Trivariate Statistical Analysis of Extreme Rainfall Events via Plackett Family of Copulas

Shih-Chieh Kao
Purdue University
October 29, 2007

Outline

- Background and Motivation
- Research Objectives
- Introduction to Copulas
- Current Choices of Trivariate Copulas
- Plackett Family of Copulas
- Temporal Distribution of Design Rainfall
- Conclusions

Background and Motivation

- Many hydrologic variables are indexed in space and time, and are co-dependent.
  - The assumption of independence is not realistic
- Univariate stochastic approaches are not capable of addressing multivariate problems.
  - Infinite possibilities of joint distributions exist for fixed marginals
- The need to characterize dependence structure
  - Linear correlation coefficient is not a complete measure.
- Explore use of copulas as a solution
- Constructing higher order (>2) stochastic models is an unresolved problem

Research Objectives

- Given depth and duration, use conditional expectation to develop the temporal distribution for design rainfall
  - (1) Capture peak properties
  - (2) Develop temporal accumulation curves
- Construct a trivariate copula preserving bivariate dependencies for analyzing Indiana rainfall
- Explore the nuances of compatibility problem
  - Not any given set of bivariate dependencies has a valid trivariate copula
- Examine the use of Plackett family of copulas at the trivariate level

Basic Probabilistic Definition

- Univariate (for variable X)
  - Cumulative density function (CDF) and probability density function (PDF)

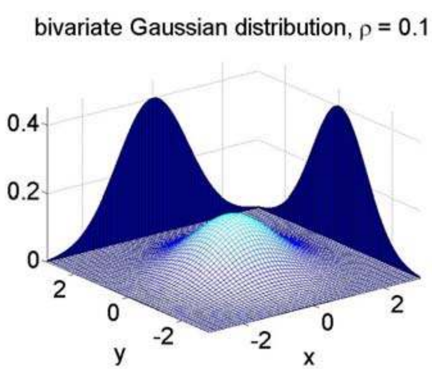
$$F_X(x) = P[X \le x] \qquad f_X(x) = \frac{\partial}{\partial x} F_X(x)$$

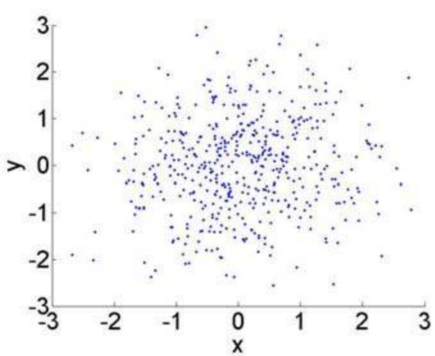
- Bivariate (for variables X and Y)
  - joint-CDF and joint-PDF

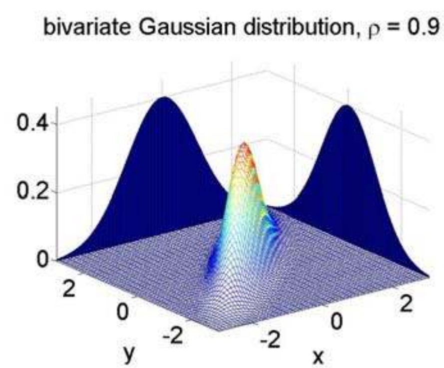
$$H_{XY}(x, y) = P[X \le x, Y \le y] \qquad h_{XY}(x, y) = \frac{\partial^2 H_{XY}(x, y)}{\partial x \partial y}$$

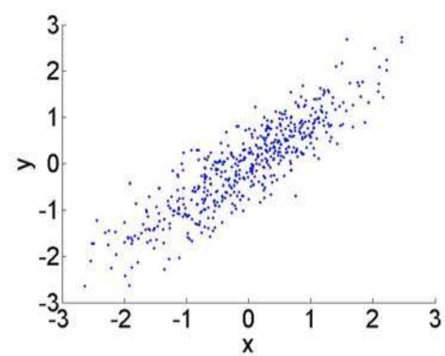
  - Marginal distributions

$$f_X(x) = \int_{-\infty}^{\infty} h_{XY}(x, y)\,dy \qquad f_Y(y) = \int_{-\infty}^{\infty} h_{XY}(x, y)\,dx$$







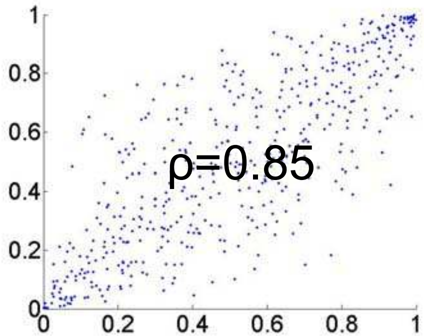


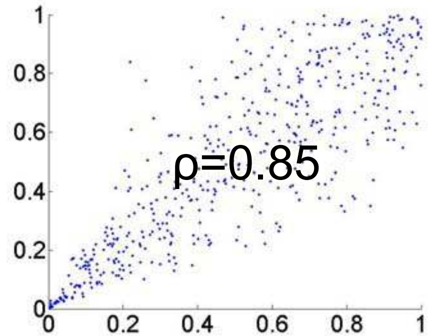
Concept of Dependence Structure

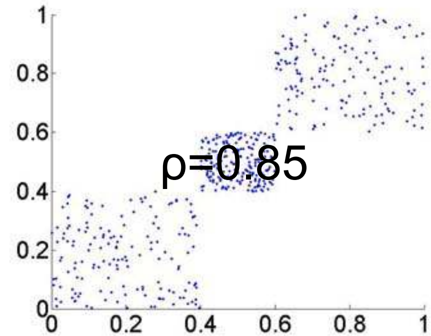
- Conventionally quantified by the linear correlation coefficient ρ

$$\rho_{XY} = \frac{E[(X-\bar{x})(Y-\bar{y})]}{Std[X]Std[Y]}$$

  - Can not correctly describe association between variables







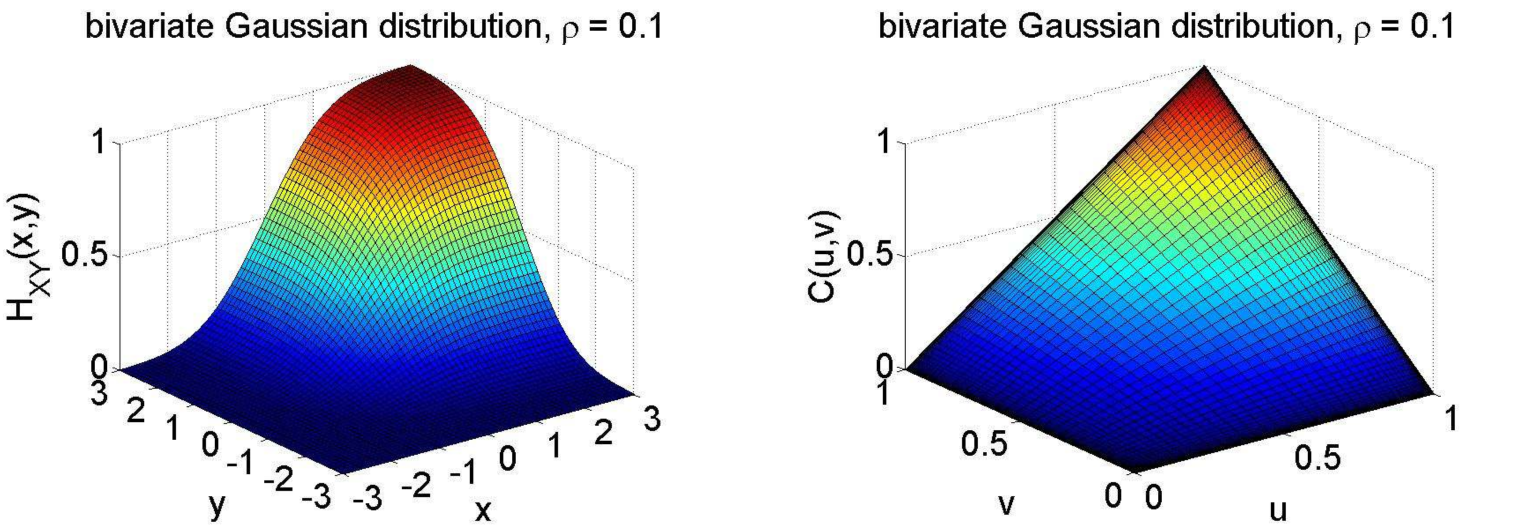
  - Only valid for Gaussian (or some elliptic) distributions
  - A better tool is required to characterize dependence => copulas

Introduction to Copulas

- A copula C(u,v) is a function comprised of margins u & v from [0,1]×[0,1] to [0,1].
  – Sklar (1959) showed that for continuous marginals u and v, there exists a unique copula C such that

$$H_{XY}(x,y) = C_{UV}(F_X(x), F_Y(y)) = C_{UV}(u,v)$$

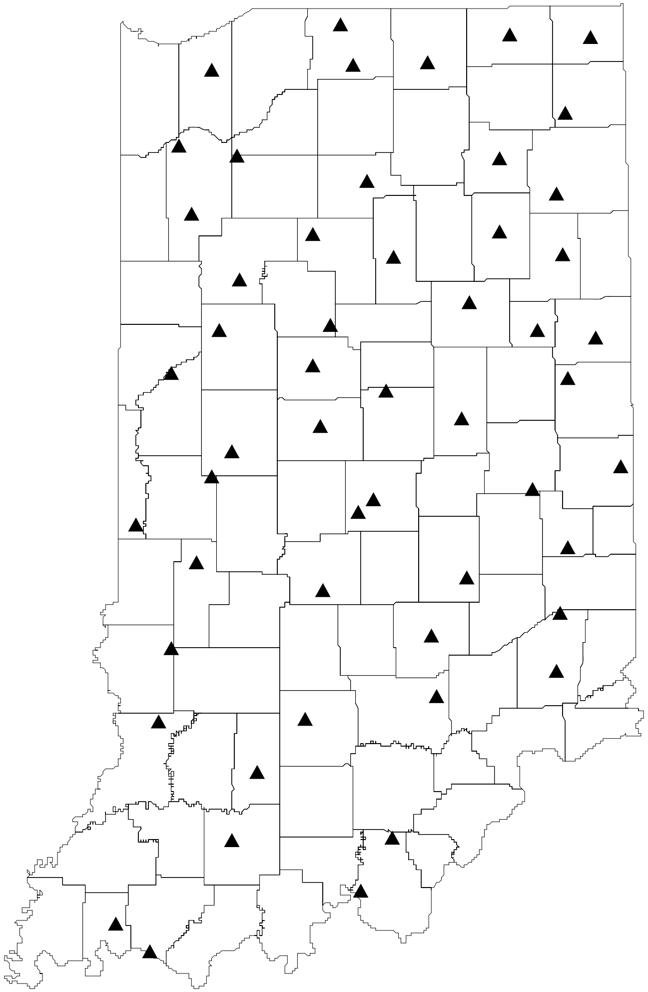
  – Transformation from $[-\infty,\infty]^2$ to $[0,1]^2$



  – Provides a complete description of dependence structure

Data Source & Study Area

- Nation Climate Data Center, Hourly Precipitation Dataset (NCDC, TD 3240 dataset)
- 53 Co-operative Rainfall Stations in Indiana with record length greater than 50 years
- Minimum rainfall hiatus: 6 hours
- About 4800 events per station
- Annual maximum cumulative probability (AMP) definition for selecting annual series



Difficulties in Constructing Higher-order Copulas

- Preserving mutual dependencies

$$\begin{cases} C_{UVW}(1,v,w) = C_{VW}(v,w) \\ C_{UVW}(u,1,w) = C_{UW}(u,w) \\ C_{UVW}(u,v,1) = C_{UV}(u,v) \end{cases}$$

  – Drawback of Archimedean copulas

$$\varphi_\theta\left(C_{UVW}(u,v,w)\right) = \varphi_\theta(u) + \varphi_\theta(v) + \varphi_\theta(w)$$

    - Only one bivariate dependence can be preserved

- Compatibility problem
  – Q1: Is it possible to have all perfect positive dependencies at the bivariate level? (i.e. $\rho_{XY} = 1$, $\rho_{YZ} = 1$, and $\rho_{XZ} = 1$)
  – Q2: Is it possible to have all perfect negative dependencies at the bivariate level? (i.e. $\rho_{XY} = -1$, $\rho_{YZ} = -1$, and $\rho_{XZ} = -1$)
  – Not any set of given bivariate dependencies has valid copulas

Current Choices of Trivariate Copulas

- Archimedean Copulas
  - Grimaldi and Serinaldi 2006a; Zhang and Singh, 2007b, 2007c
- Fully-nested copulas
  - Grimaldi and Serinaldi, 2006b, 2007

$$\begin{cases} \varphi_1\left(C_{UVW}\left(w, C_{UV}\left(u,v\right)\right)\right) = \varphi_1\left(w\right) + \varphi_1\left(C_{UV}\left(u,v\right)\right) \\ \varphi_2\left(C_{UV}\left(u,v\right)\right) = \varphi_2\left(u\right) + \varphi_2\left(v\right) \end{cases}$$

  - Not all bivariate dependencies can be preserved
- Salvadori and De Michele (2006)
  - Special case of "conditional copulas" (Chakak and Koehler, 1995)
  - Sequence of variables is not interchangeable
- Meta-elliptical copulas
  - Genest *et al*., 2007; Renard and Lang, 2007
  - Extension of multivariate Gaussian distribution
  - Lack of parameter on the trivariate level

Constant Cross Product Ratio Theory

- Constant cross product ratio theory (2-Plackett copulas)
  - For any given point (u,v) in $[0,1]^2$

$$\psi_{UV} = \frac{P[U \le u, V \le v]P[U > u, V > v]}{P[U > u, V \le v]P[U \le u, V > v]}$$

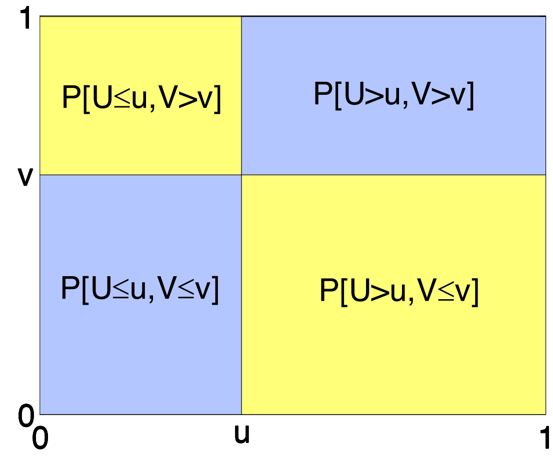
  - In terms of copulas $C_{UV}(u,v)$

$$C_{UV}(u,v) = \frac{[1+(\psi_{UV}-1)(u+v)] - \sqrt{[1+(\psi_{UV}-1)(u+v)]^2 - 4uv\,\psi_{UV}(\psi_{UV}-1)}}{2(\psi_{UV}-1)}$$

  - $\Psi = 1$, independent

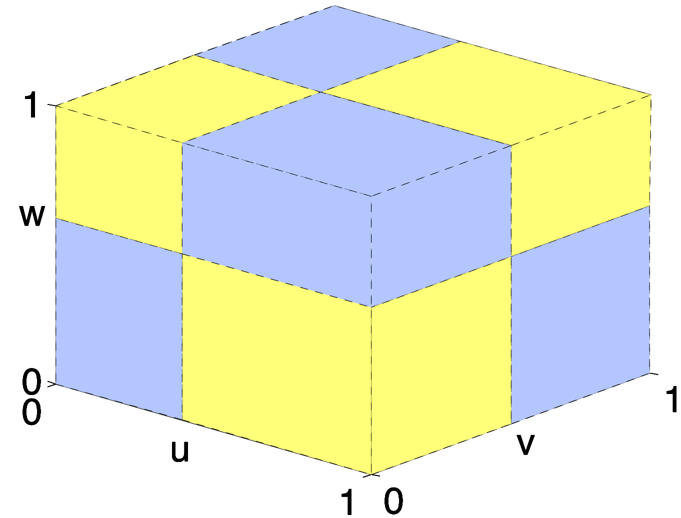
    $\Psi > 1$, positive dependent ($\Psi \to \infty$, totally positive)

    $\Psi < 1$, negative dependent ($\Psi \to 0$, totally negative)
  - Parameter estimation
    - Maximum likelihood
    - Median approach – $n_{00}n_{11}/n_{01}n_{10}$



Trivariate Plackett Family of Copulas (I)

- 3-Plackett
  - In $[0, 1]^3$ $\quad \psi_{UVW} = \dfrac{P_{000}P_{011}P_{101}P_{110}}{P_{111}P_{100}P_{010}P_{001}}$



  - Solve the 4-th order polynomial

$$\psi_{UVW}(a_1 - z)(a_2 - z)(a_3 - z)(a_4 - z) - z(z - b_1)(z - b_2)(z - b_3) = 0$$

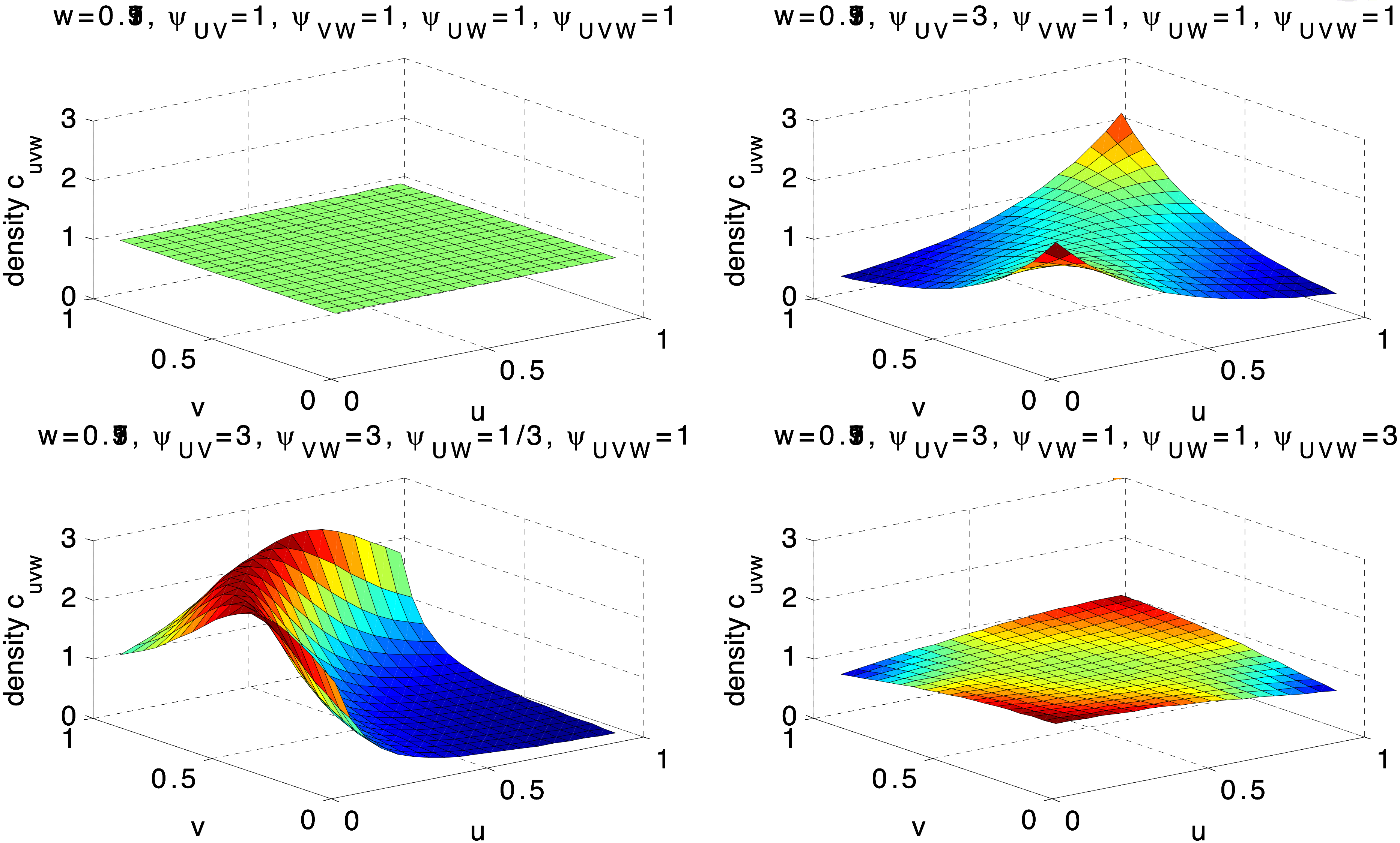
$$\begin{cases} a_1 = C_{VW}(v,w), \quad a_2 = C_{UW}(u,w), \quad a_3 = C_{UV}(u,v) \\ a_4 = 1 - u - v - w + C_{UV}(u,v) + C_{VW}(v,w) + C_{UW}(u,w) \\ b_1 = C_{UW}(u,w) + C_{VW}(v,w) - w \\ b_2 = C_{UV}(u,v) + C_{VW}(v,w) - v \\ b_3 = C_{UW}(u,w) + C_{UV}(u,v) - u \end{cases}$$

  - When compatible, only one solution in (Fréchet bounds)

$$\max(0, b_1, b_2, b_3) \le z \le \min(a_1, a_2, a_3, a_4)$$

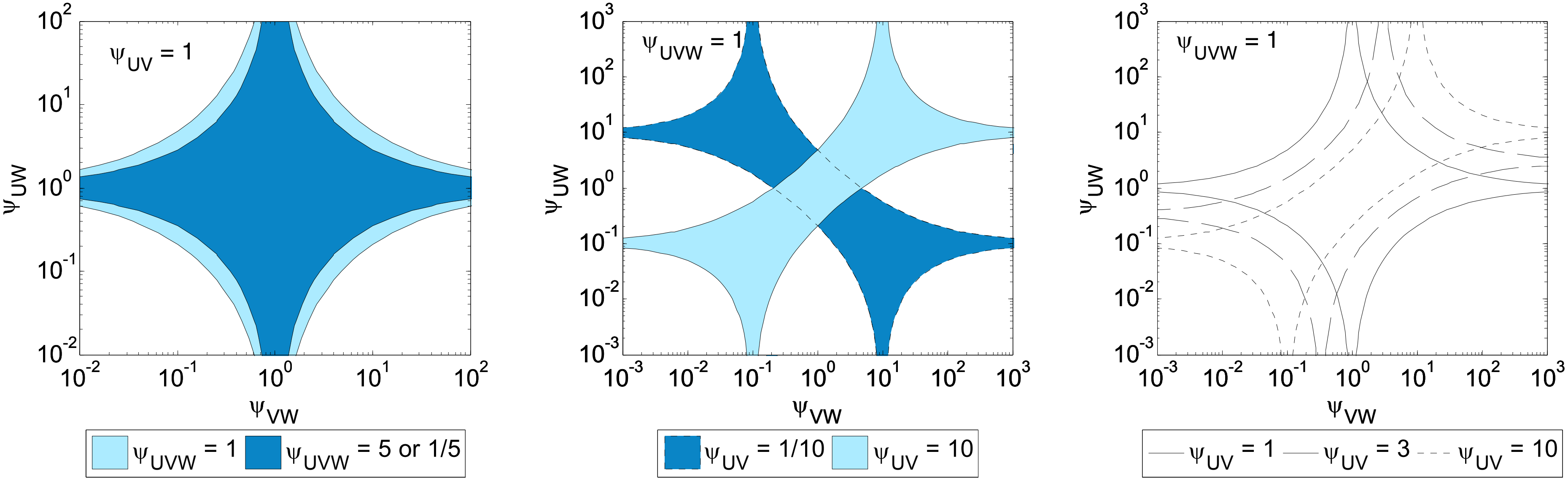
  - Implicit procedure for computing copula density
  - Parameter estimation
    - Maximum likelihood
    - Median approach – $n_{000}n_{011}n_{101}n_{110}/n_{111}n_{100}n_{010}n_{001}$

Trivariate Plackett Family of Copulas (II)



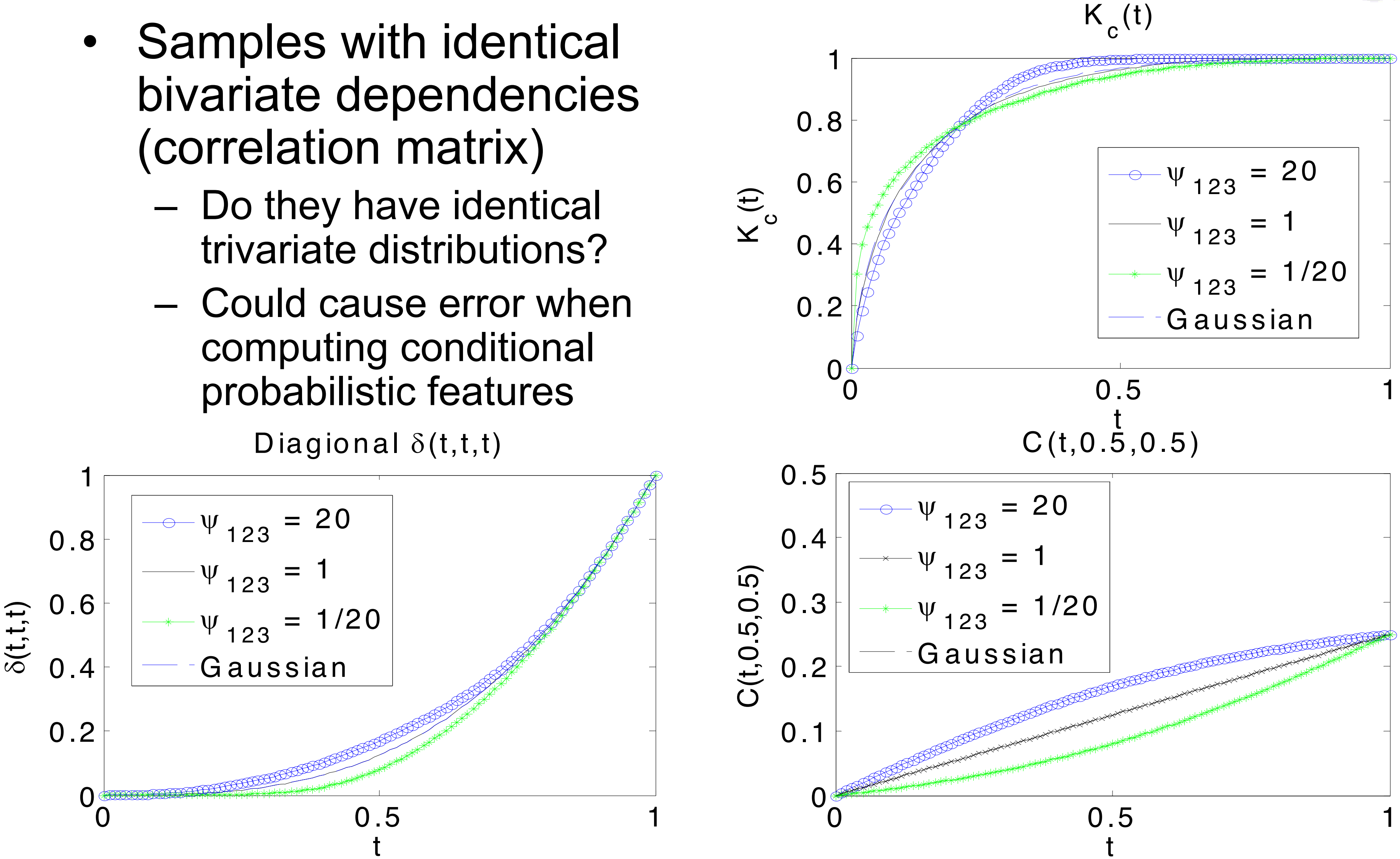
Feasible Region for Valid 3-Plackett

- Problem: bivariate distributions may not be compatible, i.e. probability measure V_C in $[a_1, b_1, c_1]$ and $[a_2, b_2, c_2]$ might be negative! – Open Question.
- Feasible region of Plackett parameters
  - Adopt numerical approach to find when copula density is greater or equal to zero in $[0,1]^3$



Comparison between Plackett and Gaussian

- Samples with identical bivariate dependencies (correlation matrix)
  - Do they have identical trivariate distributions?
  - Could cause error when computing conditional probabilistic features



Temporal Distribution of Design Rainfall

- Given depth (P) and duration (D), what is the corresponding temporal distribution of design rainfall?
  - Conditional expectation
  - Two applications
    - Capture the peak features
    - Develop the temporal accumulation curves

- Selected variables for analysis:
  - Depth (volume), P (mm)
  - Duration, D (hour)
  - Peak Intensity, I (mm/hour)
  - Percentage Time to Peak, T_p (%)
  - Percentage cumulative accumulation at each 10% temporal ordinates, A_{10}, A_{20}, …, A_{90} (%)

- 50 out of 53 stations are valid for 3-Plackett

Marginal Distributions

- Parameter estimation
  - Maximum likelihood (ML) & method of moments (MOM)
- Goodness-of-fit
  - Chi-square and Kolmogorov-Smirnov (KS) test
- Selection of marginal distribution:
  - Akaike Information Criterion (AIC) & Bayesian Information Criterion (BIC)

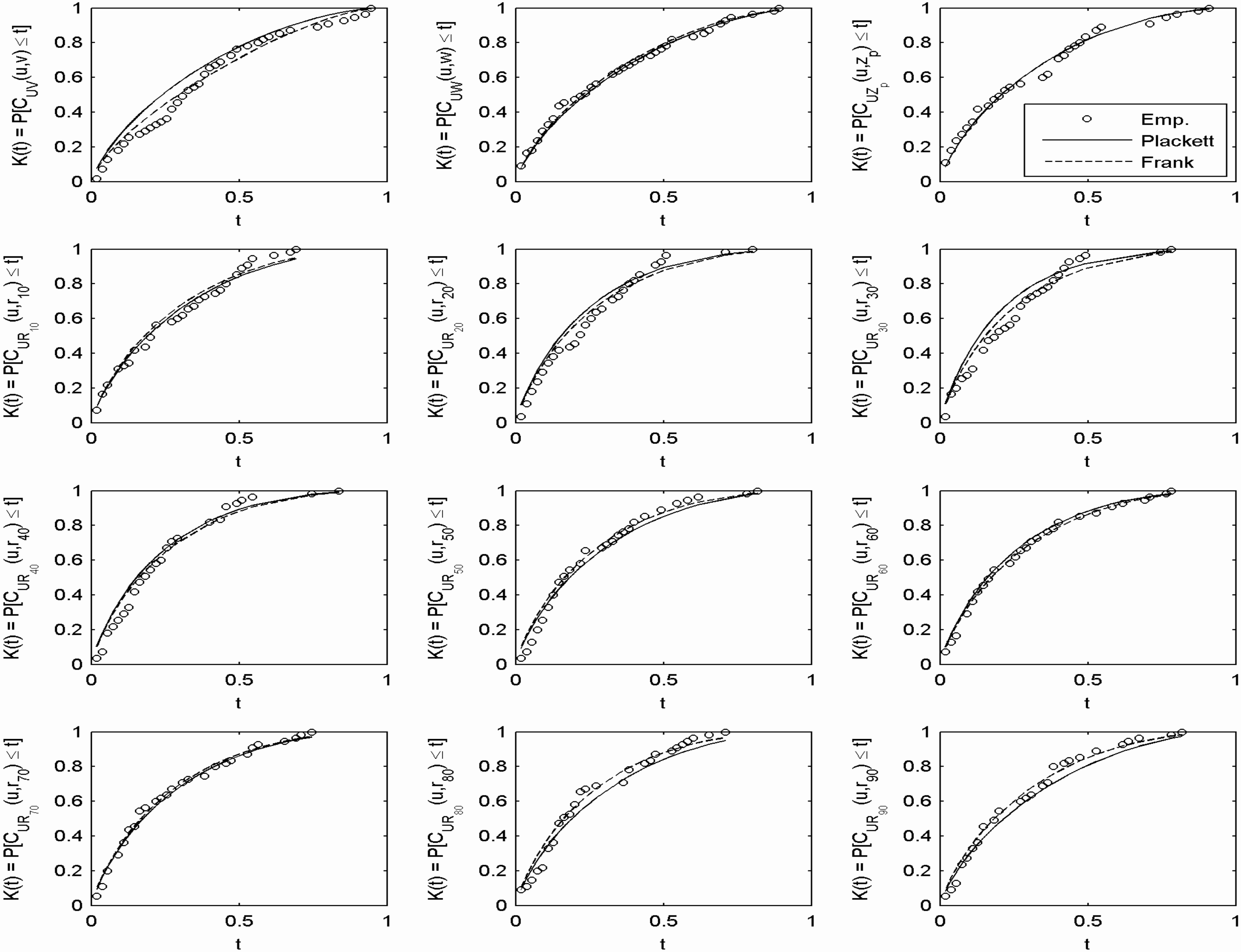
	Number of stations with the minimum AIC				Number of stations with the minimum BIC			
	GEV	LN	P3	LP3	GEV	LN	P3	LP3
Depth, P	7	38	1	7	4	47	0	2
Duration, D	1	46	6	0	1	50	2	0
Intensity, I	9	41	2	1	2	49	1	1

  - P, D, and I: Log-normal (LN) distribution
  - T_p and A_k: Beta (β) distribution

Bivariate Dependence Structure

	Kendall's τ		Frank's θ esimated by				Plackett's ψ estimated by			
			ML		Kendall's τ		ML		Median	
	mean	stdev	mean	stdev	mean	stdev	mean	stdev	mean	stdev
P vs. D	0.336	0.077	3.554	1.031	3.410	0.975	5.140	2.335	4.468	2.394
P vs. I	0.246	0.097	2.410	1.055	2.389	1.029	3.270	1.449	2.926	1.474
P vs. T_p	0.047	0.100	0.392	0.959	0.429	0.926	1.323	0.551	1.518	0.838
P vs. A_k	-0.053	0.103	-0.503	0.966	-0.494	0.947	0.883	0.469	0.917	0.581
D vs. I	-0.196	0.093	-1.971	0.962	-1.863	0.927	0.439	0.198	0.554	0.341
D vs. T_p	0.030	0.085	0.246	0.868	0.272	0.776	1.214	0.477	1.421	0.608
D vs. A_k	-0.076	0.093	-0.726	0.910	-0.710	0.866	0.783	0.364	0.746	0.395

- Parameter estimation
  - Maximum likelihood (ML)
  - Non-parametric approach using Kendall's tau τ
  - Median approach
- Goodness-of-fit
  - Multidimensional KS test
  - Rosenblatt's transformation test (RTT)
    - $Z_1 = \Phi^{-1}(P[U \le u])$, $Z_2 = \Phi^{-1}(P[V \le v | U = u])$
    - Test if $S = Z_1^2 + Z_2^2$ follows chi-square distribution (λ^2_2)



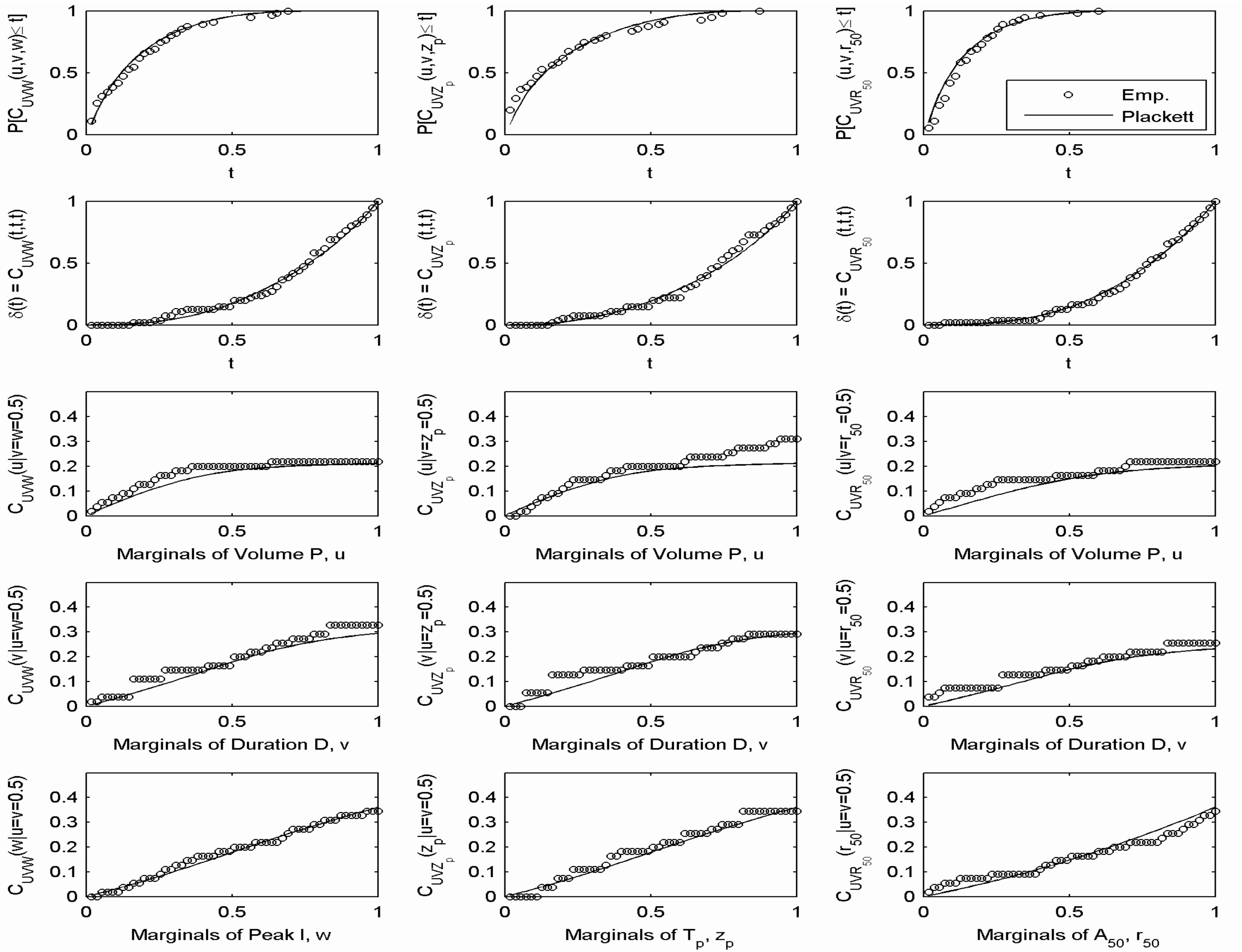
Emp.
Plackett
Frank
K(t) = P[C_UV(u,v) ≤ t]
K(t) = P[C_UW(u,w) ≤ t]
K(t) = P[C_UZp(u,z_p) ≤ t]
K(t) = P[C_UR10(u,r_10) ≤ t]
K(t) = P[C_UR20(u,r_20) ≤ t]
K(t) = P[C_UR30(u,r_30) ≤ t]
K(t) = P[C_UR40(u,r_40) ≤ t]
K(t) = P[C_UR50(u,r_50) ≤ t]
K(t) = P[C_UR60(u,r_60) ≤ t]
K(t) = P[C_UR70(u,r_70) ≤ t]
K(t) = P[C_UR80(u,r_80) ≤ t]
K(t) = P[C_UR90(u,r_90) ≤ t]
t

Trivariate Dependence Structure

Variables	Plackett's ψ mean	Plackett's ψ stdev	Variables	Plackett's ψ mean	Plackett's ψ stdev
P vs. D vs. I	1.163	0.531	P vs. D vs. T_p	1.438	1.088
P vs. D vs. A_{10}	1.162	0.722	P vs. D vs. A_{60}	1.392	1.174
P vs. D vs. A_{20}	1.149	0.623	P vs. D vs. A_{70}	1.403	1.200
P vs. D vs. A_{30}	1.432	1.271	P vs. D vs. A_{80}	1.228	0.765
P vs. D vs. A_{40}	1.379	1.180	P vs. D vs. A_{90}	1.228	0.880
P vs. D vs. A_{50}	1.458	1.542			

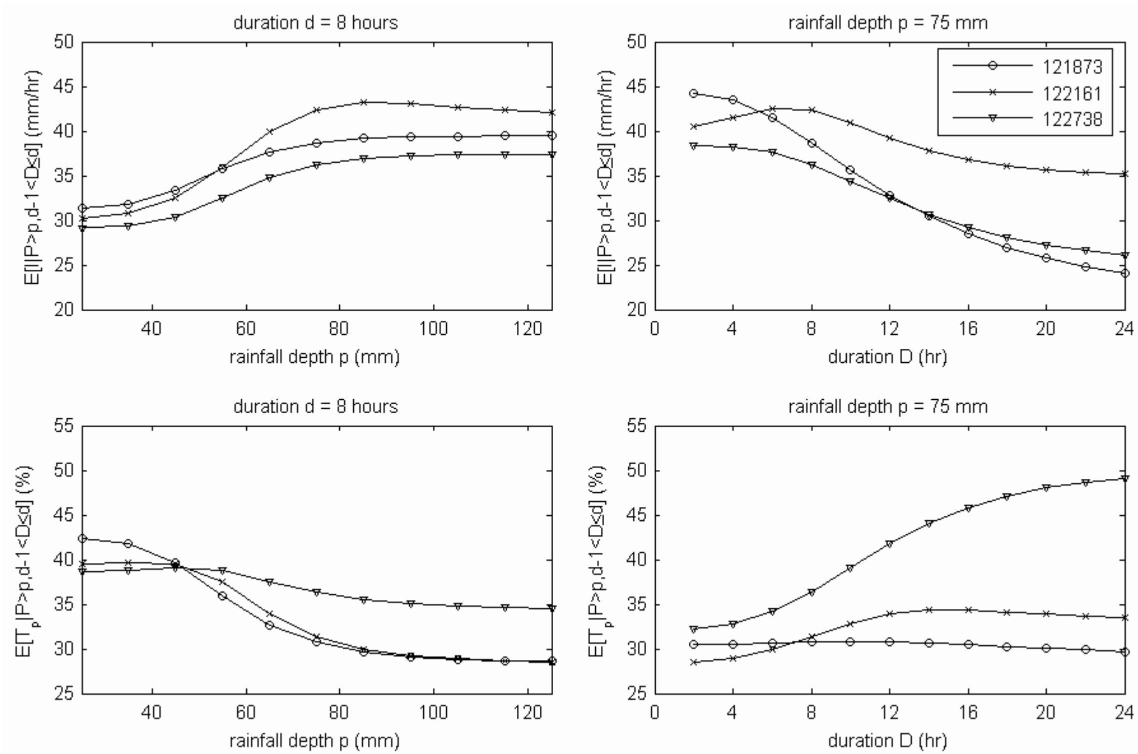
- Parameter estimation
  - Only by maximum likelihood (ML)
  - Sample size is not sufficient for median approach. Cases with zero observation may exist
- RTT test
  - $Z_1 = \Phi^{-1}(P[U \le u])$, $Z_2 = \Phi^{-1}(P[V \le v | U=u])$, $Z_3 = \Phi^{-1}(P[W \le w | U=u, V=v])$
  - Test if $S = Z_1^2 + Z_2^2 + Z_3^2$ follows chi-square distribution (λ^2_3)

	Number of stations invalid for 3-Plackett		Number of stations rejected by RTT at 5% significance level	
	median approach	maximum likelihood	median approach	maximum likelihood
C_{UVW} (P vs. D vs. I)	3/53	22/53	4/50	4/31
C_{UVR90} (P vs. D vs. A_{90})	0/53	1/53	0/53	0/52


$P[C_{UVW}(u,v,w) \le t]$
$P[C_{UVZ_p}(u,v,z_p) \le t]$
$P[C_{UVR_{50}}(u,v,r_{50}) \le t]$
Emp.
Plackett
t
$\delta(t) = C_{UVW}(t,t,t)$
$\delta(t) = C_{UVZ_p}(t,t,t)$
$\delta(t) = C_{UVR_{50}}(t,t,t)$
$C_{UVW}(u|v=w=0.5)$
$C_{UVZ_p}(u|v=z_p=0.5)$
$C_{UVR_{50}}(u|v=r_{50}=0.5)$
Marginals of Volume P, u
$C_{UVW}(v|u=w=0.5)$
$C_{UVZ_p}(v|u=z_p=0.5)$
$C_{UVR_{50}}(v|u=r_{50}=0.5)$
Marginals of Duration D, v
$C_{UVW}(w|u=v=0.5)$
$C_{UVZ_p}(z_p|u=v=0.5)$
$C_{UVR_{50}}(r_{50}|u=v=0.5)$
Marginals of Peak I, w
Marginals of T_p, z_p
Marginals of A_{50}, r_{50}

Rainfall Peak Attributes

- Given depth (P) and duration (D), compute the conditional expectation of peak intensity (I) and percentage time to peak (T_p)
  - Peak intensity increases with total depth, decreases with duration
  - Time to peak increases with duration, decreases with total depth

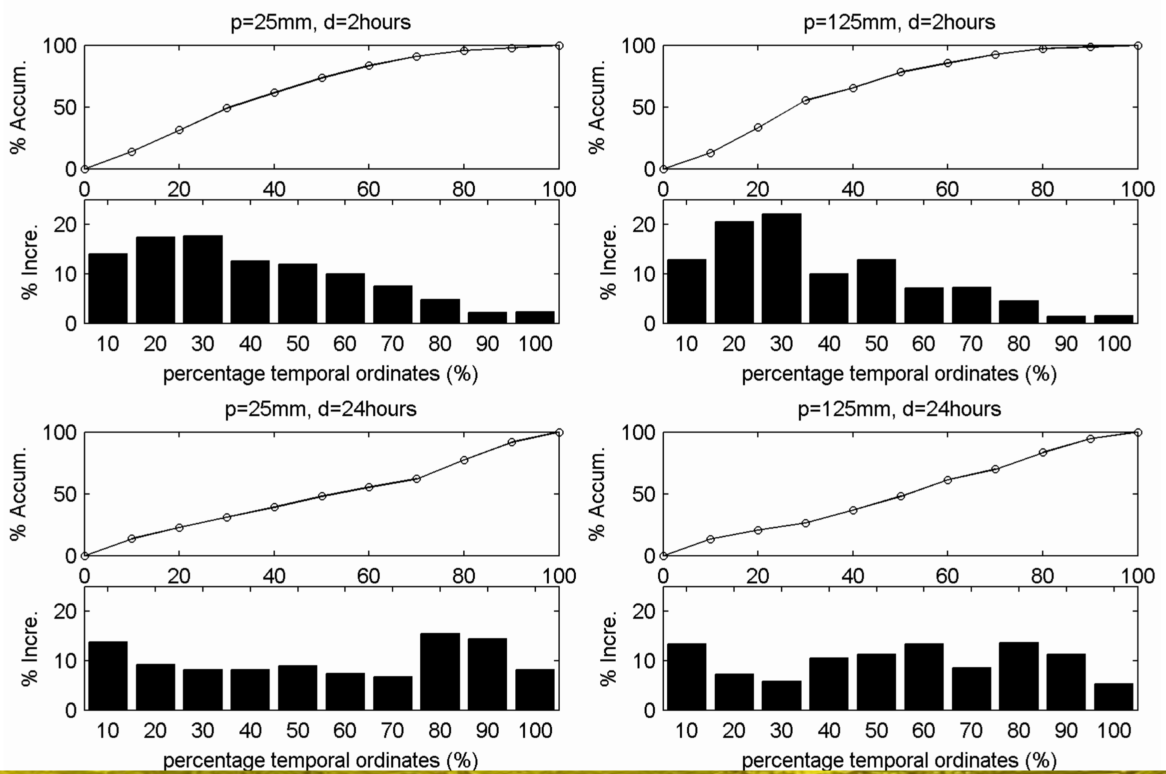


$$E\left[I \mid P > p_D, d_D - 1 < D < d_D\right]$$

$$E\left[T_p \mid P > p_D, d_D - 1 < D < d_D\right]$$

Temporal Accumulation Curves

- Given depth (P) and duration (D), compute the conditional expectation of percentage accumulations at each 10% temporal ordinates (A_{10}, A_{20}, …, A_{90})
- Results are sensitive to the quality of data



$$E\left[R_k \mid P > p_D, d_D - 1 < D < d_D\right]$$

Conclusions (I)

- Plackett family of copulas, along with the underlying cross product ratio theory, was found to be a suitable trivariate dependence model in constructing rainfall temporal distribution.

- The feasibility region for Plackett parameters that would result in valid 3-copulas has been identified numerically in this study.

- Not every sets of given bivariate dependencies have a corresponding valid 3-copula (even for Gaussian copulas). The compatibility of given bivariate dependencies needs to be investigated.

Conclusions (II)

- Marginal distributions
  - Log-normal distribution is found suitable for depth, duration, and peak intensity
  - Beta (β) distribution is found suitable for percentage time to peak and percentage cumulative accumulation at each 10% temporal ordinates

- Dependence Structure
  - Plackett family is found to be a suitable dependence model both on the bivariate and trivariate levels.

- When given depth and duration, it can be observed that peak intensity (I) increases with depth, decreases with duration, while time to peak (T_p) increases with duration, decreases with depth.

- The analytical proof for trivariate Plackett family remains an "*Open Question*".

Questions?