2009 IEEE International Conference on Data Mining Workshops

Motivating Complex Dependence Structures in Data Mining: A Case Study with Anomaly Detection in Climate

Shih-Chieh Kao[1], Auroop R. Ganguly[1], and Karsten Steinhaeuser[1,2]
[1] Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[2] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
e-mail: {kaos, gangulyar, steinhaeuserkj}@ornl.gov

Abstract *- While data mining aims to identify hidden knowledge from massive and high dimensional datasets, the importance of dependence structure among time, space, and between different variables is less emphasized. Analogous to the use of probability density functions in modeling individual variables, it is now possible to characterize the complete dependence space mathematically through the application of copulas. By adopting copulas, the multivariate joint probability distribution can be constructed without constraint to specific types of marginal distributions. Some common assumptions, like normality and independence between variables, can also be relieved. This study provides fundamental introduction and illustration of dependence structure, aimed at the potential applicability of copulas in general data mining. The case study in hydro-climatic anomaly detection shows that the frequency of multivariate anomalies is affected by the dependence level between variables. The appropriate multivariate thresholds can be determined through a copula-based approach.*

I. INTRODUCTION

Due to the need for handling massive and high dimensional datasets, many statistical methods, such as multivariate regression analysis [1], multivariate analysis of variance (MANOVA [2]), principal component analysis, clustering analysis, geostatistics (e.g., Kriging method [3]), autoregressive integrated moving average model (ARIMA [4]) and Bayesian analysis, have been adopted in data mining to abstract useful information hidden in large datasets. For the sake of simplification, some of these methods assume the randomness components to be normally distributed or spatially/temporally/inter-variably independent. However, the actual data are mostly spatio-temporal correlated with non-trivially individual probability distributions across multiple variables. Though the importance of dependence has been emphasized by studies such as spatio-temporal data mining [5, 6], the influence of dependence structure on various domains and applications could be even boarder.

One major challenge toward modeling the multivariate probability space is our lack of an effective mathematical tool to characterize the dependence structure between variables. Comparing to the procedures of univariate statistical analysis, after obtaining general ideas from the basic moment-based statistics (i.e., mean, standard deviation, coefficients of skewness and kurtosis), one can go one step further to identify the most appropriate univariate probability density function (PDF, like Gaussian, Student *t*, extreme value distributions, or kernel-based probability density estimator) to study all statistically-relevant problems such as risk, statistical similarity, extreme value and frequency, anomaly detection, and data simulation. In other words, the 1-dimensional PDF provides better capabilities in various applications than single-value moments.

However, unlike PDF toward moments, what corresponded to cross-moments (e.g., correlation coefficient) is not obvious. While different metrics like Pearson's correlation coefficient ρ, Gini's measure of association γ, Kendall's concordance measure τ, Spearman's rank correlation r [7] and mutual information (MI [8]) have been used extensively to describe the correlation and dependence in different applications, they are all single-value statistics with case-specific limitations. It is likely that once the pattern of dependence between variables becomes very complicated (e.g., non-linear processes or dependent only to a localized region), the conventional metrics may not be sufficient to detect the hidden association. What is needed is a detailed mathematical description to the entire dependence space, and such capability was not noticed until recently. As we show in Section 2, copulas are one satisfactory solution to model the dependence structure.

For domains like hydro-meteorological and climatic analyses, a generalizable description of dependence structure is desirable. There exist huge amounts of remote sensing data, gauge observations and climatic modeling outputs with complicated spatio-temporal and inter-variable dependence for investigation, in which most of the natural hydrologic variables (e.g., precipitation and streamflow) are far from the scope of normal distribution. Some natural hazards like extreme storms and droughts also involve a great amount of correlated meteorological variables (temperature, precipitable water, evaporation, wind speed, and so on), which make the risk assessment for climate change even more challenging. The long-distance teleconnection is also among the first priority to investigate as it may eventually lead to uncertainty reduction in climatic projections.

Since the characterization of dependence structure via copulas is relatively new to the general data mining community, this study aims to provide a fundamental introduction to the potential application of copulas in data mining. The definition, illustration, and background information of dependence structure and copulas are provided in Section 2. Focusing on the domain of climate extremes analysis, a case study utilizing copulas in climatic anomaly detection will be presented in Section 3. We show

978-0-7695-3902-7 2009

DOI 10.1109/ICDMW.2009.37

how the positive/negative dependence levels between variables affect the number of detections of multivariate anomalies. A copula-based method for threshold adjustment to detect the same amount of anomalies under various dependence levels is also proposed and tested. Finally, the summary and concluding remarks are presented in Section 4. Through taking climate as the case study, we would like to point out that this method is also potentially applicable to other problems as well, with examples in fraud detection, marketing, bioinformatics, earlier disease detection [9], and intrusion detection [10].

II. Dependence Structure and Copulas

Though the most central theory of copulas was proposed early in 1959 [11], copulas did not receive much attention until the recent decade. The successful implementation of copula-based statistical analysis on multivariate financial dataset suggests its general applicability in other domains as well. Copulas are also becoming popular in water resources and hydro-climatic analysis [12-14]. This Section aims to provide some basic explanations and illustrations to help understanding the concept of dependence structure and copulas. More mathematical details can be found in standard textbooks of copulas like [7, 15].

A. Definition and Illustration of Copulas

The first usage of "Copula" is attributed to Sklar [11] in a theorem describing how one-dimensional distribution functions can be combined to form multivariate distributions. For *d*-dimensional continuous random variables $\{X_1,...,X_d\}$ with joint cumulative distribution function (CDF) $H_{X_1,...,X_d}$ and marginal CDFs $u_j = F_{X_j}(x_j)$, $j = 1,...,d$, Sklar showed that there exists one unique *d*-copula $C_{U_1,...,U_d}$ such that:

$$C_{U_1,...,U_d}(u_1,...,u_d) = H_{X_1,...,X_d}(x_1,...,x_d) \qquad (1)$$

Since u_j can also be interpreted as the transformation of x_j from $[-\infty,\infty]$ to $[0,1]$, copulas $C_{U_1,...,U_d}$ is a mapping of $H_{X_1,...,X_d}$ from $[-\infty,\infty]^d$ to $[0,1]^d$. The consequence of this transformation is that the marginal CDFs are segregated from $H_{X_1,...,X_d}$, and hence $C_{U_1,...,U_d}$ becomes only relevant to the association between variables. In other words, $C_{U_1,...,U_d}$ gives a complete mathematical characterization of the entire dependence structure.

An illustration is shown in Figure 1, where the standard bivariate Gaussian joint-PDFs with ρ = 0 and ρ = 0.5 are plotted in (a) and (d), and the corresponding realizations are shown in (c) and (f). Though variables *X* and *Y* in (a) are independent, it can hardly be identified from the shape of joint-PDF h_{XY} in (a), mainly because h_{XY} is a mixture of both marginal PDFs and dependence structure. The advantage of copulas can be clearly seen from the copula density ($\partial^2 C_{UV} / \partial u \partial v$) plots in (b) and (e). When *X* and *Y* are independent, the corresponding copula density will be a horizontal surface, indicating equal probability in any pair of (*u*, *v*). On the other hand, when *X* and *Y* are positively dependent (ρ = 0.5), more copula density will fall in near the main diagonal *u* = *v*, and much less density in regions [*u* < 0.5, *v* > 0.5] and [*u* > 0.5, *v* < 0.5]. It means that there will be more probability for low-low and high-high pairs of (*u*, *v*), but less probability for low-high and high-low ones, which corresponds to the anticipation of positive correlation. Therefore, the plot of copula density can be used effectively to examine the dependence pattern, just as the role of PDFs for marginal variables.

A further illustration of some commonly used copulas (densities) is shown in Figure 2. Student *t* copulas with degree of freedom ν = 2 is illustrated in (a), Frank copulas in (b), and Clayton copulas in (c). The copula parameters are determined so that all these copulas will have the same correlation coefficient ρ = 0.5. From the distinct appearances of various copula densities in Fig. 1(e) and Figure 2, it is obvious that the Pearson's correlation coefficient ρ, or other single-value dependence measures, may not be sufficient to capture the complete dependence space. The use of copula functions will provide more flexibility in handling various types of challenges.

For Student *t* copulas, since they are derived from the multivariate Student *t* distribution, the feature of heavier tails is retained. Comparing Fig. 2(a) to Fig. 1(e), it can be observed that more copula densities are fallen in the four tail regions, and hence Student *t* copulas are useful to model heavy-tail dependence structure in the multivariate extreme value analysis. Though both Gaussian and Student *t* copulas (or more generally, meta-elliptical copulas [16]) are derived from the well-known multivariate Gaussian and Student *t* distributions, they do not have an explicit mathematical expression to work with. Therefore, other choices of copulas are also of great interest since they are easier to manipulate and are more computationally efficient, which is a desired feature in data mining applications. Both Frank and Clayton copulas shown in Fig. 2(b) and (c) belong to a special copula class - Archimedean copulas, which will be discussed in more details in the following subsection.

B. Family of Archimedean Copulas

Among various types of copulas, one-parameter Archimedean copulas have attracted the most attention owing to their convenient properties. For an Archimedean copula, there exists a generator φ such that the following relationship holds:

$$\varphi(C(u,v)) = \varphi(u) + \varphi(v) \qquad (2)$$

The generator φ is a continuous and strictly decreasing function defined in [0,1], and $\varphi(1) = 0$. When $\varphi(t) = -\ln(t)$, the copula in (2) becomes $C(u,v) = uv$, which is a special case when the variables are independent (e.g., Fig. 1(b)). Frank and Clayton families of copulas are formulated as:

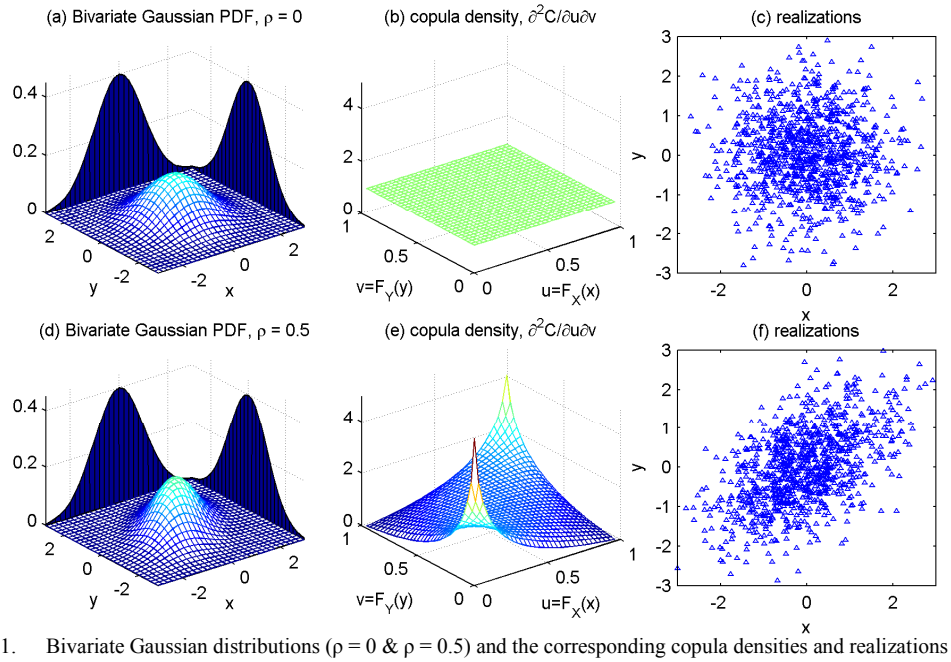

Figure 1. Bivariate Gaussian distributions (ρ = 0 & ρ = 0.5) and the corresponding copula densities and realizations

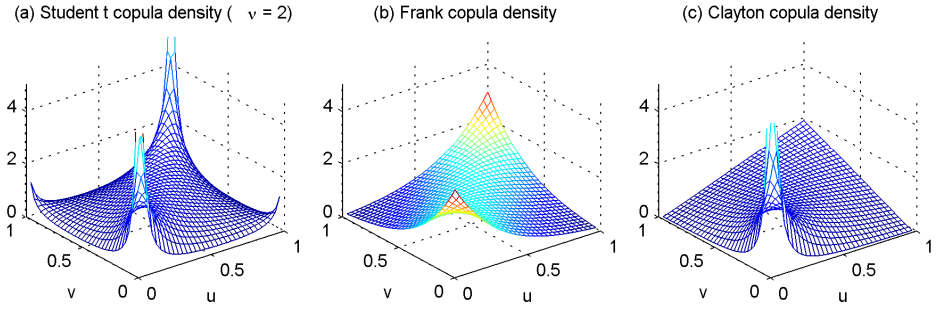

Figure 2. Illustration of (a) Student *t* copulas (degree of freedom $\nu = 2$), (b) Frank copulas, and (c) Clayton copulas. All three copulas have the same Pearson's correlation coefficient ρ = 0.5

Frank family of Archimedean copulas:

$$\begin{cases} \varphi(t) = -\ln((e^{-\theta t}-1)/(e^{-\theta}-1)) \\ C(u,v) = -\dfrac{1}{\theta}\ln(1+\dfrac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}) \\ \theta \in (-\infty,0)\cup(0,\infty) \end{cases} \tag{3}$$

Clayton family of Archimedean copulas:

$$\begin{cases} \varphi(t) = (t^{-\theta}-1)/\theta \\ C(u,v) = (\max(u^{-\theta}+v^{-\theta}-1,0))^{-1/\theta} \\ \theta \in [-1,0)\cup(0,\infty) \end{cases} \tag{4}$$

The appearance of Frank copulas can be seen in Fig. 2(b), in which the copula density is symmetric to both diagonals $u = v$ and $u + v = 1$. It has similar shape comparing to the Gaussian copulas in Fig. 1(e), but with explicit mathematical expressions that are easier to operate. Frank family of copulas is a popular choice for modeling bivariate dependence, and is found suitable for several types of hydrologic dependence structure [12].

Comparing to Frank copulas, Clayton copulas are symmetric only to the main diagonal $u = v$. As illustrated in Fig. 2(c), there will be much denser pairs of (u, v) in the low-low region than in the high-high region. Therefore, it is potentially suitable for cases with imbalanced local dependence within two tail regions. By transforming the original variable in a reversed order (i.e., $\hat{X} = -X$), the Clayton copulas can be fitted in different directions in order to reach an optimal use.

In Eqs. (3) and (4), θ represents the dependence parameter, and it can be estimated through the conventional maximum likelihood (ML) method [17]. Nevertheless, the ML estimator is not only related to dependence structure, but also determined by the goodness-of-fit of marginal distributions. Hence, the estimation errors of marginal variables may propagate to the dependence parameter. The existence of marginal outliers will also have a significant influence and may cause the estimator to be biased. Alternatively, a non-parametric procedure (NP) that is unique to the family of Archimedean copulas [18] can be

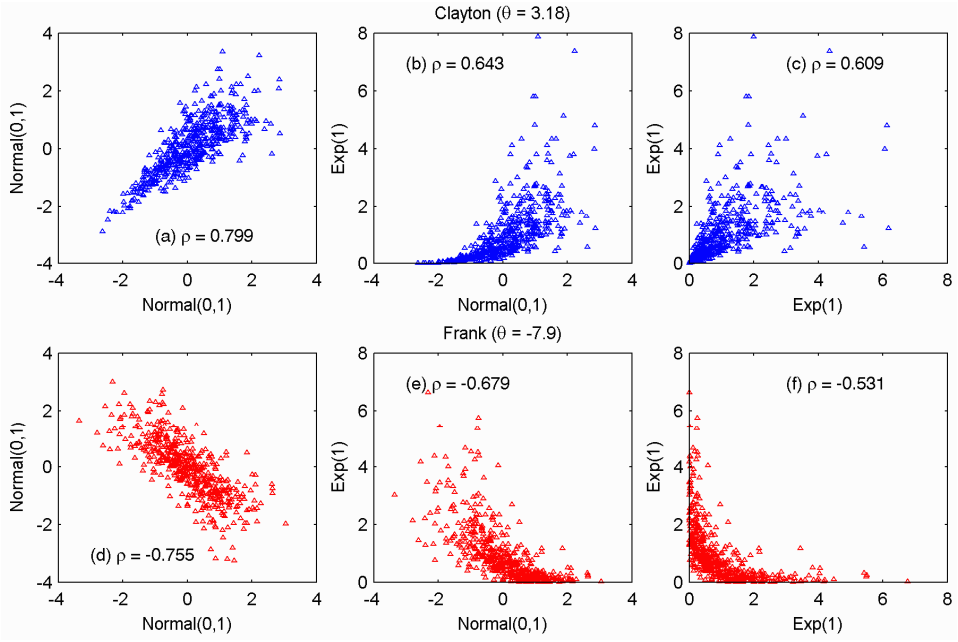

Figure 3. The capability of copulas in random number generation. All simulated patterns (500 data points in each panel) are combinations of Normal(0,1) and Exp(1) marginals, and Clayton(3.18) and Frank(-7.9) copulas

utilized. To obtain a NP estimator of θ, one starts with equating Kendall's concordance measure τ to φ as:

$$\begin{aligned}\tau &= 1+4\int_0^1(\varphi(t)/\varphi'(t))dt \\ &= 1-4[D_1(-\theta)-1]/\theta \quad \text{(Frank)} \\ &= \theta/(\theta+2) \quad \text{(Clayton)}\end{aligned} \tag{5}$$

where D_1 is the 1st order Debye function $D_1(\theta)=\int_0^\theta(t/\theta(e^t-1))dt$. Meanwhile, the sample Kendall's $\hat{\tau}$ can be estimated by:

$$\hat{\tau}=(c-d)/(n(n+1)/2) \tag{6}$$

where n represents the sample size, c denotes the number of concordant pairs $(x_2\text{-}x_1)(y_2\text{-}y_1) > 0$, and d denotes the number of discordant pairs $(x_2\text{-}x_1)(y_2\text{-}y_1) < 0$. By solving $\tau=\hat{\tau}$, the NP estimator θ can be obtained. This procedure proceeds independently from the analysis of marginal variables, and is more computationally efficient for large datasets. If Kendall's τ is regarded as a general cross-moment, then NP is conceptually similar to the well-known method of moment in estimating PDF parameters. Except using Kendall's τ, the dependence parameter can be estimated in a similar manner via other rank-based dependence measure such as Spearman's rank correlation r.

C. Empirical Copulas

To evaluate the suitability of a selected copula with estimated parameter, one can utilize empirical copulas as the observed dependence structure for evaluation. Similar to the concept of plotting position formula used in univariate statistical analysis (e.g. Weibull formula), empirical copulas are rank-based empirically joint cumulative probability measures [7]. For sample size n, the d-dimensional empirical copula C_n is:

$$C_n(k_1/n,k_2/n,...,k_d/n)=a/n \tag{7}$$

where a is the number of samples $\{x_1,...,x_d\}$ with $x_1 \le x_{1(k_1)}$, …, $x_d \le x_{d(k_d)}$, and $x_{1(k_1)}$, …, $x_{d(k_d)}$ with $1 \le k_1,...,k_d \le n$ are the order statistics from the sample. The C_n can be applied in different goodness-of-fit tests, including multidimensional Kolmogorov- Smirnov (KS) test [19], tests based on the probability integral transformation [20], kernel-based smoothing techniques [21], and cross product ratio model [22].

D. Random Number Generation

Copulas also provide a convenient way for generating correlated random variables. To generate jointly-distributed random variables (x,y) from a given joint CDF H_{XY}, the first step is to generate independently uniformly-distributed random pairs (u,t) from 0 to 1. Since the conditional probability $P[V \le v \mid U=u]$ will be independent to U, by equating $P[V \le v \mid U=u]=t$, v can be evaluated, and random pairs (u,v) will possess the dependence structure of H_{XY}. By transforming (u,v) via the inverse CDFs $x=F_X^{-1}(u)$ and $y=F_Y^{-1}(v)$, random variables (x,y) can be obtained. This general expression can be shown as [7]:

$$P[V \le v \mid U = u] = \frac{\partial C(u,v)}{\partial u} = t \quad (8)$$

For Frank and Clayton ($\theta > 0$) copulas, the explicit equations can be further obtained (thanks to the simplicity provided by Archimedean copulas):

$$v = \frac{1}{\theta}\ln\left(\frac{1-t+te^{u\theta}}{1-t+te^{(u-1)\theta}}\right) \text{ (Frank)} \quad (9)$$

$$v = \left(1-u^{-\theta}+t^{-\theta/(1+\theta)}u^{-\theta}\right)^{-1/\theta} \text{ (Clayton, } \theta > 0) \quad (10)$$

An example of random number generation is shown in Figure 3, where various combinations of a positive dependence structure by Clayton copulas (θ = 3.18), a negatively dependence structure by Frank copulas (θ = -7.9), standard normal marginals (mean 0 and standard deviation 1) and exponential marginals (mean 1) are illustrated. The dependence parameters are evaluated so that the theoretical correlation coefficient for Clayton copulas is 0.8 and for Frank copulas is -0.8. After simulated 500 data points in each case, the sample correlation coefficient ρ is also reported.

It can be observed from Figs. 3(a) and 3(d) that sample correlations are close to the theoretical ones when copulas are associated with normal distributions. Nevertheless, by replacing normal marginals to exponential distributions (b, c, e and f), the sample correlations become weaker, even if all (a-c) and (d-f) cases share the same dependence structures. It highlights the difficulty in identifying dependence patterns with the interference of marginal distributions. The various patterns simulated in Fig. 3 also demonstrate the flexibility of copula-based random number generation techniques in many different applications.

III. CASE STUDY

The specific challenges of implementing data mining techniques in climate analysis are:

- Climate data contains multiple variables (e.g., temperature, pressure, wind speed, humidity, precipitable water and precipitation). Each variable has its own type of distribution, seasonal variability, and long-term non-stationary trend.
- The extensive datasets range across various temporal (6-hourly, daily, monthly, or annually) and spatial (mostly ranging from 1° to 5° square grids) resolutions, resulting in Terabytes to Petabytes of data with unknown spatio-temporal dependence structure and long-distance teleconnection
- The Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) [23] is an internationally joint effort of 23 global circulation models (GCMs). Associated with various emission scenarios, there are plenty of data for analysis, comparison and assessment.
- Most of the hydro-meteorological variables are governed by non-linear processes with non-intuitive mechanisms.

With the progressing data mining capabilities [24], methods such as anomaly detection, similarity measures, classification, clustering, association rule, social network and spatio-temporal data mining are potential to explore. It is our hope that via the strength of different data mining techniques, useful and hidden insights can be discovered, the complexity of climate data can be reduced, and eventually our understanding toward the interwoven climate system will be improved. Copulas are applied for multivariate climatic anomalies detection (e.g., co-occurrence of hot and dry weather events) to demonstrate how dependence affects the number of detections, and what can be improved via the use of copula-based techniques.

A. Data Source and Pre-processing

The National Centers for Environmental Prediction - Department of Energy Atmospheric Model Intercomparison Project Reanalysis (NCEP2 [25]) is adopted in this study. NCEP2 data are provided at 1.9°x1.9° spatial resolution (total of 18,048 grid cells), beginning from 1979 until present. Reanalysis data such as NCEP2 are usually treated as proxy of observations and are used extensively in various hydro-meteorological studies. Monthly temperature $\widetilde{X}_{k,l}^{(i,j)}$, precipitation $\widetilde{Y}_{k,l}^{(i,j)}$, and precipitable water $\widetilde{Z}_{k,l}^{(i,j)}$ on the 7,194 land grid cells are selected to study the observed climate anomalies (months that are both abnormally hot and dry). Notation $\widetilde{X}_{k,l}^{(i,j)}$ is interpreted as the monthly temperature at grid (i,j) in month k of year l, in which $i = 1,...,192$, $j = 1,...,94$, $k = 1,...,12$, and $l = 1979,...,2008$.

In order to alleviate the influence of seasonal variability, we perform the z-score transformation [26]. Taking temperature as an example, the transformation follows Eqs. (11-13):

$$\overline{X}_k^{(i,j)} = \sum_{l=1979}^{2008} \widetilde{X}_{k,l}^{(i,j)} / n \quad (11)$$

$$Sx_k^{(i,j)} = \sum_{l=1979}^{2008} (\widetilde{X}_{k,l}^{(i,j)} - \overline{X}_k^{(i,j)})^2 / (n-1) \quad (12)$$

$$X_{k,l}^{(i,j)} = (\widetilde{X}_{k,l}^{(i,j)} - \overline{X}_k^{(i,j)}) / Sx_k^{(i,j)} = X_t^{(i,j)} \quad (13)$$

In (11), the average temperature $\overline{X}_k^{(i,j)}$ for month k (January, February, …) at each grid (i,j) is computed, and fed to (12) to calculate the corresponding standard deviation $Sx_k^{(i,j)}$. In (13), both $\overline{X}_k^{(i,j)}$ and $Sx_k^{(i,j)}$ are used to normalize the original $\widetilde{X}_{k,l}^{(i,j)}$ to be a unit-free time series $X_t^{(i,j)}$, in which t denotes the t^{th} month since January, 1979. Same

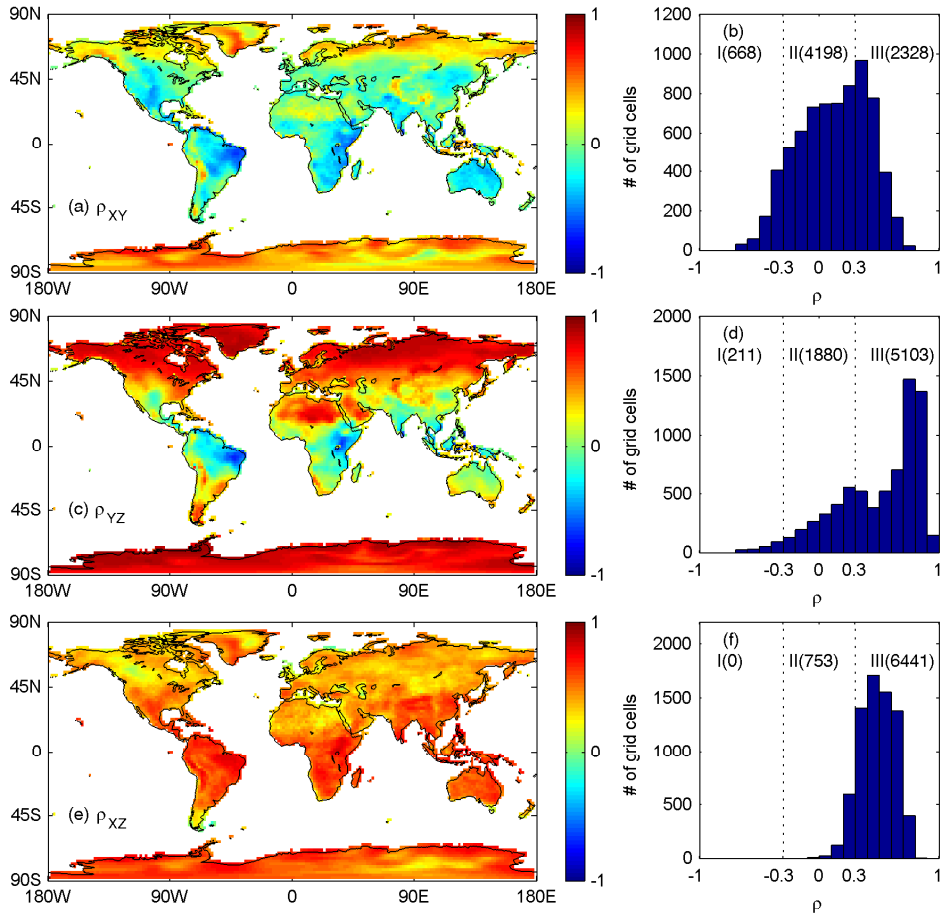

Figure 4. Global maps and histograms of correlation coefficient ρ for (Plots a & b) normalized temperature (X) versus normalized precipitation (Y), (Plots c & d) Y versus normalized precipitable water (Z), and (3) X versus Z. Number of grid points in Region I (ρ < -0.3), Region II (-0.3 < ρ < 0.3), and Region III (ρ > 0.3) are also marked on the histograms.

procedure holds to derive the normalized precipitation $Y_t^{(i,j)}$, and precipitable water $Z_t^{(i,j)}$.

B. Dependence Structure

Taking ρ as the starting point, the correlation coefficient between each pair of normalized variables is computed. The results are illustrated as maps and histograms in Figure 4. To assist discussion, three categories are defined, namely Region I (ρ < -0.3, negatively dependent), Region II (-0.3 < ρ < 0.3, near independent), and Region III (ρ > 0.3, positively dependent). The total number of grid cells fallen in each region is also reported. It can be observed that the correlation ranges widely from negative to positive between *X* & *Y*, and mostly positive between *X* & *Z*, and *Y* & *Z*. The correlation between normalized precipitation and precipitable water is especially strong that there is no negative correlation in Region I.

To further investigate the dependence structure and construct the joint probability distributions, the copula-based approach can be applied at each grid. Selecting the grid nearest Miami, FL (79.67W~81.56W, 24.76N~26.67N) as an example, the fitting of marginal distributions and dependence structure are shown in Figure 5. Without involving further assumptions of parametric CDFs, the kernel density estimators were utilized to derived the marginals $u = F_X$, $v = F_Y$ and $w = F_Z$. It can be observed from Fig. 5(a-c) that even with the z-score transformation, variables can still be non-Gaussian distributed.

As for the dependence structure, Frank family of Archimedean copulas (Eq. 3) is chosen. The dependence parameters are estimated via the NP approach (Eq. 5), and three bivariate copulas, C_{UV}^{Frank}, C_{UW}^{Frank} and C_{VW}^{Frank} between each pair of marginals are derived. In addition, the empirical copulas (Eq. 7) are computed for model verification. Fig. 5 (d-f) illustrate the difference between Frank and empirical copulas, and the small differences (~±0.02) suggest the suitability of Frank copulas. By combining both marginal distribution and dependence structure, the bivariate joint-CDF can be modeled, i.e. $P[X \le x, Y \le y] = C_{UV}(F_X(x), F_Y(y))$. The joint-CDF will be a handy tool in solving all kinds of statistical-related problems.

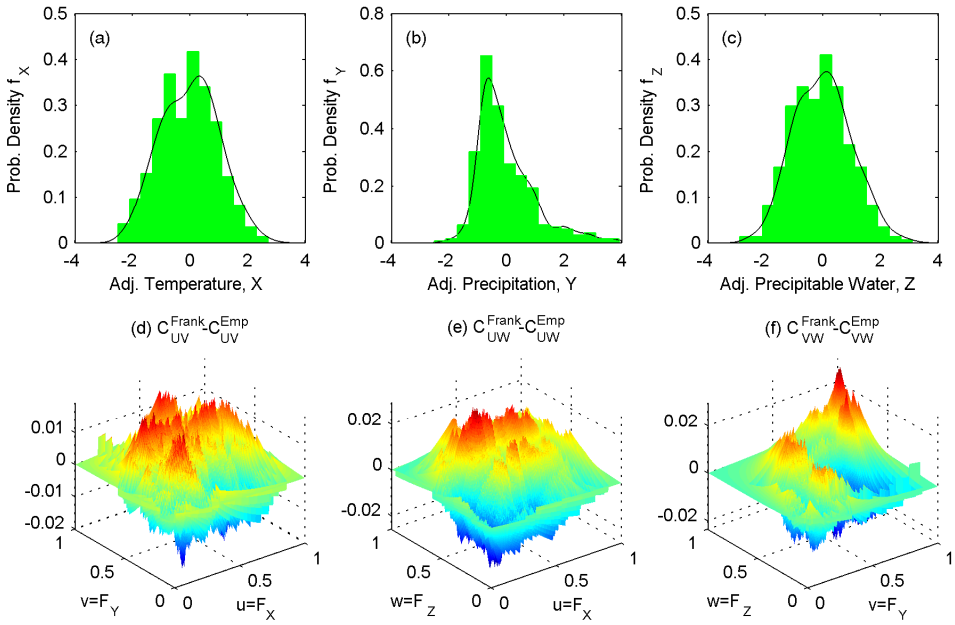

Figure 5. Taking the grid cell containing Miami as an example, (a-c) show the histograms and kernel density fitting of adjusted temperature (X), precipitation (Y), precipitable water (Z), and (d-f) show the differences between fitted Frank copulas and empirical copulas of each pair of variables.

C. *Climatic Anomaly Detection*

Since the correlation between hydro-climatic variables spans across a wide range, there is a need to investigate how different levels of dependence will affect the detection of multivariate anomalies. We will focus on identifying the observed heat wave and drought events. In other words, months with higher temperature, less precipitation, and less precipitable water with respect to the given thresholds are identified and categorized as anomalies.

In case 1, a fixed threshold q = 20% is set. The upper 20% temperature percentile ($x_{80\%}^{(i,j)}$), lower 20% precipitation percentile ($y_{20\%}^{(i,j)}$) and precipitable water percentile ($z_{20\%}^{(i,j)}$) at each grid cell are identified. Clearly, the observed frequency for univariate events such as $\{X^{(i,j)} \geq x_{80\%}^{(i,j)}\}$ will naturally be q. By using these three thresholds, the observed frequencies of bivariate anomalies $\{X^{(i,j)} \geq x_{80\%}^{(i,j)}, Y^{(i,j)} \leq y_{20\%}^{(i,j)}\}$, $\{X^{(i,j)} \geq x_{80\%}^{(i,j)}, Z^{(i,j)} \leq z_{20\%}^{(i,j)}\}$, and $\{Y^{(i,j)} \leq y_{20\%}^{(i,j)}, Z^{(i,j)} \leq z_{20\%}^{(i,j)}\}$ are computed for each regions defined in Fig. 4. The results are then summarized in Table 1.

It can be seen that the observed frequencies in different regions vary a lot. When looking for dry and wet events, more cells in Region I will be detected since X and Y are negatively correlated (more probability for high-low pairs). On the other hand, much less cells will be detected in Region III. When looking for dry and low precipitable water events, more cells will be detected in Region III because of less probability for low-low pairs. Since it is nearly independent in Region II, the theoretical frequency will be around 4%, and it is confirmed in Table 1. This result suggests that the anomalies in different regions need to be interpreted differently. A hot and dry event will appear to be much rare in Region III than Region I. If the dependence structure between variables is not considered properly, we may not detect the appropriate multivariate anomaly pairs for analysis.

TABLE I. SUMMARY OF BIVARIATE CLIMATIC ANOMALY DETECTION

	Regions		
	I	**II**	**III**
Case 1, fixed threshold			
$\{X^{(i,j)} \geq x_{80\%}^{(i,j)}, Y^{(i,j)} \leq y_{20\%}^{(i,j)}\}$	8.41%	4.39%	1.43%
$\{X^{(i,j)} \geq x_{80\%}^{(i,j)}, Z^{(i,j)} \leq z_{20\%}^{(i,j)}\}$	8.38%	3.67%	0.52%
$\{Y^{(i,j)} \leq y_{20\%}^{(i,j)}, Z^{(i,j)} \leq z_{20\%}^{(i,j)}\}$	-	5.98%	8.96%
Case 2, adjusted threshold			
$\{X^{(i,j)} \geq x_{1-q1}^{(i,j)}, Y^{(i,j)} \leq y_{q1}^{(i,j)}\}$	4.81%	4.48%	4.47%
$\{X^{(i,j)} \geq x_{1-q2}^{(i,j)}, Z^{(i,j)} \leq z_{q2}^{(i,j)}\}$	4.52%	4.24%	4.85%
$\{Y^{(i,j)} \leq y_{q3}^{(i,j)}, Z^{(i,j)} \leq z_{q3}^{(i,j)}\}$	-	4.15%	4.45%

One solution is to identify an adjusted threshold q_1 for different dependence levels so that the similar amount of information can be detected. We propose to use the copula-based joint probability to achieve this goal. Generally, one needs to solve:

$$P[V \leq q_3, W \leq q_3] = q^2 \tag{14}$$

For Frank family of copulas, Eq. (14) can be further simplified as:

$$q_3 = -\frac{1}{\theta}\ln(1-\sqrt{(e^{-\theta}-1)(e^{-\theta q^2}-1)}) \qquad (15)$$

In other words, the probability contained by the new threshold q_3 should be the same as they are independent. Adopting the adjusted thresholds, the observed frequency is re-calculated and reported in Table 1. It is shown that similar amounts of anomalies are detected in all regions, suggesting the applicability of the proposed method. One can also assign different thresholds in Eq. (14), i.e. $P[V \le q_2, W \le q_3] = q^2$, and the same effect can be achieved for all (q_2, q_3).

IV. Conclusion and Future Work

To alleviate some common assumptions such as normality and independence between variables in data mining, we introduce the concept of dependence structure and the use of copulas in this study. Copulas provide a complete mathematical description to the entire dependence space, and hence have better capabilities for potential applications in various domains. The case study suggests that the dependence level should be accounted for to correctly detect the multivariate anomalies. Given the great flexibility of copulas, we expect that more data mining techniques would benefit from this robust method. Our future work will focus on a case-by-case comparison between existing data mining techniques and the copula-induced algorithms.

Acknowledgment

This research was funded by the Laboratory-Directed Research and Development (LDRD) Program of the Oak Ridge National Laboratory (ORNL), managed by UT Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725. The United States Government retains, and the publisher, by accepting this submission for publication, acknowledges that the United States Government retains, a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this submission, or allow others to do so, for United States Government purposes.

References

[1] D. E. Hershberger and H. Kargupta, Distributed Multivariate Regression Using Wavelet-based Collective Data Mining. *J. Parallel Distrib. Comput.*, 61(3), pages 372-400, 2001.

[2] J. H. Heinrichs and J.-S. Lim, Integrating Web-based Data Mining Tools with Business Models for Knowledge Management, *Decis. Support Syst.*, 35(1), pages 103-112, 2003.

[3] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*, Oxford University Press, 1990.

[4] B. Kovalerchuk and E. Vityaev, *Data Mining in Finance: Advances in Relational and Hybrid Methods*, Springer, 2000.

[5] M. Celik, S. Shekhar, J. P. Rogers, and J. A. Shine, Sustained Emerging Spatio-temporal Co-occurrence Pattern Mining: A Summary of Results, *In 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '06.),* pages 106-115, 2006.

[6] S. Shekhar, R. R. Vatsavai, and M. Celik, Spatial and Spatiotemporal Data Mining: Recent Advances, Data Mining: Next Generation Challenges and Future Directions, AAAI Press, 2008.

[7] R. B. Nelsen, *An Introduction to Copulas*, Springer, New York, 2006.

[8] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu and G. Ostrouchov, Relative Performance of Mutual Information Estimation Methods for Quantifying the Dependence among Short and Noisy Data, *Phys. Rev. E*, 76, 026209, 2007.

[9] W.-K. Wong, A. Moore, G. Cooper, and M. Wanger, Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks, *In Eighteenth national conference on Artificial intelligence*, pages 217-223, 2002.

[10] E. Eskin, Anomaly Detection over Noisy Data using Learned Probability Distributions, *In Proceedings of the International Conference on Machine Learning*, pages 255-262, 2000.

[11] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, 8, pages 229-231, 1959.

[12] S.-C. Kao and R. S. Govindaraju, Trivariate Statistical Analysis of Extreme Rainfall Events via Plackett Family of Copulas, *Water Resour. Res.*, 44, W02415, 2008.

[13] R. Maity and D. N. Kumar, Probabilistic Prediction of Hydroclimatic Variables with Nonparametric Quantification of Uncertainty, *J. Geophys. Res.*, 113, D14105, 2008.

[14] G. Kuhn, S. Khan, A. R. Ganguly and M. L. Branstetter, Geospatial-temporal Dependence among Weekly Precipitation Extremes with Applications to Observations and Climate Model Simulations in South America, *Adv. Water Resour.*, 30(12), pages 2401-2423, 2007.

[15] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, London, 1997.

[16] C. Genest, A.-C. Favre, J. Béliveau and C. Jacques, Metaelliptical Copulas and Their Use in Frequency Analysis of Multivariate Hydrological Data, *Water Resour. Res.*, 43, W09401, 2007.

[17] C. Genest, K. Ghoudi and L.-P. Rivest, A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions, *Biometrika*, 82(3), pages 543-552, 1995.

[18] L. Zhang and V. P. Singh, Bivariate Flood Frequency Analysis Using the Copula Method, *J. Hydrol. Eng.*, 11(2), pages 150-164, 2006.

[19] R. Saunders and P. Laud, The Multidimensional Kolmogorov Goodness-of-fit Test, *Biometrika*, 67(1), page 237, 1980.

[20] W. Breymann, A. Dias and P. Embrechts, Dependence Structures for Multivariate High-frequency Data in Finance, *Quant. Finance*, 3(1), pages 1-14, 2003.

[21] J.-D. Fermanian, Goodness-of-fit Tests for Copulas, *J. Multivariate Anal.*, 95, pages 119-152, 2005.

[22] C. Wallace and D. Clayton, Estimating Relative Recurrence Risk Ratio, *Genet. Epidemiol.*, 25(4), pages 293-302, 2003.

[23] Intergovernmental Panel on Climate Change, *Climate Change 2007: Fourth Assessment Report*, 2007.

[24] M. W. Berry and M. Browne, *Lecture Notes in Data Mining*, World Scientific Publishing, NJ, 2006.

[25] M. Kanamitsu, W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino and G. L. Potter, NCEP-DOE AMIP-II Reanalysis (R-2), *Bulletin of the American Meteorological Society*, 83(11), pages 1631-1643, 2002.

[26] P.-N. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster and A. Torregrosa, Finding Spatio-temporal Patterns in Earth Science Data, *In Proceedings of KDD Workshop on Temporal Data Mining*, 2001.