water resources and hydrologic systems, as well as in several geophysical areas.

Using three different samples of extreme rainfall criteria, including annual maximum volume (AMV), annual maximum peak intensity (AMI), and annual maximum cumulative probability (AMP), Kao and Govindaraju (2007) characterized extreme rainfall events using hourly precipitation data from Indiana, United States. Results of their study have implications for current hydrologic design in that they provided better estimates of design rainfall. Gebremichael and Krajewski (2007) explored the use of copulas to construct the joint distribution between the sampling error and the corresponding rainfall rate. Taking 15-minute radar-rainfall data for the Mississippi River basin in the central United States as an example, the approach (1) estimated the marginal distribution functions in a parametric way; (2) used these with a number of copula functions in search of the one most appropriate; (3) used the maximum likelihood to estimate the parameters of copulas; and (4) selected the best-fitted parametric copula function as the one that gave the largest likelihood. Results showed that the approach had important implications for the interpretation and propagation of remote sensing precipitation uncertainties.

Based on a non-Archimedean Plackett copula family derived using the theory of constant cross-product ratio, Kao and Govindaraju (2008) showed that the Plackett family not only performed well at the bivariate level, but also allowed trivariate stochastic analysis where the lower-level dependencies between variables can be fully preserved while allowing for specificity at the trivariate level as well. The authors proposed a numerical method to estimate the feasible range of Plackett parameters. The trivariate Plackett family of copulas was then applied to study a total of 53 hourly rain gauges from the Hourly Precipitation Database (TD 3240) of the National Climate Data Center in Indiana. Results of this study suggested that while the constant cross-product ratio theory was conventionally applied to discrete type random variables, it was also applicable to continuous random variables, and that it provided further flexibility for multivariate stochastic analyses of rainfall.

Evin and Favre (2008) proposed a new stochastic point rainfall model (Neyman–Scott cluster process) considering the dependence between cell depth and duration using cubic copula, and explored the properties of this class of copulas and suggested several families of this kind attaining a large range of dependence. They derived first-, second-, and third-order moments of the modified Neyman–Scott rectangular pulses model. Hourly rainfall data from Belgium and America were employed to fit the model by these theoretical moments and obtained successful results for two rainfall series with different climates. Generating long series of synthetic rainfall and the observed rainfall data and under specific