Bivariate Frequency Analysis of Extreme Rainfall Events via Copulas

Shih-Chieh Kao
Purdue University
February 2007

Outline

- Background and motivation
- Brief introduction to copulas
- Previous work
- Selection of extreme events
- Analysis of marginal distributions
- Analysis of dependence structure
- Model applications
- Conclusions

Background and Motivation

- Extreme rainfall behavior
  - Basis for hydrologic design
  - Conventionally analyzed only by “depth”
  - Pre-specified artificial duration (filter), *not* the real duration of extreme rainfall event
  - Hard to represent other rainfall characteristics, e.g. peak intensity
- Definition of extreme event in multi-variate sense is not clear
- Dependence exists between rainfall characteristics (e.g. volume(depth), duration, peak intensity)
- Explore the use of copulas

Basic Probability Definitions

- 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

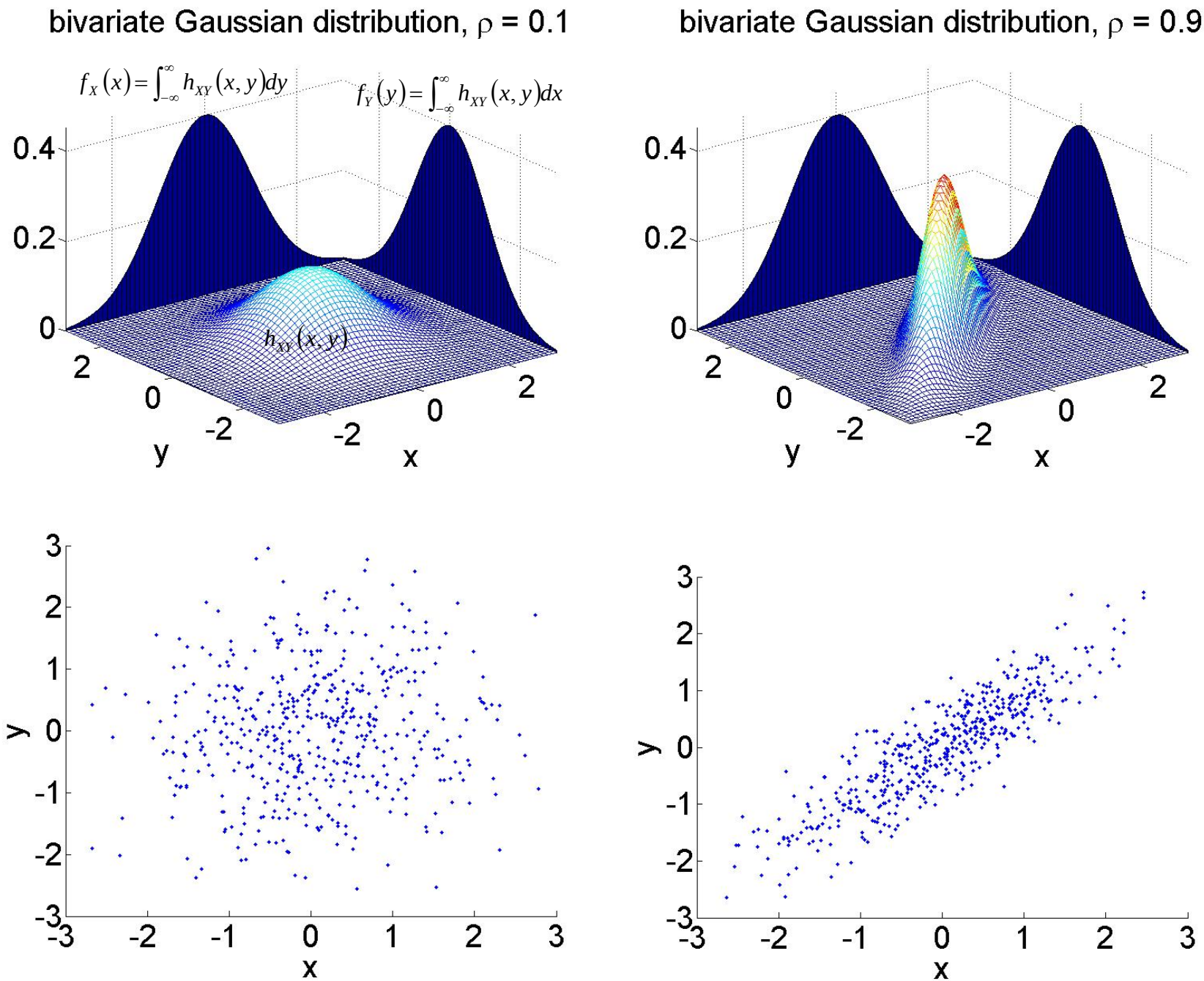
$$H_{XY}(x,y) = P[X \le x, Y \le y] \qquad h_{XY}(x,y) = \frac{\partial^2 H_{XY}(x,y)}{\partial x \partial y}$$

  - Marginal distirbutions

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

  - Marginals (univariate CDF)

$$u = F_X(x) = \int_{-\infty}^{x} f_X(x)dx \qquad v = F_Y(y) = \int_{-\infty}^{y} f_Y(y)dy \qquad 0 \le u, v \le 1$$

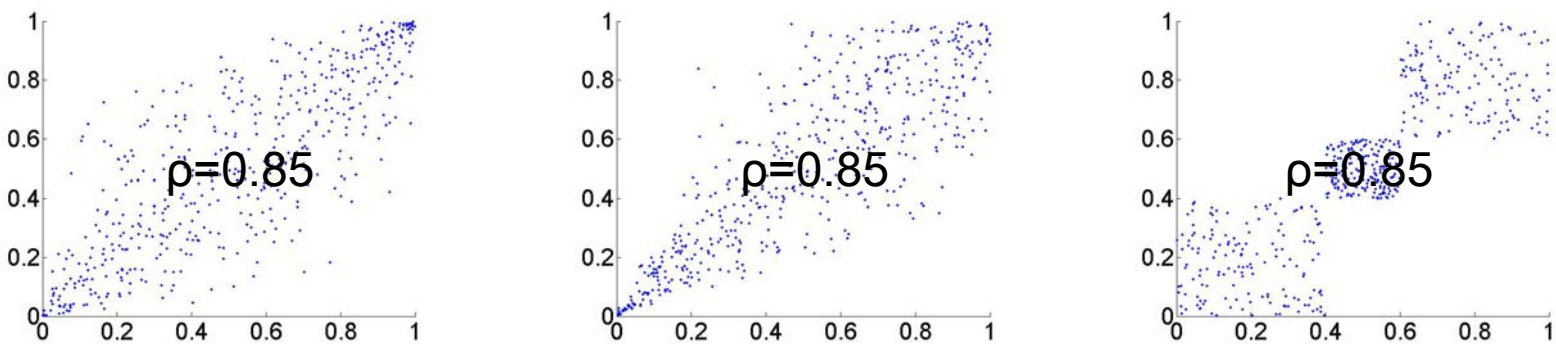

bivariate Gaussian distribution, ρ = 0.1
$f_X(x) = \int_{-\infty}^{\infty} h_{XY}(x,y)dy$
$f_Y(y) = \int_{-\infty}^{\infty} h_{XY}(x,y)dx$
$h_{XY}(x,y)$
0.4
0.2
0
2
0
-2
y
-2
0
2
x
bivariate Gaussian distribution, ρ = 0.9
0.4
0.2
0
2
0
-2
y
-2
0
2
x
3
2
1
0
-1
-2
-3
y
-3
-2
-1
0
1
2
3
x
3
2
1
0
-1
-2
-3
y
-3
-2
-1
0
1
2
3
x

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



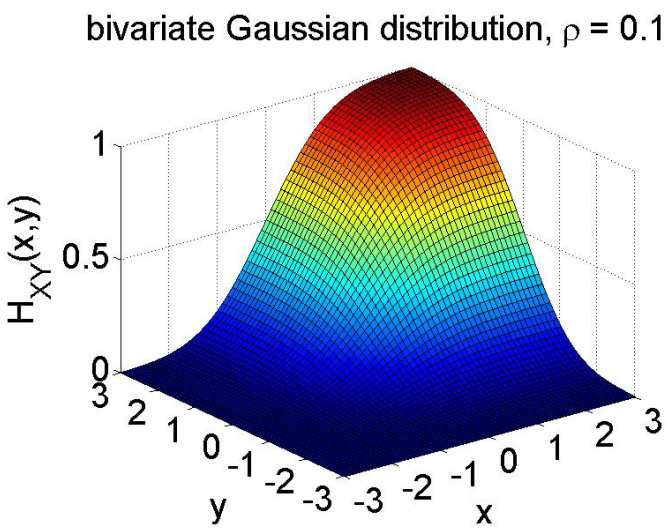
  – Only valid for Gaussian (or some elliptic) distributions
  – A better tool is required

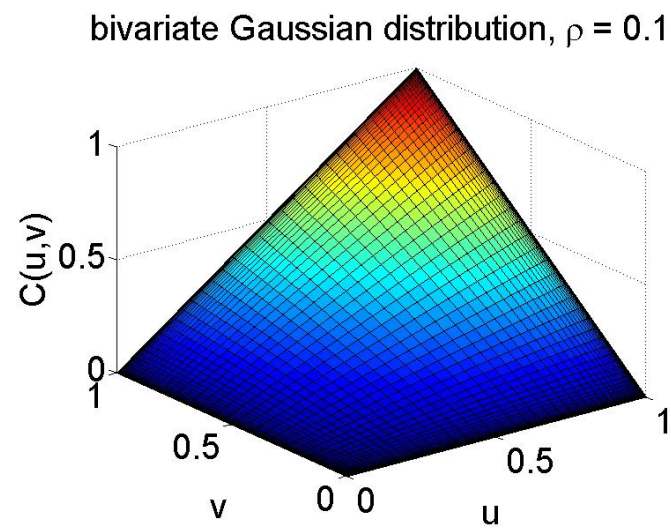
Introduction to Copulas

- 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

Archimedean Copulas (I)

- Archimedean Copulas
  - There exists a generator φ(t), such that

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

  - When φ(t) = -ln(t), C(u,v) = uv. (Independent case)
  - Commonly used 1-parameter Archimedean Copulas:
    - Frank family

$$\varphi(t) = -\ln\left[\left(e^{-\theta t} - 1\right)/\left(e^{-\theta} - 1\right)\right] \qquad C(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)}{e^{-\theta} - 1}\right)$$

    - Clayton family

$$\varphi(t) = \left(t^{-\theta} - 1\right)/\theta \qquad C(u,v) = \max\left(\left[u^{-\theta} + v^{-\theta} - 1\right]^{-\frac{1}{\theta}}, 0\right)$$

    - Genest-Ghoudi family

$$\varphi(t) = \left(1 - t^{1/\theta}\right)^{\theta} \qquad C(u,v) = \left\{\max\left(1 - \left[\left(1 - u^{1/\theta}\right)^{\theta} + \left(1 - v^{1/\theta}\right)^{\theta}\right], 0\right)\right\}^{\theta}$$

    - Ali-Mikhail-Haq family

$$\varphi(t) = \ln\{[1 - \theta(1 - t)]/t\} \qquad C(u,v) = \frac{uv}{1 - \theta(1 - u)(1 - v)}$$

Archimedean Copulas (II)

- Distribution function of copulas $K_C(t)=P[C_{UV}(u,v)\leq t]$
  - Offers cumulative probability measure for $\{(u,v)\in[0,1]^2 \mid C(u,v)\leq t\}$

$$K_C(t)=t-\frac{\varphi_\theta(t)}{\varphi_\theta'(t)}$$

- Concordance measure - Kendall's tau τ

$$\tau_{X,Y}=P[(X_1-X_2)(Y_1-Y_2)>0]-P[(X_1-X_2)(Y_1-Y_2)<0]$$

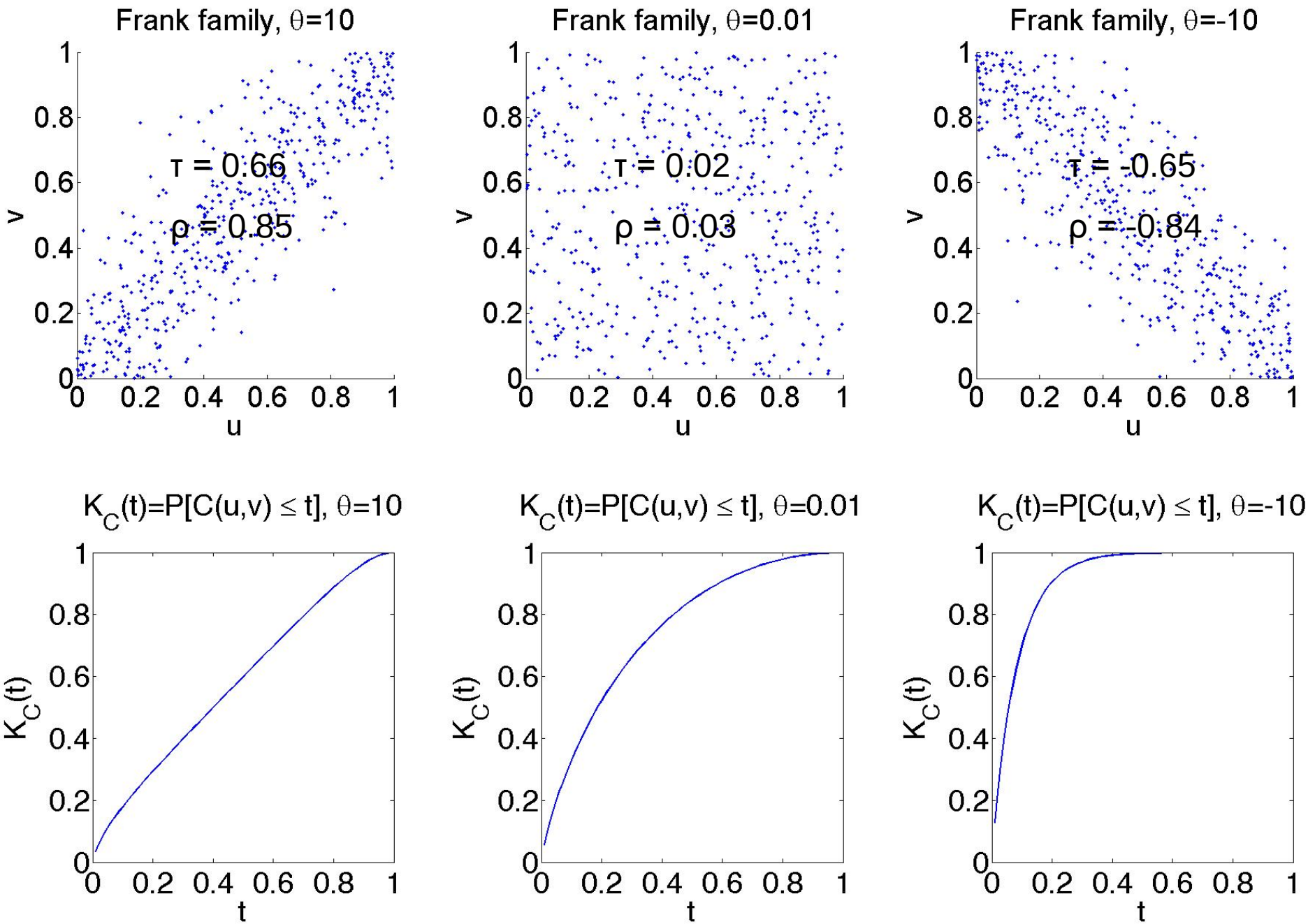
  - 1: total concordance, -1: total discordance, 0: zero concordance
  - Sample estimator (c: concordant pairs, d: discordant pairs, n: number of samples)

$$\hat{\tau}=(c-d)\Big/\binom{n}{2}$$

  - For Archimedean copulas

$$\tau=1+4\int_0^1\frac{\varphi_\theta(t)}{\varphi_\theta'(t)}dt$$

  - Non-parametric estimation of dependence parameter θ


Frank family, θ=10
τ = 0.66
ρ = 0.85
Frank family, θ=0.01
τ = 0.02
ρ = 0.03
Frank family, θ=-10
τ = -0.65
ρ = -0.84
u
v
K_C(t)=P[C(u,v) ≤ t], θ=10
K_C(t)=P[C(u,v) ≤ t], θ=0.01
K_C(t)=P[C(u,v) ≤ t], θ=-10
K_C(t)
t

Empirical Copulas

- Empirical copulas C_n

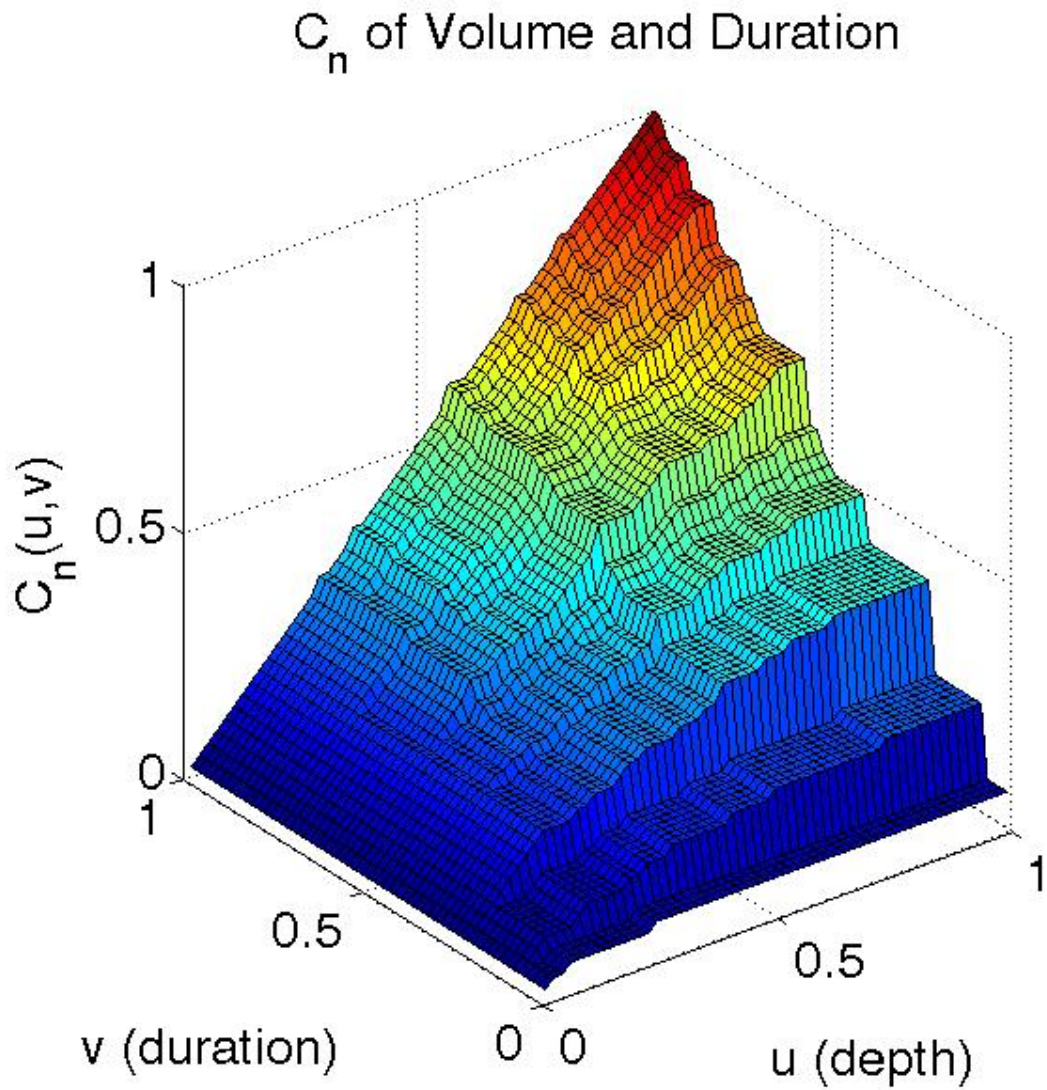
$$C_n\left(\frac{i}{n},\frac{j}{n}\right)=\frac{a}{n}$$

  – a: number of pairs (x,y) in the smaple with $x \le x_{(i)}$, $y \le y_{(i)}$

- Empirical distribution function K_{C_n}

$$K_{C_n}\left(\frac{k}{n}\right)=\frac{b}{n}$$

  – b: number of pairs (x,y) in the sample with $C_n(i/n,j/n) \le k/n$



Applications of Copulas in Hydrology

- Flood Frequency Analysis
  - Favre *et al*. (2004): Assessment of combined risk
  - De Michele *et al*. (2005): Dam spillway adequacy assessment
  - Grimaldi and Serinaldi (2006): The use of asymmetric copula in multi-variate flood frequency analysis
  - Zhang and Singh (2006): Conditional return period

- Return period assessment using bivariate model
  - Salvadori and De Michele (2004): Concept of secondary return period using distribution function K_C

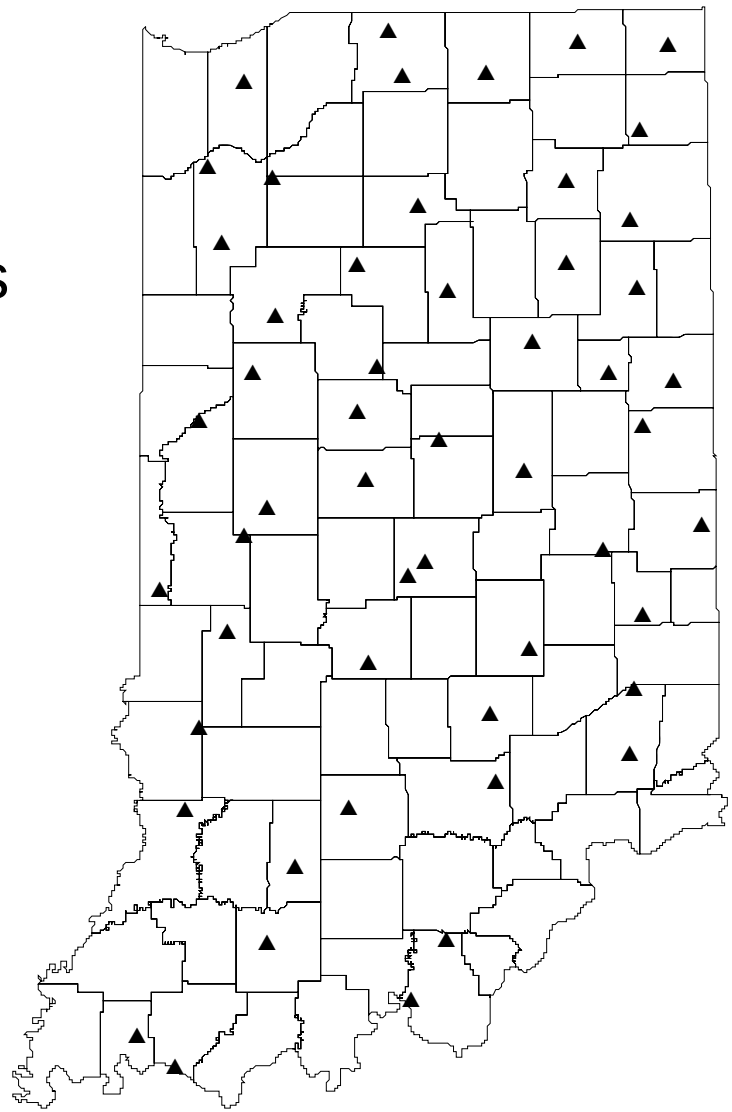
- Probablistic structure of storm surface runoff
  - Kao and Govindaraju (2007): Quantifying the effect of dependence between rainfall duration and average intensity on surface runoff

Copulas in Rainfall Frequency Analysis

- Rainfall frequency analysis
  - De Michele and Salvadori (2003, 2006)
    - Stochastic models for regular rainfall events
    - 2 rainfall stations in Italy with 7 years data
  - Grimaldi and Serinaldi (2006)
    - Extreme rainfall analysis
    - Relationship between design rainfall depth and the actual features of rainfall events
    - 10 rainfall stations in Italy with 7 years data
  - Zhang and Singh (2006)
    - Bivariate extreme rainfall frequency analysis using depth, duration and average intensity
    - 3 rainfall stations in Louisiana with 42 years data
- Unanswered questions
  - Data used for analysis may not be sufficient
  - Definition of “extreme events” in multi-variate sense?
  - Can results be applied for a large region?

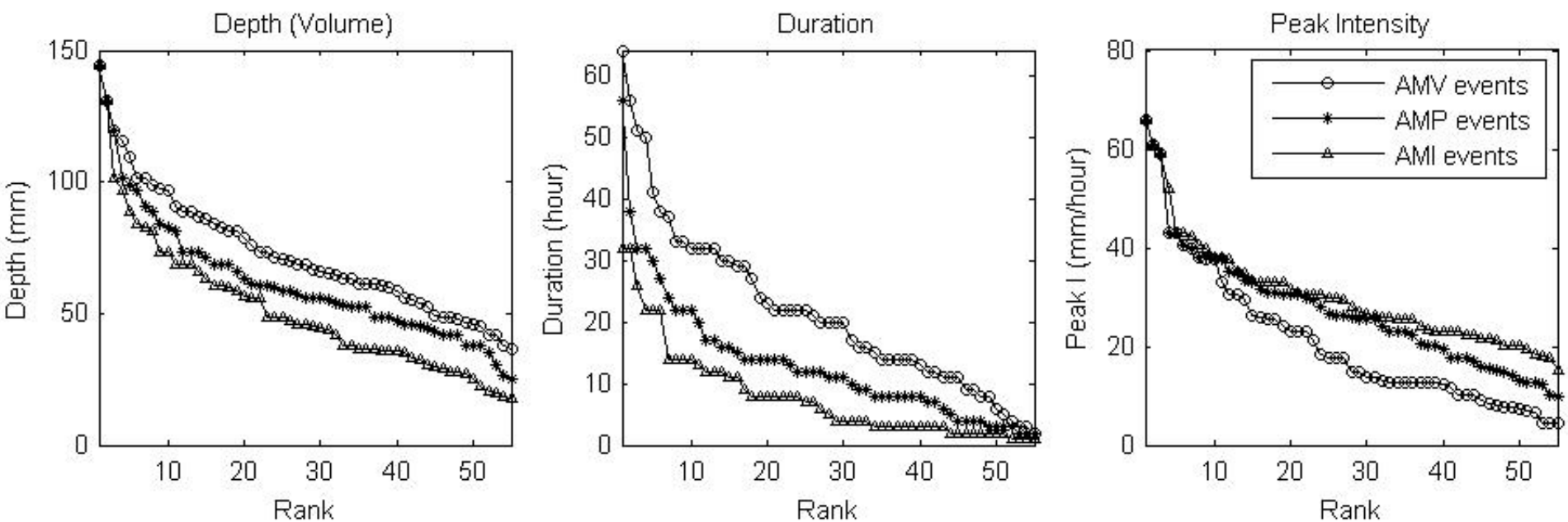
Data Source & Study Area

- 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
- Selected variables for analysis:
  - Depth (volume), P (mm)
  - Duration, D (hour)
  - Peak Intensity, I (mm/hour)
- Marginals:
  - $u=F_P(p)$, $v=F_D(d)$, $w=F_I(i)$



Definitions of Extreme Events

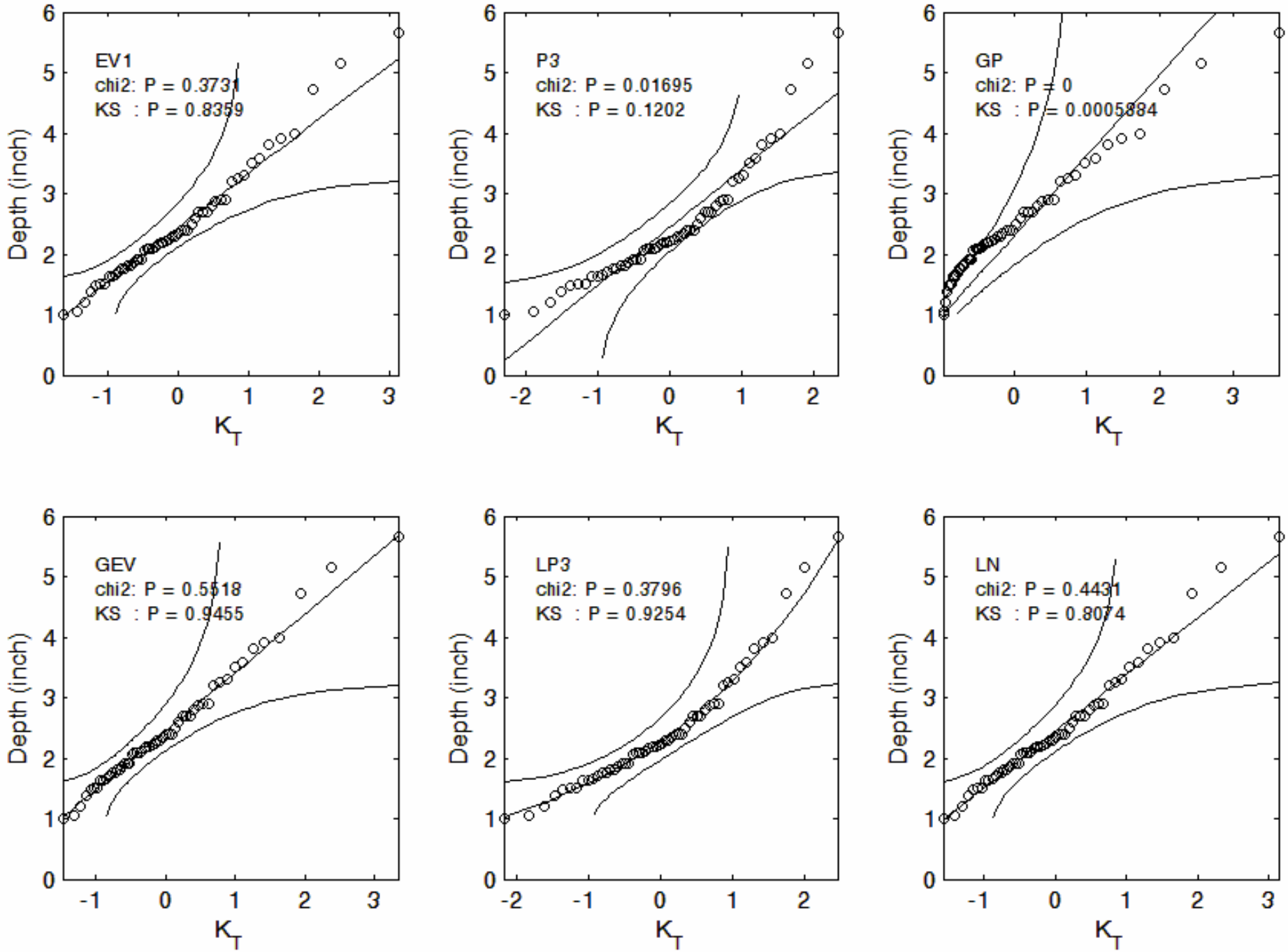
- Hydrologic designs are usually governed by depth (volume) or peak intensity
- Annual maximum volume (AMV) events
  - Longer duration
- Annual maximum peak intensity (AMI) events
  - Shorter duration
- Annual maximum cumulative probability (AMP) events
  - The use of empirical copulas C_n between volume and peak intensity
  - Wide range of durations



Analysis of Marginal Distributions (I)

- Candidate distributions
  - Extreme value type I (EV1)
  - Generalized extreme value (GEV)
  - Pearson type III (P3)
  - Log-Pearson type III (LP3)
  - Generalized Pareto (GP)
  - Log-normal (LN)
- Parameters estimated primarily by maximum likelihood (ML) or method of moments (MOM)
- Gringorton formula for empirical probabilities
- Chi-square and Kolmogorov-Smirnov (KS) test with 10% significance level

Identification of marginal distributions of volume (AMP events), Station 120132



Analysis of Marginal Distributions (II)

AMV events	Rejection rate (%) of Chi-square test						Rejection rate (%) of KS test					
	EV1	GEV	P3	LP3	GP	LN	EV1	GEV	P3	LP3	GP	LN
Depth, P	13.2	17.0	41.5	17.0	100	13.2	0.0	0.0	7.5	0.0	52.8	0.0
Duration, D	13.2	15.1	24.5	37.7	100	22.6	1.9	0.0	7.5	0.0	22.6	0.0
Intensity, I	15.1	17.0	45.3	20.8	100	11.3	0.0	0.0	1.9	0.0	54.7	0.0

AMI events	Rejection rate (%) of Chi-square test						Rejection rate (%) of KS test					
	EV1	GEV	P3	LP3	GP	LN	EV1	GEV	P3	LP3	GP	LN
Depth, P	5.7	3.8	62.3	3.8	100	1.9	0.0	0.0	11.3	0.0	45.3	0.0
Duration, D	60.4	39.6	88.7	37.7	100	28.3	15.1	0.0	45.3	0.0	45.3	0.0
Intensity, I	15.1	15.1	34.0	18.9	100	15.1	0.0	0.0	5.7	0.0	71.7	0.0

AMP events	Rejection rate (%) of Chi-square test						Rejection rate (%) of KS test					
	EV1	GEV	P3	LP3	GP	LN	EV1	GEV	P3	LP3	GP	LN
Depth, P	17.0	9.4	60.4	18.9	100	15.1	0.0	0.0	9.4	0.0	34.0	0.0
Duration, D	24.5	26.4	64.2	26.4	100	18.9	0.0	0.0	13.2	1.9	15.1	0.0
Intensity, I	7.5	17.0	43.4	18.9	100	9.4	0.0	0.0	1.9	3.8	62.3	0.0

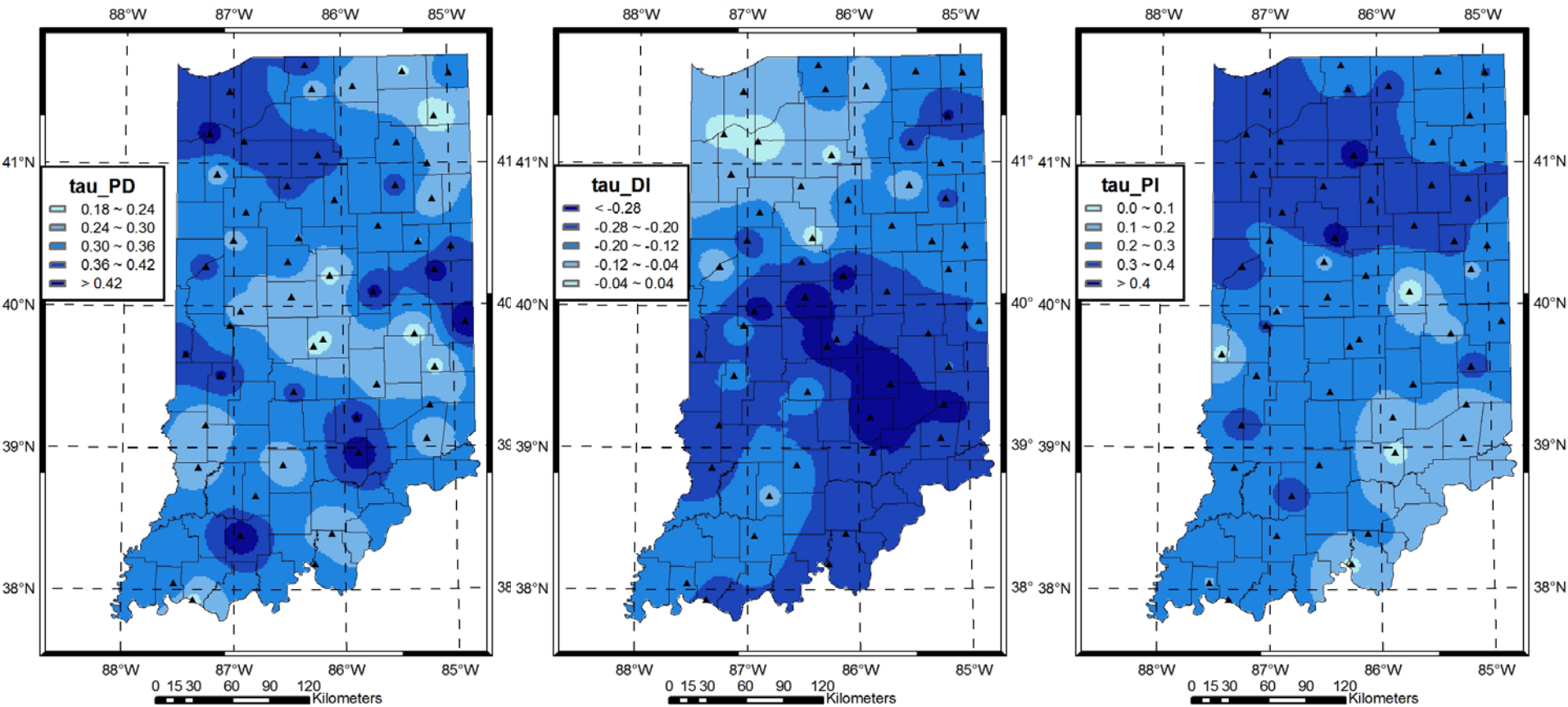
- EV1, GEV, LP3, LN provided better fit. GP provided the worst.
- Fitting for duration of AMI events did not yield very good result
- EV1 and LN could be recommended for use

Analysis of Dependence Structure (I)

- Candidate Archimedean copulas
  - Frank family
  - Clayton family
  - Genest-Ghoudi family
  - Ali-Mikhail-Haq family
- Non-parametric procedure for estimating dependence parameter
- Examination of Goodness-of-fit
  - Distribution function $K_C(t)=P[C(u,v)\leq t]$
  - Diagonal section of copulas $\delta(t)=C(t,t)$
  - Section with one marginal as median (one marginal equals 0.5)
  - Multidimensional KS test (Saunders and Laud, 1980)

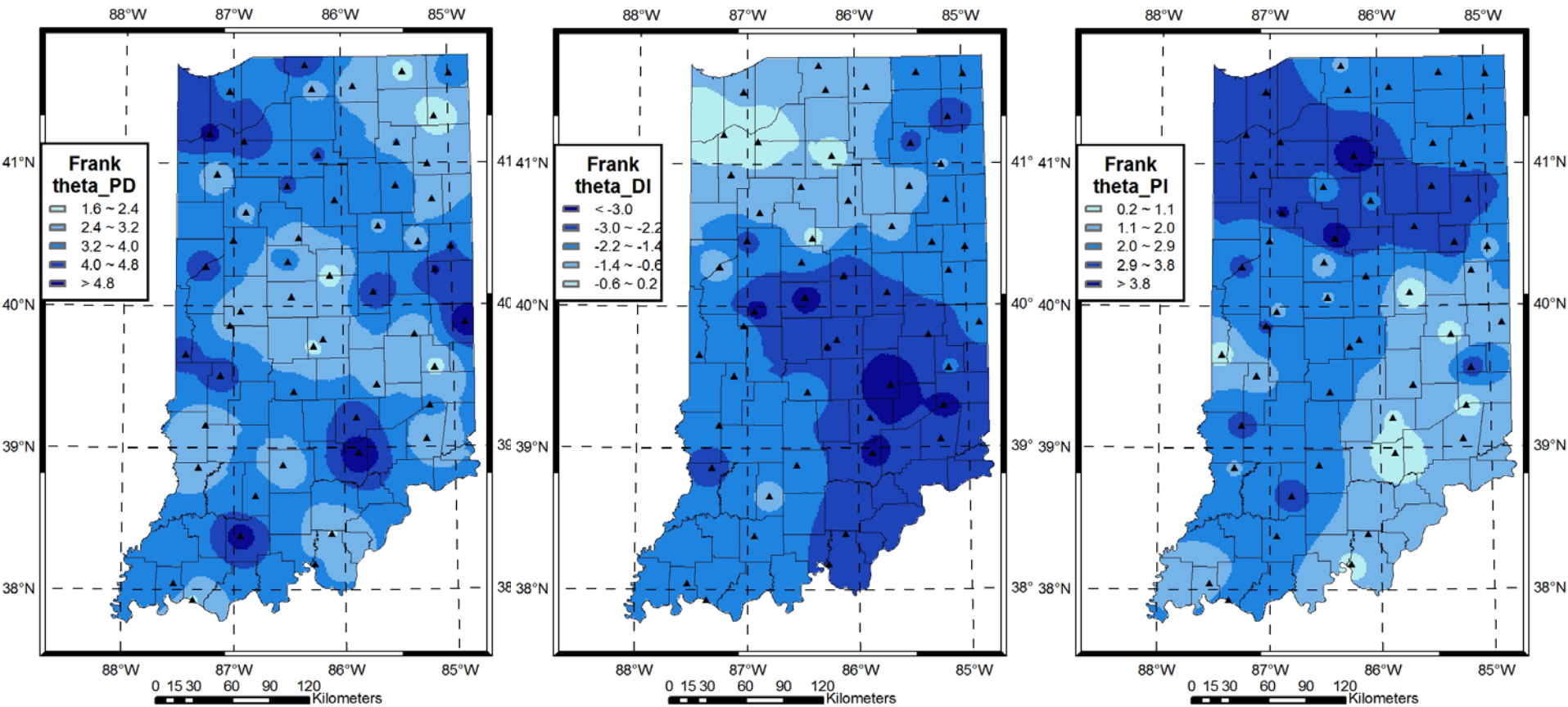
Variation of Kendall's τ

	τ_{PD}		τ_{DI}		τ_{PI}	
	mean	stdev	mean	stdev	mean	stdev
AMV events	0.183	0.084	-0.370	0.068	0.260	0.097
AMI events	0.407	0.070	-0.011	0.096	0.405	0.070
AMP events	0.324	0.078	-0.185	0.093	0.265	0.094

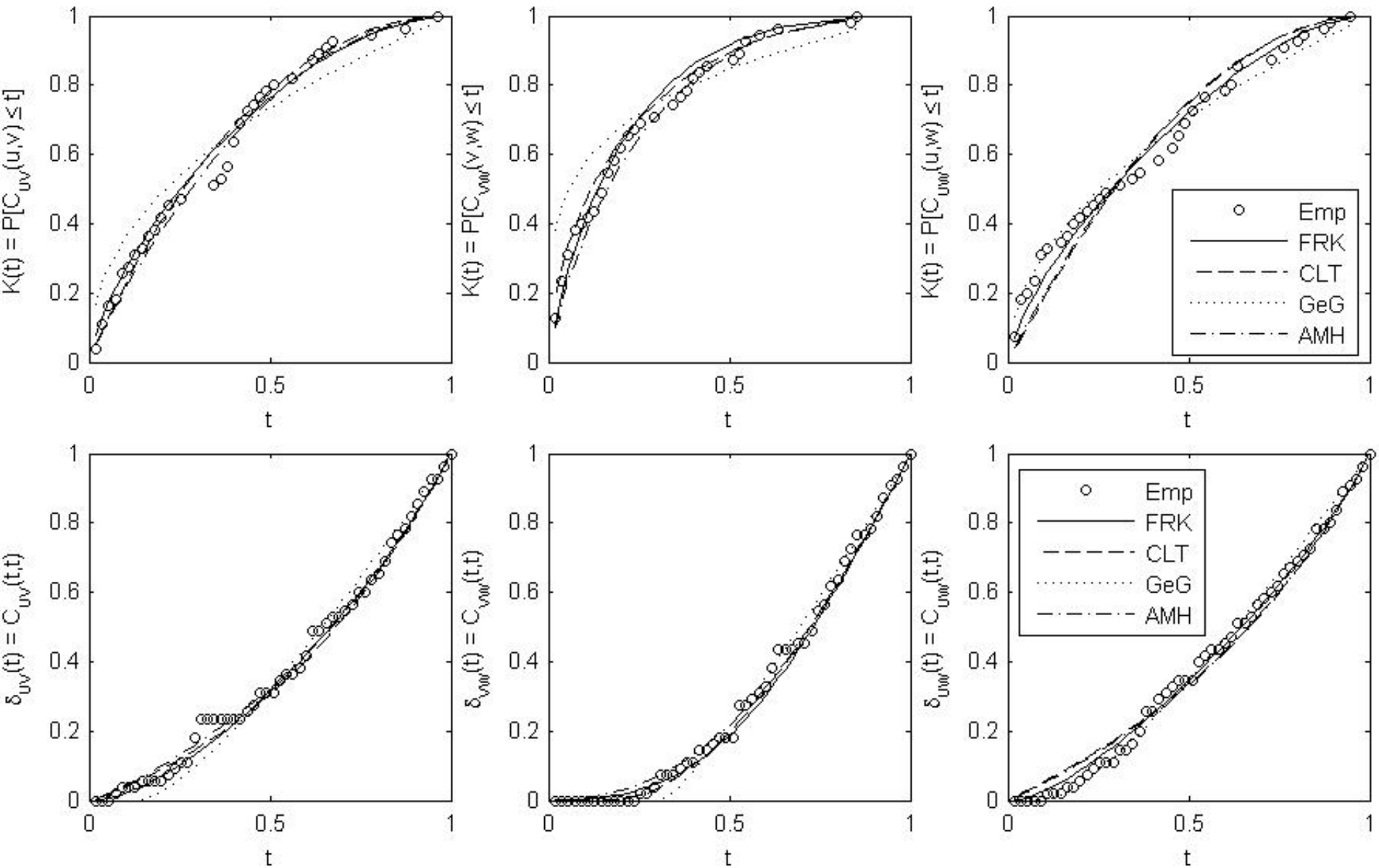


Variation of θ (Frank family)

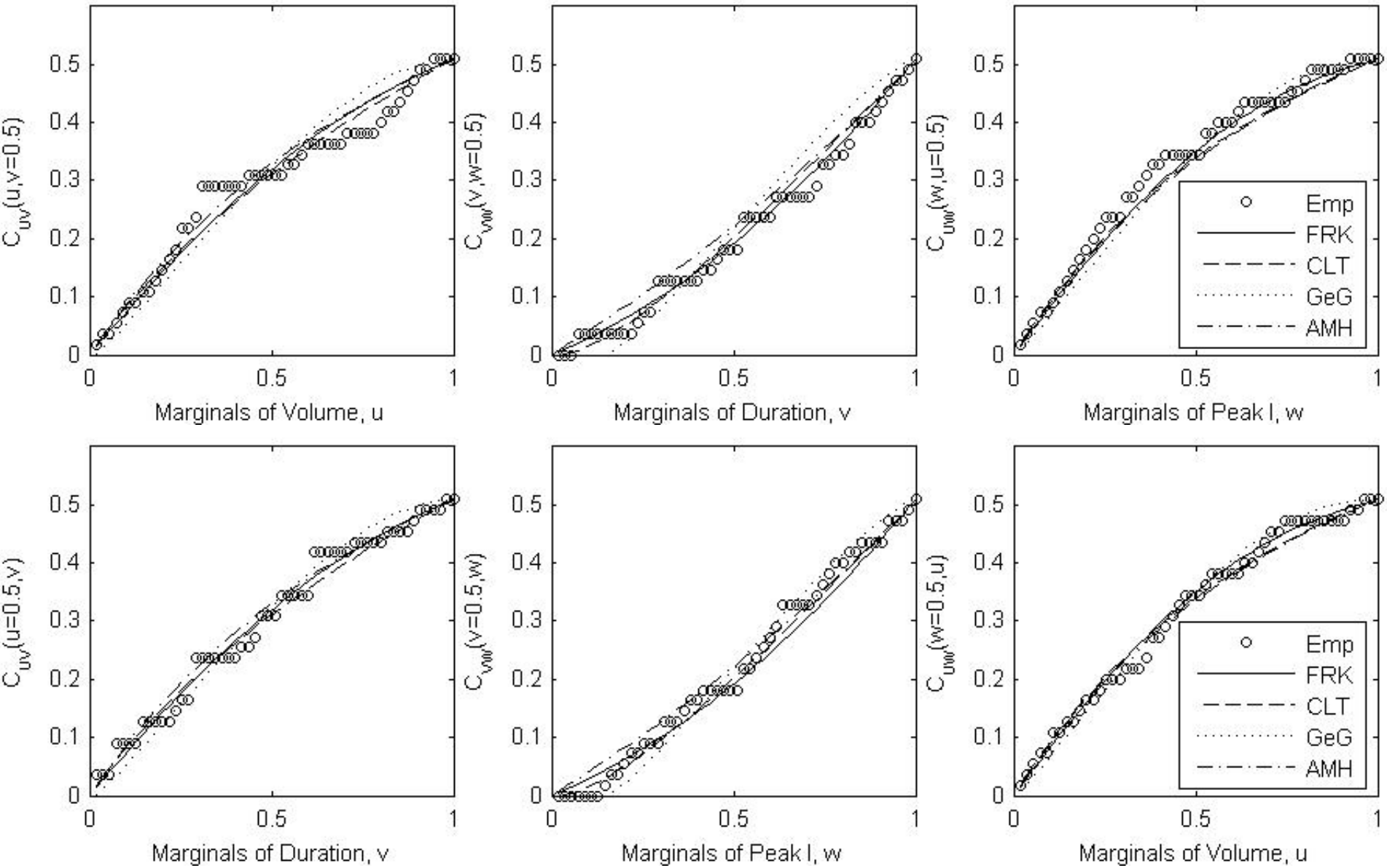
Frank family	θ_{UV}		θ_{VW}		θ_{UW}	
	mean	stdev	mean	stdev	mean	stdev
AMV events	1.726	0.825	-3.824	0.909	2.546	1.063
AMI events	4.333	1.003	-0.111	0.883	4.314	0.986
AMP events	3.410	0.975	-1.863	0.927	2.389	1.029



Assessment of Copula Performance (I)



Assessment of Copula Performance (I)



Analysis of Dependence Structure (II)

- The distribution function $K_C(t)$ provides the strictest examination of copulas

- Clayton and Ali-Mikhail-Haq families performed well for positive dependence cases (C_{UV} and C_{UW})

- Frank family of Archimedean copulas
  - performed well for both positive and negative dependence
  - passed the KS test for entire Indiana at the 10% significant level
  - recommended for use in practice

Construct Joint Distribution via Copulas

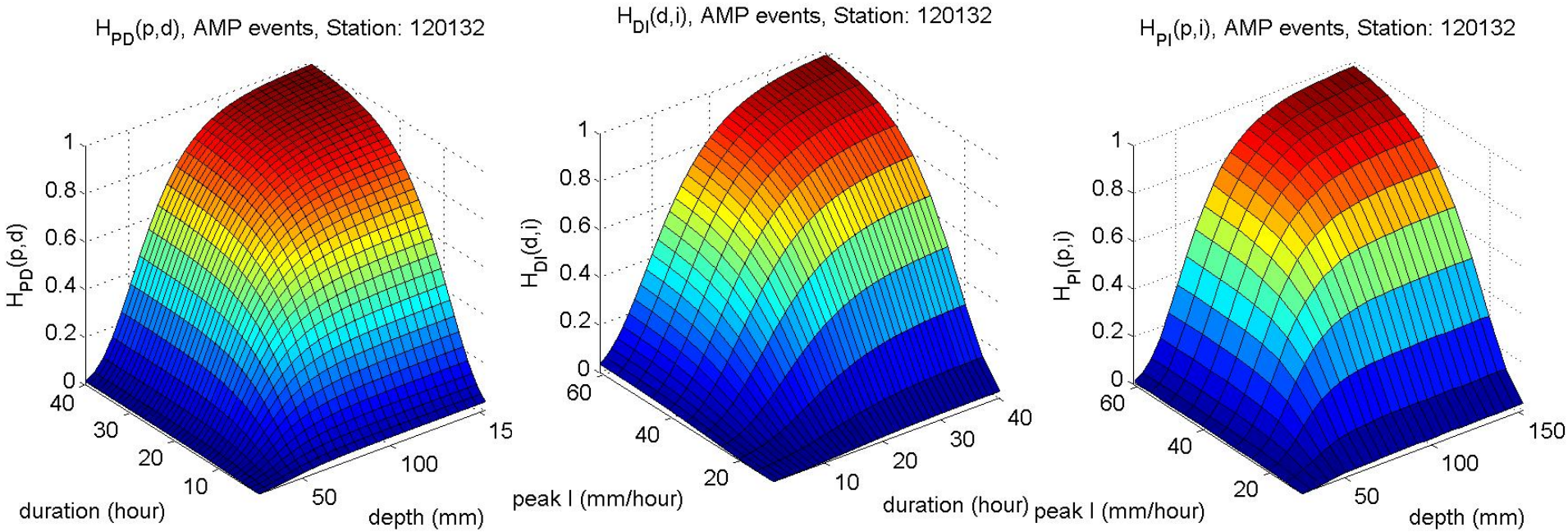
- Bivariate stochastic models

$$H_{PD}(p,d) = C_{UV}(F_P(p), F_D(d)) = C_{UV}(u,v)$$

$$H_{DI}(d,i) = C_{VW}(F_D(d), F_I(i)) = C_{VW}(v,w)$$

$$H_{PI}(p,i) = C_{UW}(F_P(p), F_I(i)) = C_{UW}(u,w)$$

- Examples using Frank family and EV1 marginals



Application 1

Estimate of depth for known duration (I)

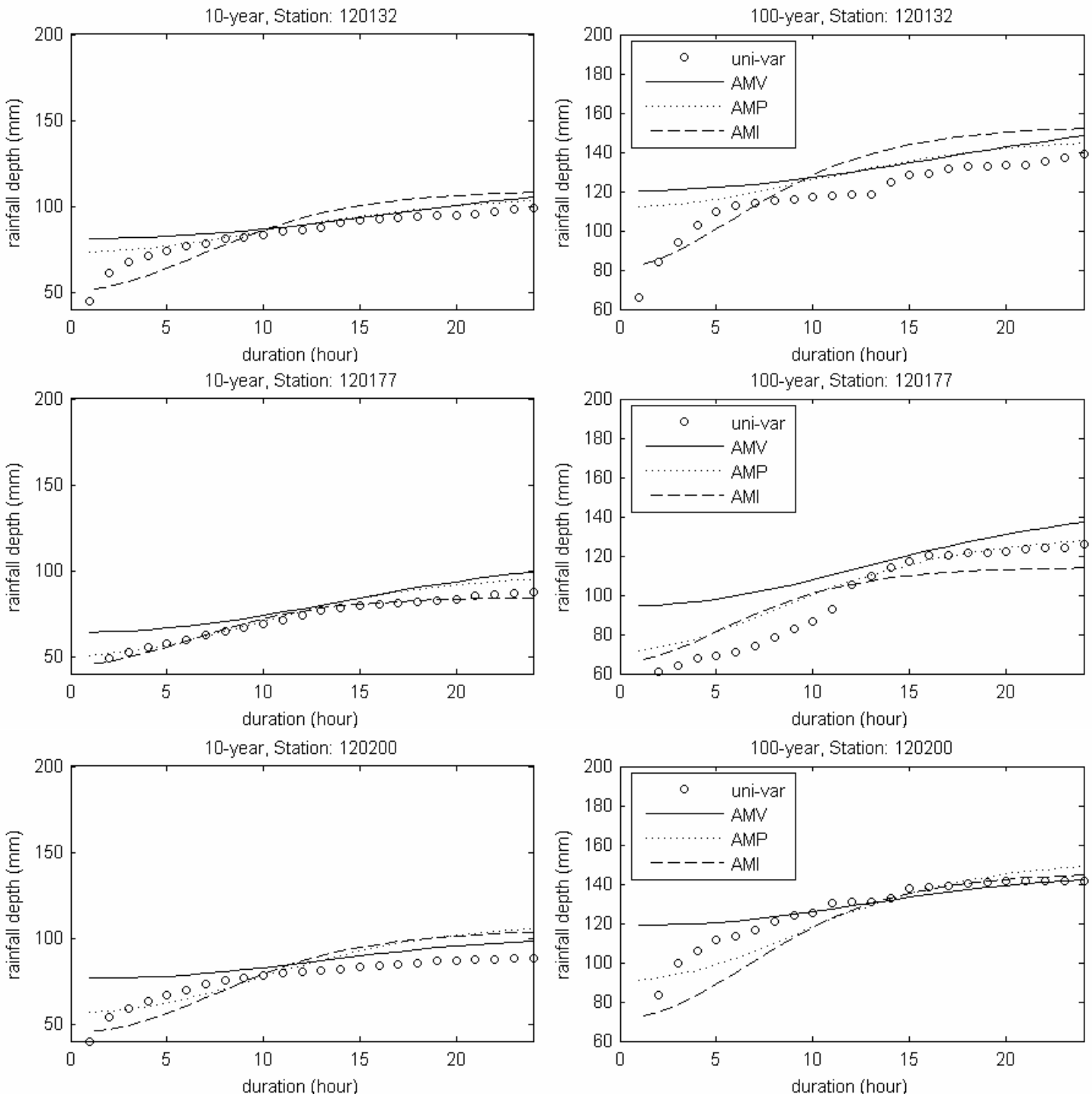
- For a known (or measured) d-hour event

$$F_P\left(p \middle| d-1 < D \le d\right) = \frac{H_{PD}(p,d) - H_{PD}(p,d-1)}{F_D(d) - F_D(d-1)}$$
$$= \frac{C_{UV}\left(F_P(p), F_D(d)\right) - C_{UV}\left(F_P(p), F_D(d-1)\right)}{F_D(d) - F_D(d-1)}$$

- Given return period T, the T-year, d-hour rainfall estimate p_T will satisfy

$$F_P\left(p_T \middle| d-1 < D \le d\right) = 1 - 1/T$$

- Comparison between bivariate and univariate depth estimates
  - Bivariate using EV1 marginals and Frank family
  - Univariate counterpart using GEV distribution (Rao and Kao, 2006)


10-year, Station: 120132
100-year, Station: 120132
10-year, Station: 120177
100-year, Station: 120177
10-year, Station: 120200
100-year, Station: 120200
rainfall depth (mm)
duration (hour)
uni-var
AMV
AMP
AMI

Estimate of depth for known duration (II)

- Similar trends were observed for durations greater than 10-hour, close to the univariate counterpart
- For durations less than 10-hour
  - Univariate approach underestimated the rainfall depth
  - AMV estimates gave the highest value
  - AMI estimates should be the best, but fitting problem existed
  - AMP estimates are recommended
- Average ratios for entire Indiana

duration	AMV/GEV	AMP/GEV	AMI/GEV
1	1.98	1.51	1.18
2	1.50	1.16	0.93
3	1.33	1.04	0.86
4	1.24	0.99	0.85
6	1.14	0.96	0.89
9	1.07	0.99	0.98
12	1.05	1.03	1.04
18	1.03	1.07	1.05
24	1.03	1.08	1.03

Application 2
Estimate of peak intensity for known duration (I)

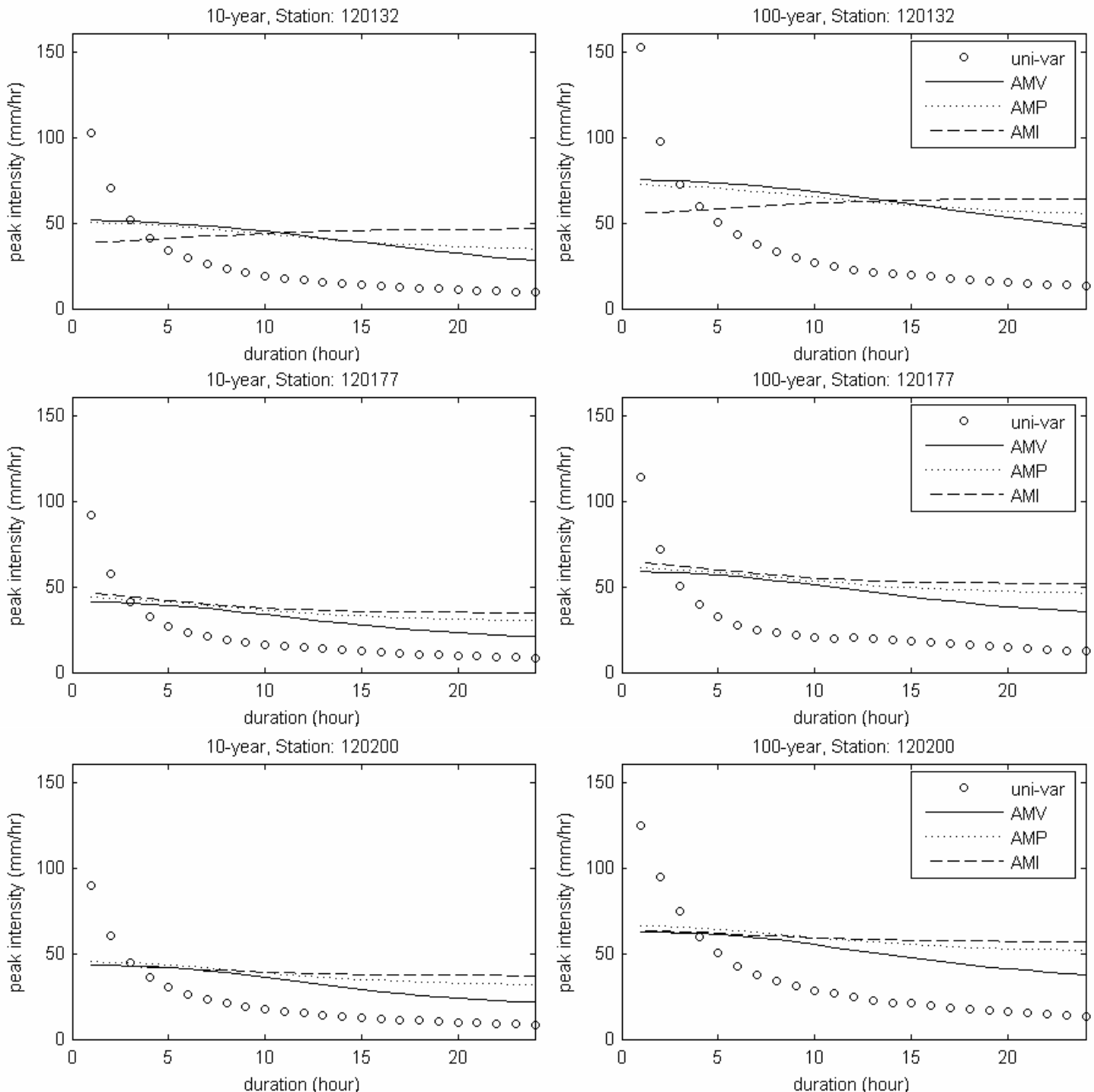
- For a known (or measured) d-hour event

$$F_I\left(i \middle| d-1 < D \le d\right) = \frac{H_{DI}(d,i) - H_{DI}(d-1,i)}{F_D(d) - F_D(d-1)}$$
$$= \frac{C_{VW}\left(F_D(d), F_I(i)\right) - C_{VW}\left(F_D(d-1), F_I(i)\right)}{F_D(d) - F_D(d-1)}$$

- Given return period T, the T-year, d-hour rainfall estimate p_T will satisfy

$$F_I\left(i_T \middle| d-1 < D \le d\right) = 1 - 1/T$$

- Comparison between bivariate and univariate depth estimates
  - Bivariate using EV1 marginals and Frank family
  - Univariate counterpart using GEV depth with Huff (1967) temporal distribution derived at each station


10-year, Station: 120132
100-year, Station: 120132
10-year, Station: 120177
100-year, Station: 120177
10-year, Station: 120200
100-year, Station: 120200
peak intensity (mm/hr)
duration (hour)
uni-var
AMV
AMP
AMI
0
50
100
150
0
5
10
15
20

Estimate of peak intensity for known duration (II)

- Similar trends were observed between AMV, AMI, and AMP estimates.
- AMI generally provided the largest estimates, unless positive dependence existed between D and I
- Univariate approach
  - Peak intensity generated by GEV depth with Huff distribution is around 4-5 times larger than the average intensity
  - Followed the IDF relationship
  - Failed to capture peak intensity
- AMP estimates are recommended

Application 3

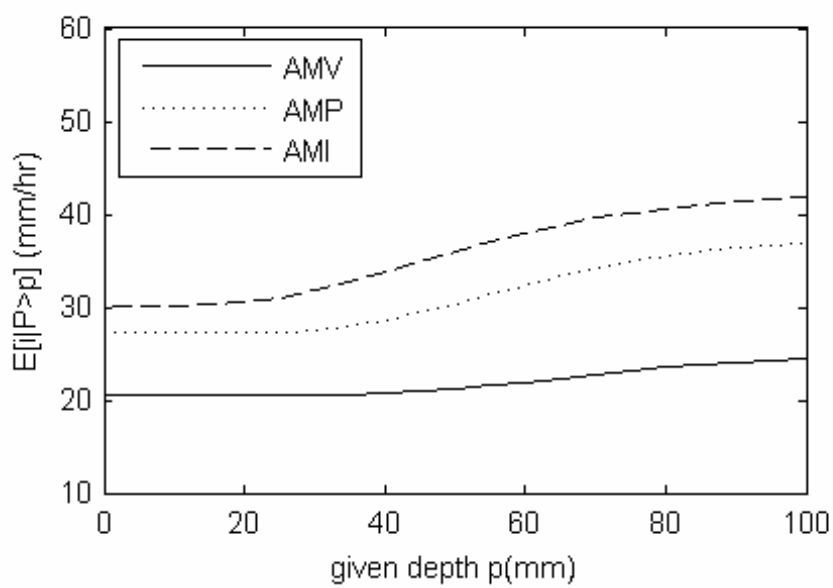
Estimate of peak intensity for known depth (I)

- For extreme events greater than a threshold p

$$F_I(i|P > p) = \frac{F_I(i) - H_{PI}(p,i)}{1 - F_P(p)} = \frac{w - C_{UW}(u,w)}{1-u}$$

- Conditional expectation E[I | P>p]

$$E[I|P > p] = \int_0^\infty i f_I(i|P > p)di = \int_0^\infty i \frac{\partial}{\partial i} F_I(i|P > p)di$$



Conclusions (I)

- Definition of extreme events
  - AMV events are generally of longer duration than AMP, following by AMI events. AMV events may therefore be less reliable for short durations.
  - For AMI definition, the hourly recording precision used in this study was found to be limiting
  - AMP criterion seems to be an appropriate indicator for defining extreme events

- Marginal distributions
  - EV1, GEV, LP3, LN were found to be appropriate marginal models for extreme rainfall
  - EV1 and LN are recommended

Conclusions (II)

- Dependence structure
  - Between P and D, positive correlated
  - Between D and I, generally negatively correlated
  - Between P and I, positive correlated
  - Frank family is recommended
  - Indiana rainfall may not be homogeneous in the multi-variate sense
- Estimate of depth for known duration
  - Similar results for durations larger than 10 hours
  - AMP estimates are recommended to use for durations less than 10 hours
- Estimate of peak intensity for known duration
  - Conventional approach fails to capture the peak intensity
  - AMP definition is recommended

Thank you for listening.
Questions?