Copulas Analysis of Victorian Precipitation Data: a Nonparametric Approach

Ummul Fahri Abdul Rauf [1, a)] and Mohd Faris Fauzi [2, b)]

Author Affiliations
[1] *Department of Mathematics, Centre for Defence Foundation Studies, National Defence University of Malaysia, 57000 Sg. Besi, Kuala Lumpur, Malaysia*
[2] *Department of Defence Science, Faculty of Technology and Defence Science, National Defence University of Malaysia, 57000 Sg. Besi, Kuala Lumpur, Malaysia*

Author Emails
[a)] *Corresponding author: ummul@upnm.edu.my*
[b)] *farisfauzi.work@gmail.com*

Abstract. This study used data from rain gauges at six (6) Victoria stations with some of the highest annual mean rainfalls. Empirical copulas are used to model the bivariate rainfall severity and duration distribution. A nonparametric density function is used to estimate the Standard Precipitation Index (SPI) prior to fitting copulas. The marginal distributions of each of the rainfall characteristics in the copula function's argument are then fitted using empirical distributions. Based on the values of Mean Absolute Error, the empirical copula can be used to represent the results obtained with standard copulas.

INTRODUCTION

In recent years, heavy flooding in Australia has posed a severe environmental challenge due to severe weather. Property, transportation networks, and other infrastructure are all affected by floods [15]. The kinetics of the flora and the faunas between upstream and downstream also affect the ecosystem [1]. As a result, there has been an increase in demand for studies that examine different characteristics of precipitation, such as the severity and duration of rainfall. These studies have made the significant impact of research disciplines such as hydrology and water resource management [2-4]. For this paper, we will examine two regions in the Australian state of Victoria that are the most flood-prone areas in the state. North-eastern and southern-western Victoria was the regions that experienced extremely high annual rainfall between 1950 and 2010 and have recently experienced massive flooding. [5].

Six (6) stations from these regions were chosen for having some of the highest annual mean rainfalls. This study uses empirical copulas to model rainfall severity and duration. For statisticians working with bivariate random variables with various marginal distributions, a copula is an extremely useful tool. Prior to installing copulas, the Standard Precipitation Index (SPI) is used to gauge the amount and duration of rain. This was accomplished making no assumptions about the prior distribution of rainfall intensity and by utilising a nonparametric density function based on the kernel density function. Because none of the parametric distributions currently in use can adequately characterise the rainfall intensity data for a number of stations in Victoria, this alternative method was used. We also used the empirical copulas in this study to estimate the dependence, duration and severity of the two rainfall characteristics. The Canonical Maximum Likelihood (CML) method is used to obtain empirical cumulative distribution functions (CMLs) from data to produce marginal distributions. A relationship between these rainfall

characteristics will be determined using the data simulated using the estimated copulas. As an alternative, we can use the Mean Absolute Error (MAE) criterion to compare the results obtained by employing the empirical copula to the results obtained by employing the theoretical copula.

This paper's outline is as follows. The introduction is the first section of this paper, followed by a description of the research area. The third section comprises the theoretical framework underlying our approach, including the Standard Precipitation Index (SPI), proposed kernel density function and Mean Absolute Error to assess goodness of fit. The application to Victorian precipitation data and results are reported in the next section. As a final note, the paper suggests some directions for future research.

STUDY AREA: VICTORIA

The State of Victoria in southeast Australia has a temperate climate characterized by mild to warm summers and cool winters. In Australia, this type of climate can be found in the southern and south-eastern coastal zones of the country. Recently, many areas of Victoria have been subjected to heavy rainfall events, after enduring severe drought conditions prevailing in most parts of the state for many years. For example, heavy rainfall events hit both Victoria and Queensland in 2010, when the number of rainy days exceeded the average annual wet days for these two states [5]. Due to these somewhat peculiar weather phenomena, opinions are sought from weather and hydrological scientists as to their probable causes, leading to a significant increase in studies of rainfall patterns. Some of these studies indicate that areas in the north-eastern and south-western Victoria are currently experiencing extreme rainfall events. Using data provided by the Australian Bureau of Meteorology [5] collected over 61 years (1950 to 2010), we choose six (6) rainfall stations located in these two regions for our case study. Table 1 summarizes these stations with their locations and annual mean rainfall.

TABLE 1. Summary of selected Stations in Victoria

Station No (mm)	Station Name	Latitude	Longitude	Annual Mean Rainfall (1950 – 2010)
83000	Archerton	36.91 °S	146.24 °E	1368.1
83033	Woods Point	37.57 °S	146.25 °E	1472.5
83073	Mount Buffalo Chalet	36.72 °S	146.82 °E	1882.6
90076	Tanybryn	38.68 °S	143.68 °E	1621.9
90083	Weeaproinah	36.84 °S	143.51 °E	1936.1
90087	Wyelangta	38.66 °S	143.45 °E	1949.8

THEORETICAL FRAMEWORK

Standard Precipitation Index

To measure the severity of rainfall levels, the Standard Precipitation Index (SPI) method was used to monitor and determine severe drought conditions. This SPI measurements were used before installing the copula. [6]. Over 60 countries are currently using the SPI in research or operational mode to monitor drought and rainfall intensities. One of the advantages of SPI is that it has a multiple time scales allow for temporal flexibility in evaluation of precipitation conditions and the water supply. The SPI for any area or region is calculated using long-term precipitation data. With multiple scales spreading out from the mean, this long-term precipitation record is fitted to a Gamma distribution. [7].

This study, on the other hand, was carried out with a nonparametric density function involving the kernel density function, with no assumptions about the prior distributions of rainfall intensity. Because none of the parametric distributions currently in use could adequately fit the rainfall intensity data collected at all stations across Victoria, this alternative method was required. A nonparametric approach involving kernel functions for SPI computation was

used to represent the variability of precipitation in the Colorado River basin [8] and [9] using the Standard Precipitation Index as a nonparametric approach for drought forecasting. The seasonal forecast of the SPI is addressed using nonparametric stochastic techniques, with the expectation of future SPI values deriving from past monthly precipitation. [9].

We used kernel density estimation to fit our precipitation data in this study. The cumulative distribution functions (cdfs) F will then be estimated using this nonparametric density estimator (y). The optimal bandwidth parameter must be chosen for this kernel smoothing technique. The following section gives insight into this technique in depth (Kernel Density Estimator). The nonparametric cdf F(y) of the rainfall event for different selected stations is then estimated using the optimal bandwidth parameter.

Using this cdf, the SPI is calculated to correspond to the precipitation such that,

$$\Phi(\text{SPI}) = F(y) \tag{1}$$

or

$$\text{SPI} = \Phi^{-1}[F(y)] \tag{2}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. Based on (1) it is therefore apparent that the SPI is the normal percentile corresponding to the precipitation cdf $F(y)$. The SPI calculation required at least 30 years of continuous monthly precipitation data. In this study, the data were first transformed into these indices in the manner described above and then subsequently used to generate the rainfall severity defined by

$$S = \sum_{i=1}^{D} \text{SPI}_i \tag{3}$$

where *D* is the rainfall duration of a severe rainfall period and *i* represents the month. Here, a severe rainfall period refers to the months where SPI values exceed 1. We use three-month moving totals for rainfall precipitation for the six rainfall stations in north-eastern and south-western Victoria, i.e., January-February-March total, then February-March-April total and repeating this pattern from January 1950 to December-2010. Seven-category classification for SPI [6] and [10] classified SPI value between -0.99 to 0.99 as near normal, positive value 1.0 to 1.49 is moderately wet, for condition very wet from 1.5 to 1.99, SPI value more than 2 considered extremely wet and a negative value of a positive value indicates a state of drought.

Kernel Density Estimator

In this paper, we used the kernel density method to fit our precipitation data. As mentioned earlier, to find the shape and structure in the data, SPI kernel smoothing generates rainfall severity and duration assuming none of the common parametric models used to model rainfall variables [11]. Kernel methods are commonly used in estimation of functions such as regression functions or probability density functions. The main advantage for a kernel approach is that it can discover structural features in a data which a parametric approach might miss.

Here, we give the definition of a univariate kernel density function given a random sample $X_1, \ldots, X_n$ with an unknown continuous, univariate probability density function $f(\cdot)$, the kernel density estimator of $f(\cdot)$ is defined by

$$f(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{4}$$

where $K(\cdot)$ is a kernel function and is the bandwidth. Referring to [11], under some mild conditions, the kernel density estimates of $f(\cdot)$ converges in probability to the true pdf. The kernel $K(\cdot)$ is usually chosen to be unimodal and symmetric about zero. In this paper we used the well-known Gaussian kernel which is defined by:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2) \tag{5}$$

For the bandwidth, we adopt the value given by [12] which is given by

$$h = 0.9An^{-1/5} \tag{6}$$

where $A =$ min {standard deviation, interquartile range}/1.34. Note that the nonparametric estimator $F(x)$ can be obtained from equation (4) by integrating the kernel density function resulting in

$$\begin{aligned} \hat{F}(x,h) &= \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} K(\frac{u - X_i}{h})\, du \\ &= \frac{1}{n} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \end{aligned} \tag{7}$$

where $\mathrm{K}(x) = \int_{-\infty}^{x} K(u)\, du$.

Empirical Copulas

The copulas essentially are mathematical objects which combine two or more marginal distributions. Some of the copulas have parameters that are based on correlations, where these parameters will have a function which known as the generator function as well as its inverse. All copulas have the same output which is known as a joint distribution. In this paper, we will focus on empirical copulas.

From Sklar Theorem [13], any bivariate cumulative distribution cdf can be expressed as:

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \tag{8}$$

for some function $C: [0,1]^2 \rightarrow [0,1]$, where F_X and F_Y denote the marginal cdfs. The function C represents the dependency between X and Y which joins the two marginal cdfs and is called a copula. In this study, we use the empirical copula [14] to obtain the joint cdf of rainfall characteristics. The empirical copula is defined by

$$C_n(u,v) = \frac{1}{n} \sum_{i=1}^{n} 1(\frac{Q_i}{n+1} \leq u, \frac{R_i}{n+1} \leq v) \tag{9}$$

where is the sample size of observations data, Q_i and R_i are the ranks of the empirical marginal cumulative distribution functions obtained from fitting the kernel cdfs to the marginal cdf $F_X(x)$and $F_Y(y)$ respectively. In equation (9) the notation 1(A) refers to the indicator function of the set A. The choice of the empirical copula is motivated by the fact that we rarely have sufficient information to determine which parametric copula will best fit bivariate precipitation data, in this case, rainfall severity and duration. Hence, we will use the data to determine the dependency structure empirically. The reader is referred to [3], for an extensive treatment on fitting parametric marginal cdfs and copulas to a real data set.

Application to Victorian Precipitation Data

This section is divided into two parts. The first part discusses the computational aspect of SPI using a nonparametric approach. In the second part we present an empirical copula analysis on rainfall severity and its duration of the selected stations listed in Table 1. Figure 1 shows a map indicating the locations of the selected rain-gauge stations that been used in this study.

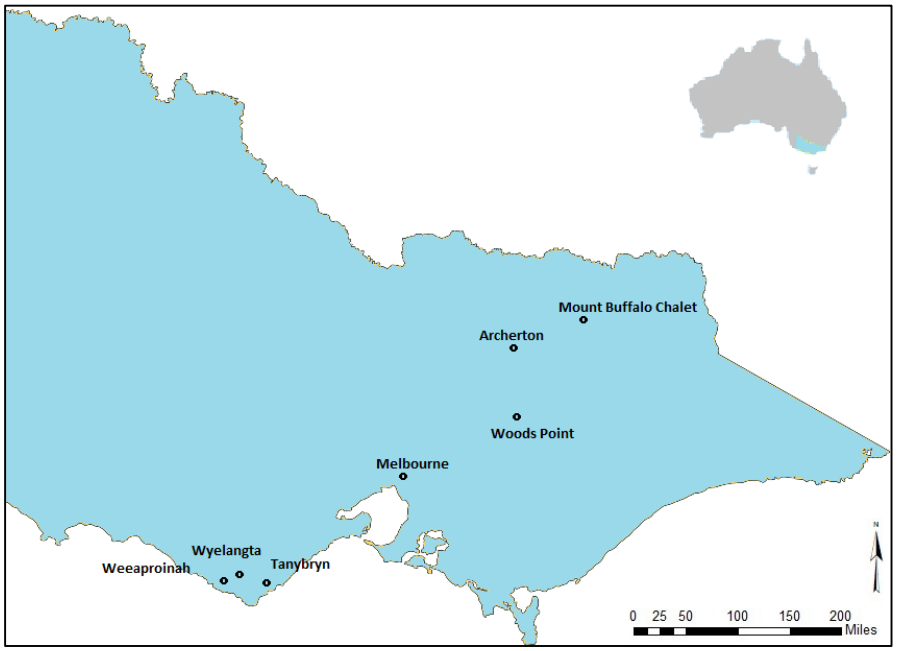

FIGURE 1. Selected Rain-gauge Stations at North-eastern and South-western Victoria

A similar method to that introduced by [5] will be used here for the SPI computation. However, in this study as discussed earlier, instead of fitting the precipitation time series to gamma probability density function, we used a kernel smoothing approach to obtain the estimated density function.

For example, Figure 2 shows the histogram from observed precipitation data for Archerton station (1950 to 2010). From the graph, it is apparent that the distribution is skewed to the right with modal rainfall intensity between 20mm to 40mm and a maximum value of about 420mm. The kernel density estimate of probability density function of rainfall intensity is displayed in Figure 3, using the Gaussian Kernel function with bandwidth computed from equation (6) to equal $h = 19.00$.

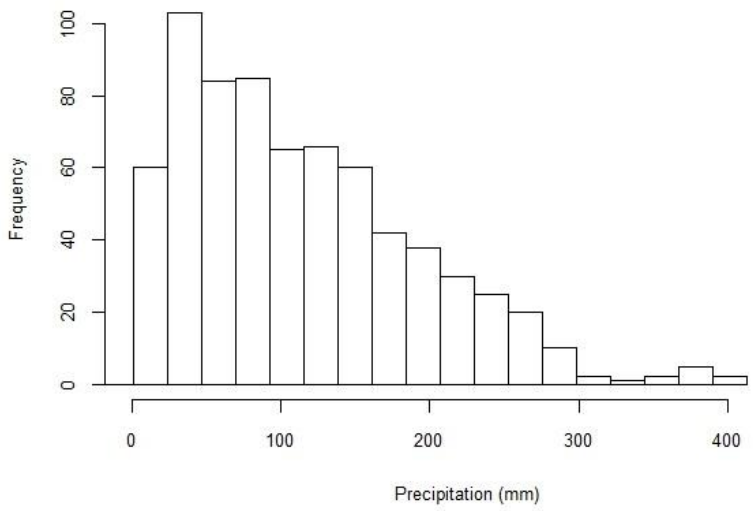

FIGURE 2. Histogram from observed Precipitation Data (monthly) – Archerton Station, Victoria (1950 to 2010)

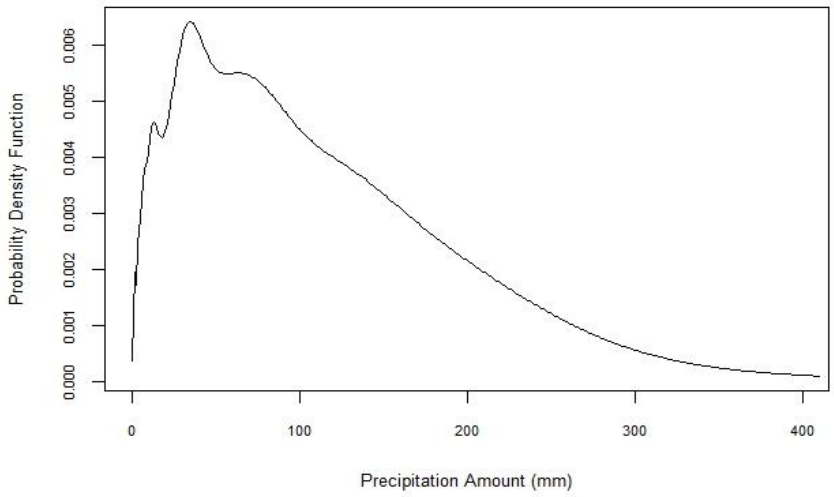

FIGURE 3. Gaussian kernel density estimate of Precipitation Data – Archerton Station, Victoria (1950 to 2010)

The nonparametric cdfs $F(y)$of the rainfall events for different selected stations were then calculated using equation (7). Finally, we obtained SPI values using equation (1) and (2). Figure 4 displays the three-monthly SPI for all six (6) stations in Victoria. Between 1950 to 1960, all stations show at least six occurrences with SPI values exceeding 2, which, according to Table 2, indicate extremely wet rainfall events during that period.

Referring to the SPI values between 2008 and 2010 show that five (5) stations, not including Tanybryn, experienced extremely wet event in 2010. However, Tanybryn station experienced very wet events with SPI values between 1.5 to 1.99.

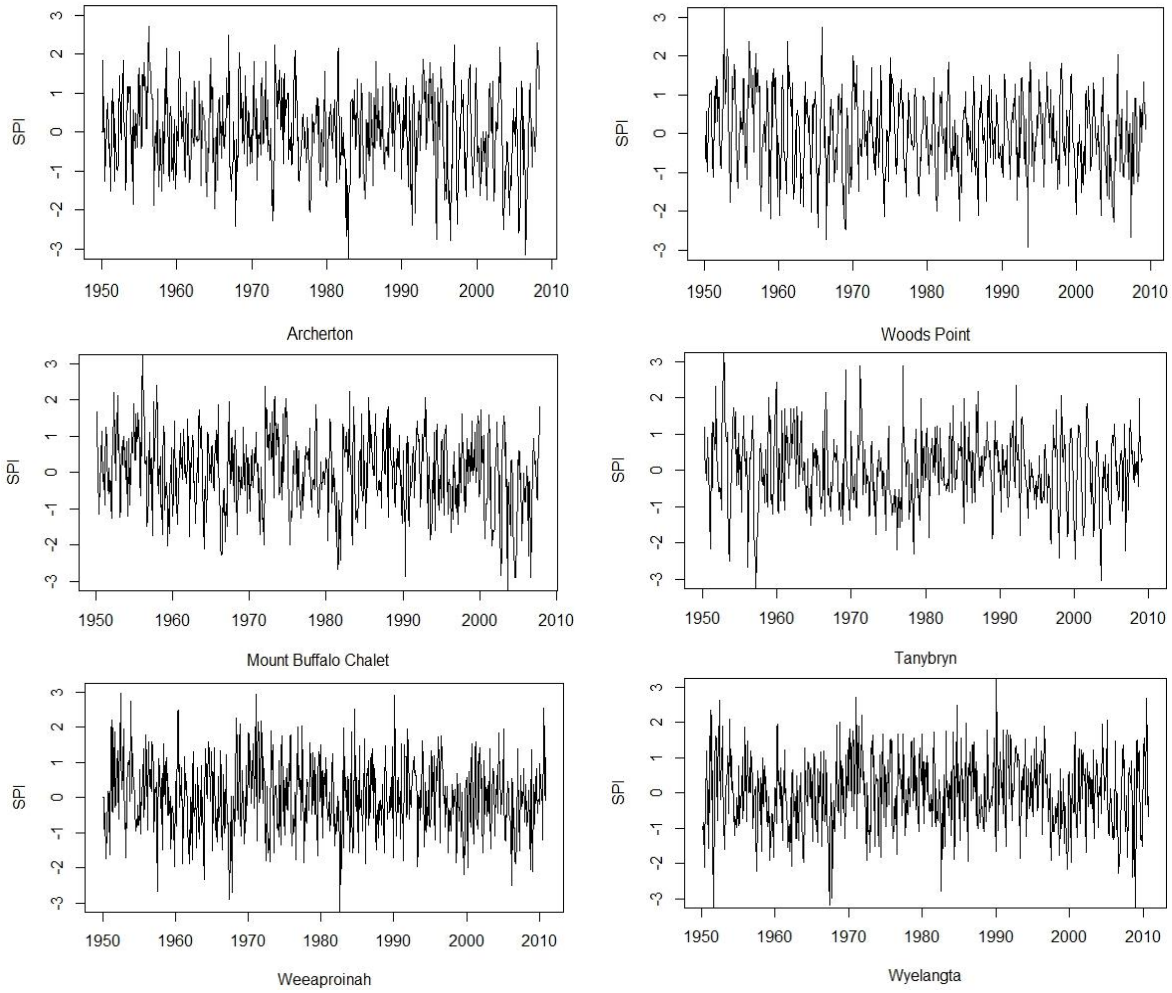

FIGURE. 4. The monthly SPI of six selected stations, Victoria (1950 to 2010)

Copulas Analysis

We compare our results obtained from the empirical copula with those obtained using three well-known theoretical copulas, i.e. Clayton, Frank and Gumbel-Hougaard. The parameters of the theoretical copulas were obtained using the Canonical Maximum Likelihood (CML) method. Mean Absolute Error (MAE) defined in equation (10) below was then used to compare both theoretical and empirical copulas. MAE is defined by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |C - C_{ni}| \tag{10}$$

where C refers to the values of the theoretical copula and C_n the empirical copula. The results are displayed in Table 3 below.

Table 3. MAE value for theoretical and empirical copula

Station Name	Types of Copulas	Mean Absolute Error
Archerton	Clayton	0.063
	Frank	0.079
	Gumbel-Hougaard	0.065
Woods Point	Clayton	0.081
	Frank	0.069
	Gumbel-Hougaard	0.068
Mount Buffalo Chalet	Clayton	0.074
	Frank	0.052
	Gumbel-Hougaard	0.066
Tanybryn	Clayton	0.079
	Frank	0.087
	Gumbel-Hougaard	0.066
Weeaproinah	Clayton	0.087
	Frank	0.085
	Gumbel-Hougaard	0.077
Wyelangta	Clayton	0.081
	Frank	0.078
	Gumbel-Hougaard	0.062

From Table 3, we may conclude that the empirical copula approximates the theoretical copulas closely and is therefore a close sample-based representation of the theoretical copula. Figure 5 compares between the values of the empirical copula and the true copula on the unit square, and again indicate a high linear correlation between the two set of values.

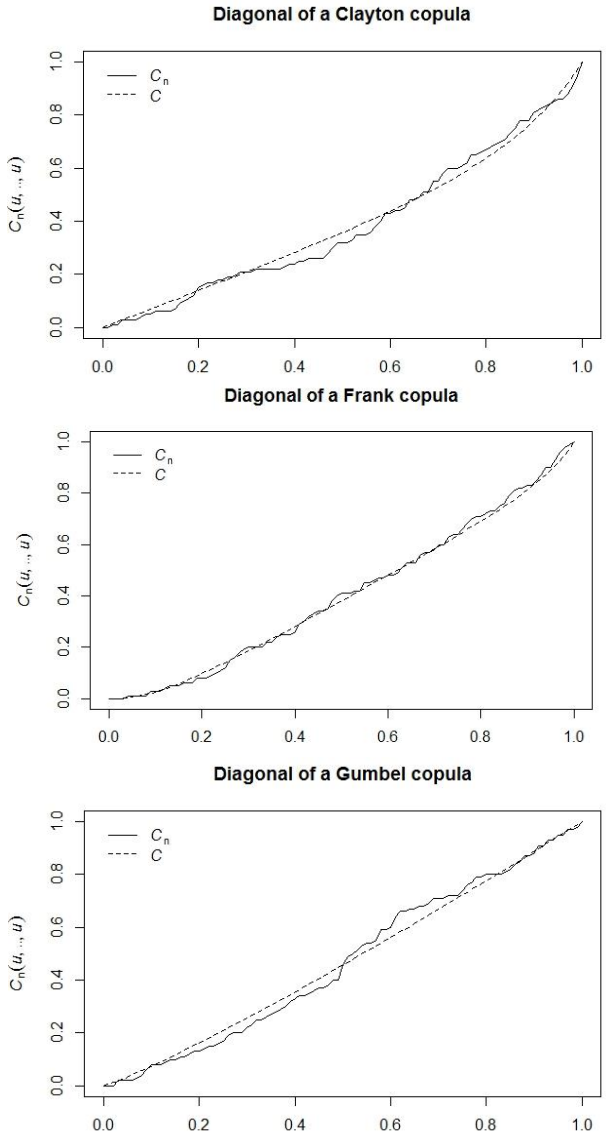

FIGURE. 5. The comparison curve of empirical copula and the true copula c.d.f on the diagonal of the unit square

CONCLUSION

At six (6) stations across Victoria, we used copula techniques to investigate two significant rainfall characteristics. The bivariate distributions of rainfall severity and duration are modelled using empirical copulas. Prior to fitting copulas, a nonparametric density function is used to estimate the Standard Precipitation Index (SPI). The marginal distributions of each of the rainfall characteristics in the copula function's argument are then fitted using an empirical distribution. We can conclude from the Mean Absolute Error values that the empirical copula can be used to represent the results obtained with standard copulas.

In conclusion, the nonparametric methodology discussed in this article has enabled the copula methodology to be extended further for the analysis of severe weather events. The analysis attempted in this paper is quite broad in scope and can easily be adapted to analyse climate data from other parts of Australia, which experience drier weather conditions or have an erratic rainfall pattern.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Bureau of Meteorology, Australia (BOM) for providing the complete historical precipitation data used in this study.

REFERENCES

1. Queensland Government, Department of Environment and Heritage Protection (2013). Flood Impacts. Retrieved from http://www.ehp.qld.gov.au/recovery/flood-impacts.html
2. U. F. Abdul Rauf and P. Zeephongsekul, “Modelling rainfall severity and duration in north-eastern Victoria using copulas”, Proceedings of the 19th International Congress on Modelling and Simulation (MSSANZ , Perth, 2011), pp. 3462-3468.
3. U. F. Abdul Rauf and P. Zeephongsekul (2014). Copula based analysis of rainfall severity and duration: a case study. Theor. Appl. Climatol., 115, 153–166
4. S. C. Kao, and R. S. Govindaraju (2008) Trivariate statistical analysis of extreme rainfall events via the placket family of copulas. Water Resour. Res. 44:1-19
5. Australian Bureau of Meteorlogy (2010). Weather and Climate Data. Retrieved from http://www.bom.gov.au/climate/data/
6. T. B. Mackee, N. J. Doesken, and J. Kleist (1993). "The relationship of drought frequency and duration to time scales." 8th Conference on Applied Climatology 179-184. Anaheim, California.
7. D. C. Edward and T. B. McKee (1997). "Characteristics of 20th century drought in United State at multiple time scales." Climatology Report 97-2, Colorado State University, Fort Collins, CO, USA.
8. T. W. Kim, B. Juan, B. Nijssen and D. Roncayolo (2006). Quantification of linkages between large-scale climatic patterns and precipitation in the Colorado river basin. J. Hydrol., 321(14):173-186.
9. A. Cancelliere, B. Bonaccorso and G. Di Mauro, Drought forecasting using the Standardized Precipitation Index (Water Resour. Manag., 2007) pp. 801–819.
10. T. B. Mackee, N. J. Doesken and J. Kleist (1995). Drought monitoring with multiple time scales. Proceedings of the 9th Conference on Applied Climatology, Boston, Massachussetts, pp. 3462-233-236.
11. M. P. Wand and M. C. Jones, Kernel Smoothing (Chapman and Hall, London, 1995).
12. B. W. Silverman, Density Estimation for Statistics and Data Analysis (Chapman and Hall, London, 1986).
13. R. B. Nelsen, An Introduction to Copulas (Springer-Verlag, New York, 1999).
14. P. Deheuvels (1979). "La function de dpendance empirique et ses propits. Un test non paramtrique d’independance" Acad.Roy.Belg.Bull.Cl.Sci. 65:274292.
15. Glass, C. E. (2013). “Dangers from Floods” in Interpreting Aerial Photographs to Identify Natural Hazards (Elsevier, 2013), pp. 111-122