Dependence

Copula

Applications

Motivation

1. What is the joint probability of concurrent heavy precipitation and high streamflow?

Motivation

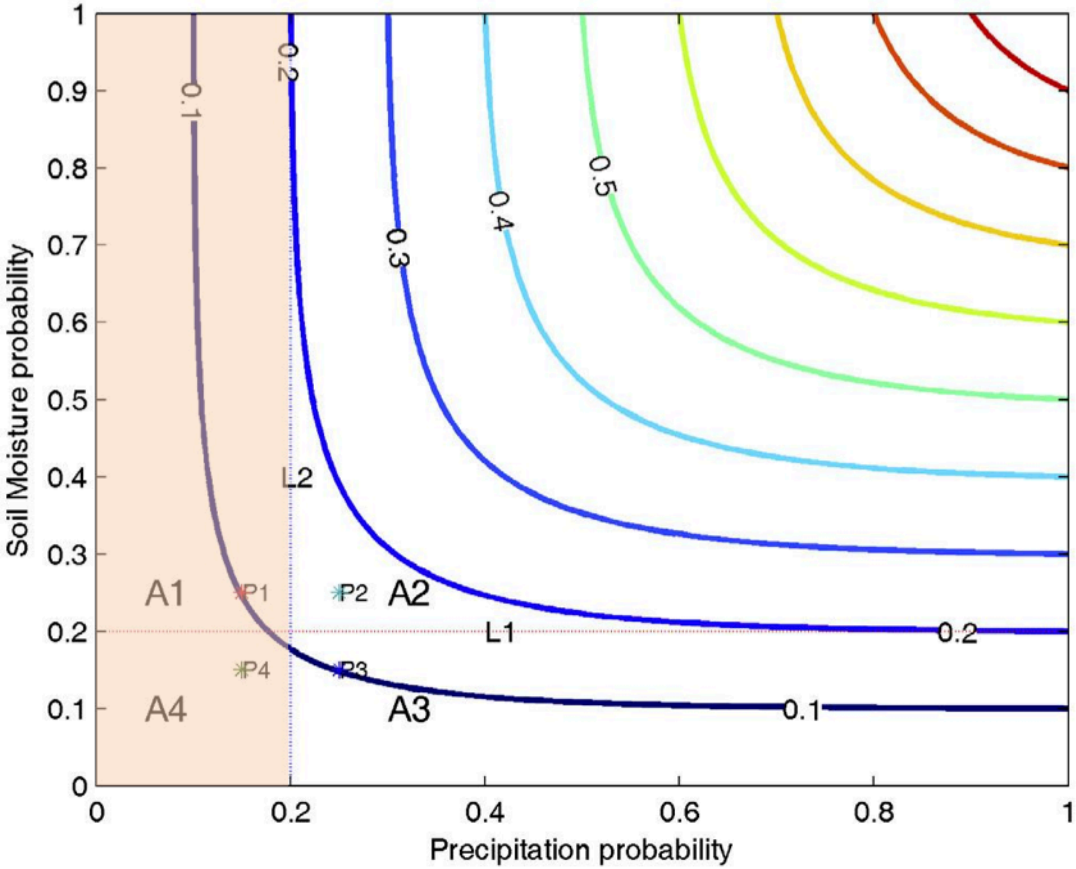
1. What is the joint probability of concurrent heavy precipitation and high streamflow?

What is the joint probability of low soil moisture and heatwave?

What is the joint probability of heatwave, drought severity and duration?

…

Motivation



Precipitation

$$p_p = P(X \leq x)$$

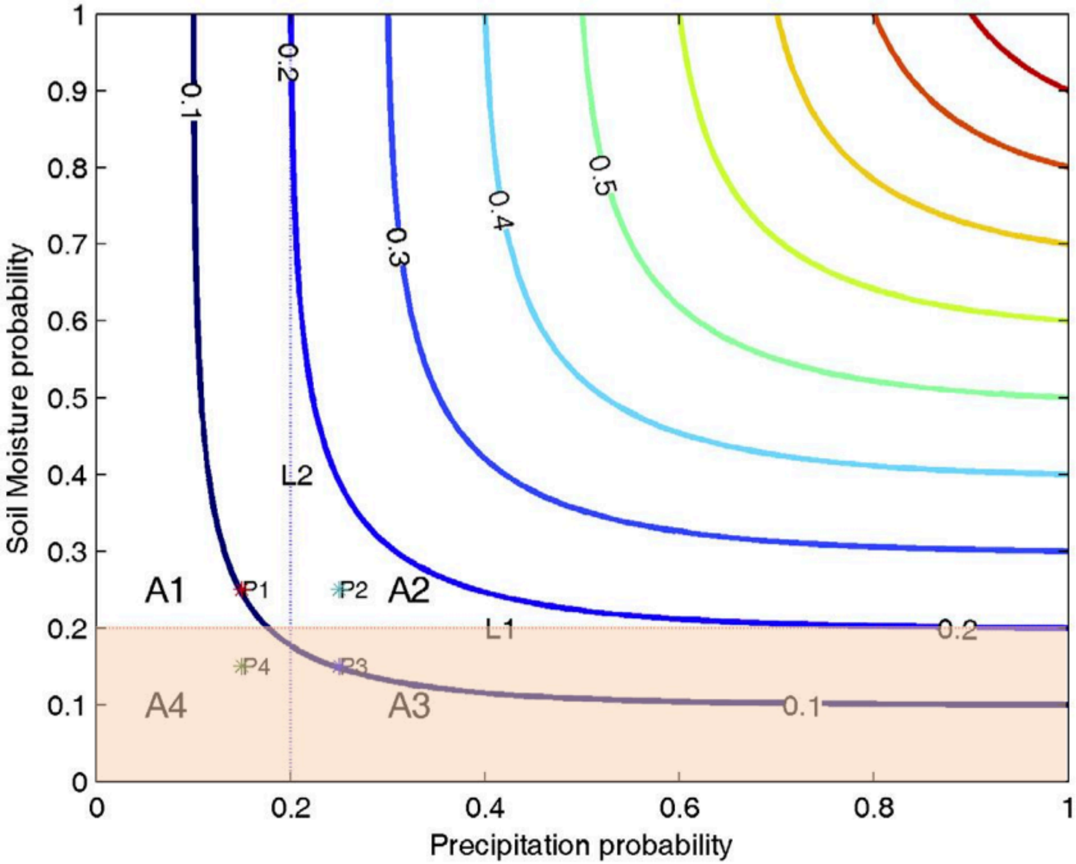
Soil moisture

$$p_{sm} = P(Y \leq y)$$

$$p_{p-sm} = P(X \leq x, Y \leq y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Motivation



Precipitation

$$p_p = P(X \leq x)$$

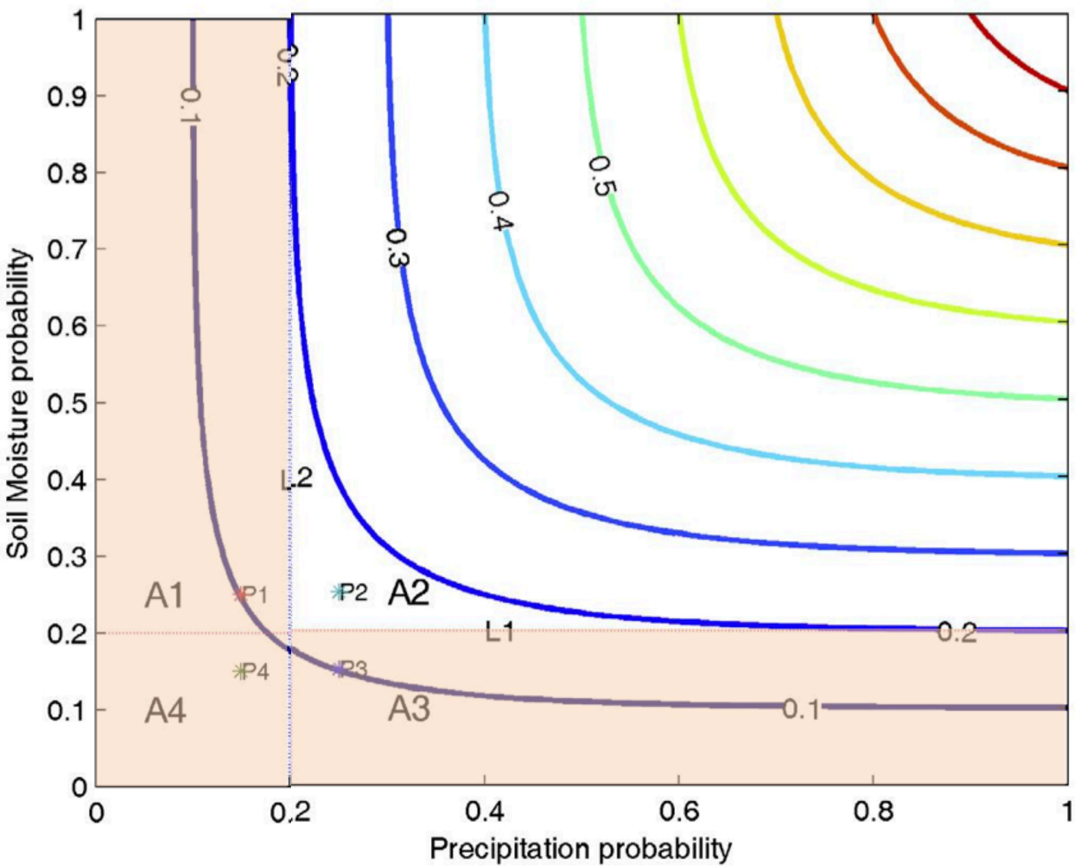
Soil moisture

$$p_{sm} = P(Y \leq y)$$

$$p_{p-sm} = P(X \leq x, Y \leq y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Motivation



Precipitation

$$p_p = P(X \leq x)$$

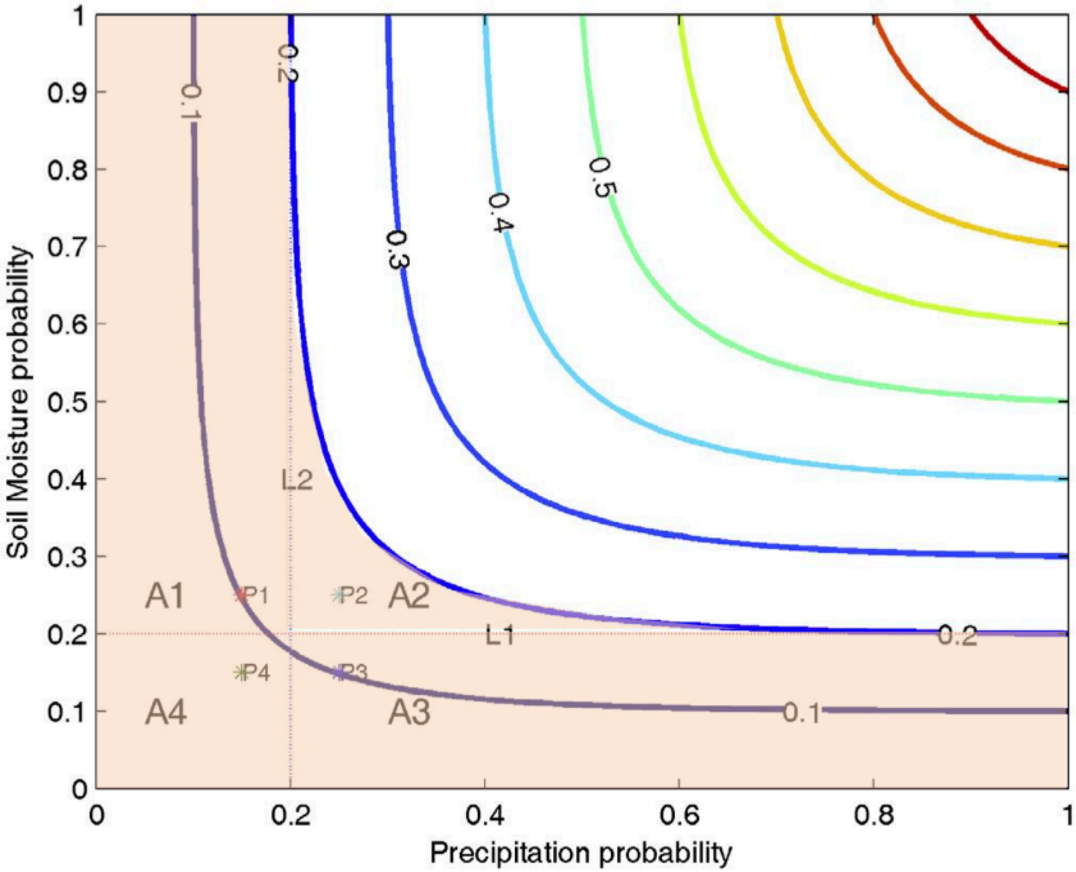
Soil moisture

$$p_{sm} = P(Y \leq y)$$

$$p_{p-sm} = P(X \leq x, Y \leq y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Motivation



Precipitation

$$p_p = P(X \leq x)$$

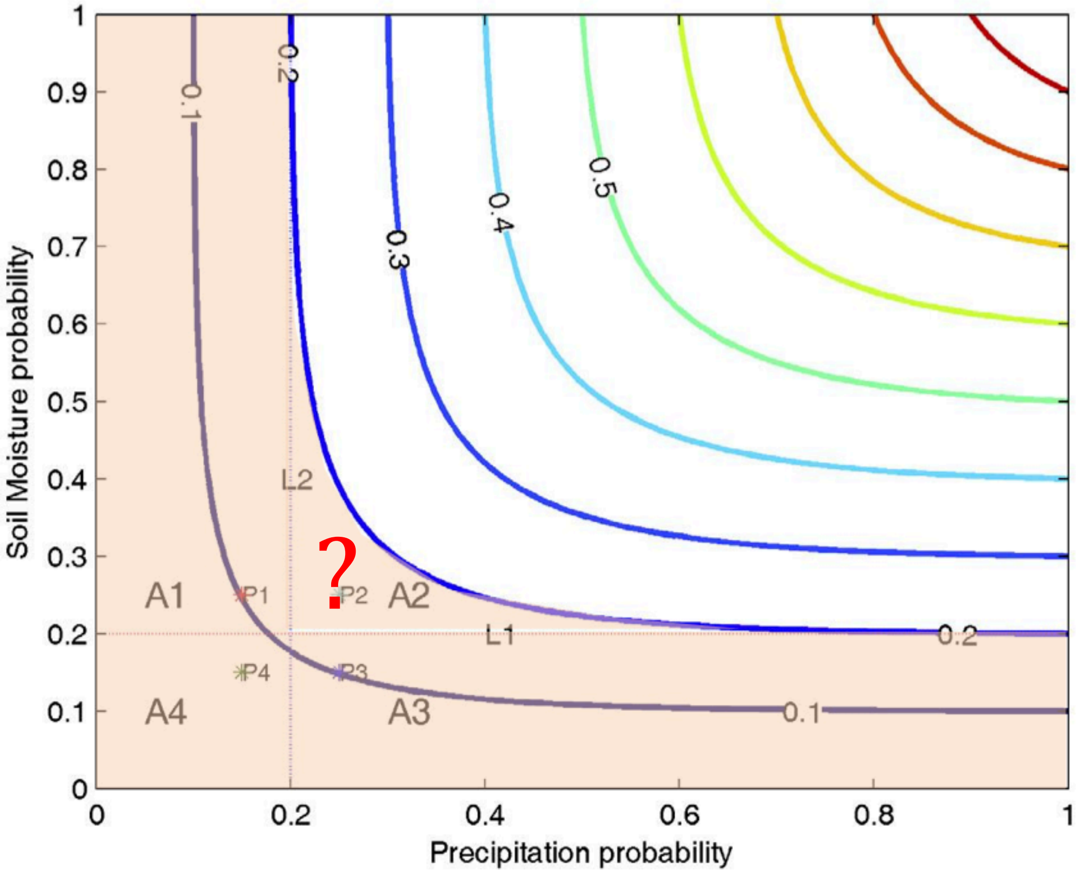
Soil moisture

$$p_{sm} = P(Y \leq y)$$

$$p_{p-sm} = P(X \leq x, Y \leq y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Motivation



Precipitation

$$p_p = P(X \le x)$$

Soil moisture

$$p_{sm} = P(Y \le y)$$

$$p_{p-sm} = P(X \le x, Y \le y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Measures of linear dependence

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

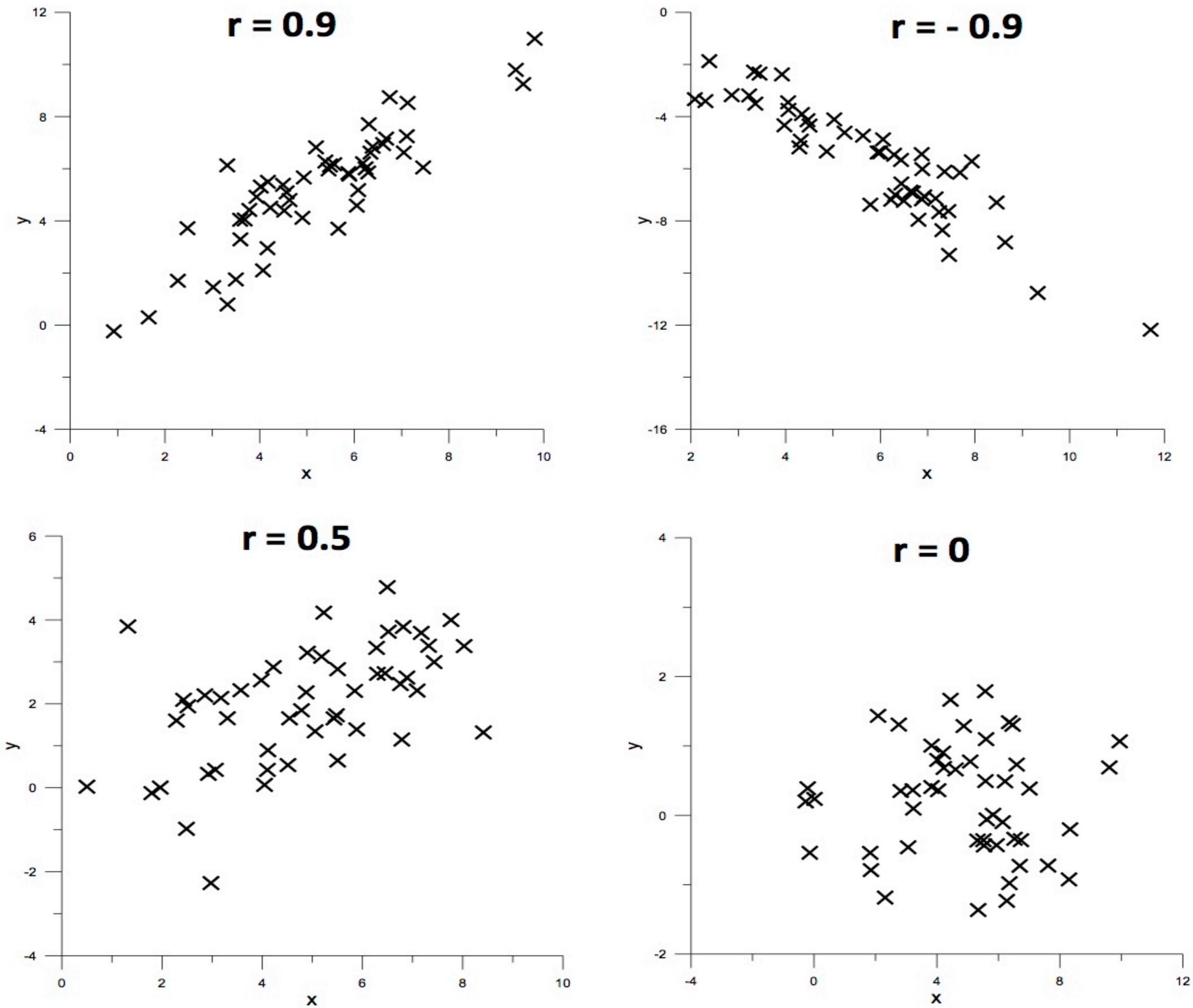
Covariance

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

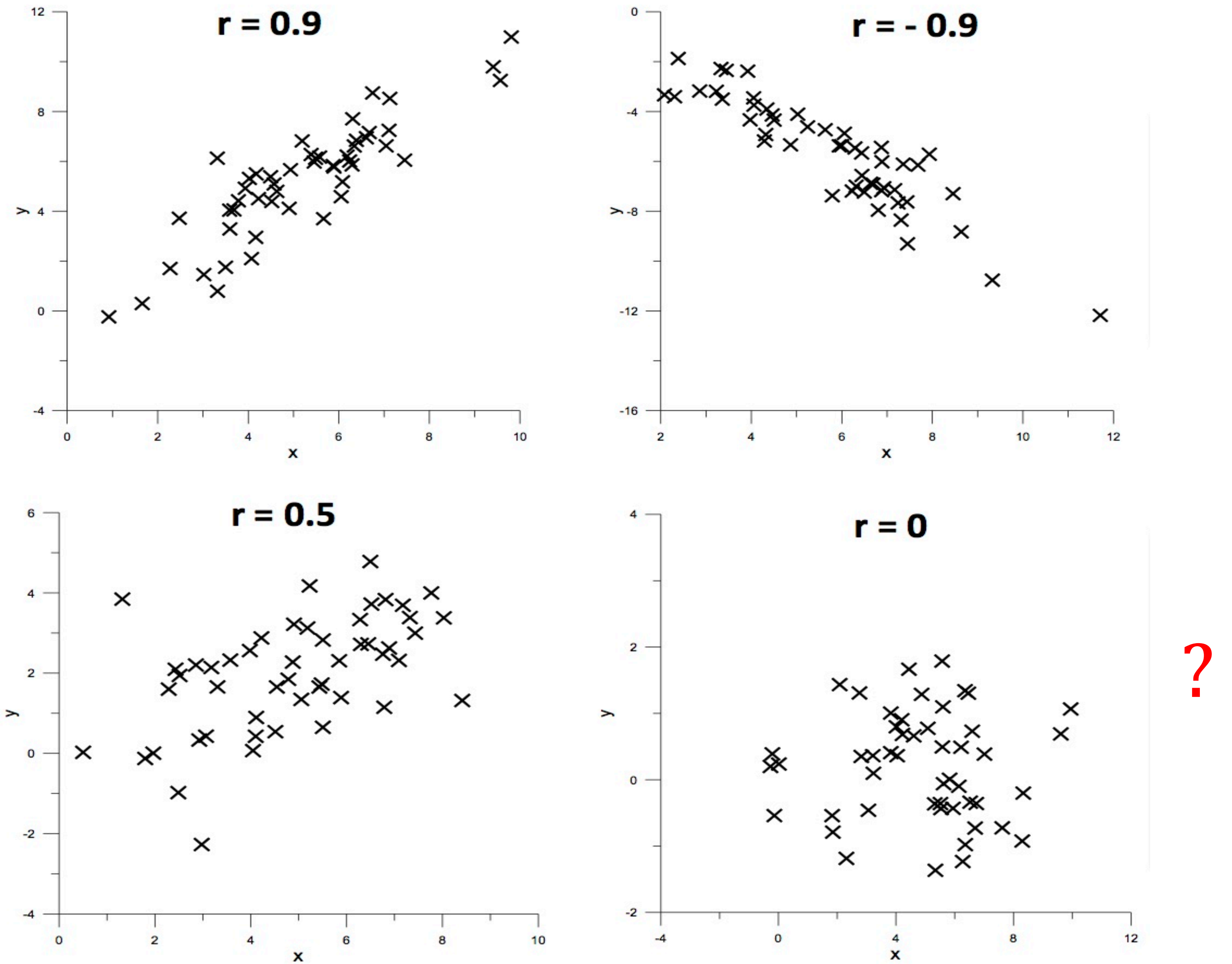
Correlation Coefficient

$$-1 \le r_{xy} \le +1$$

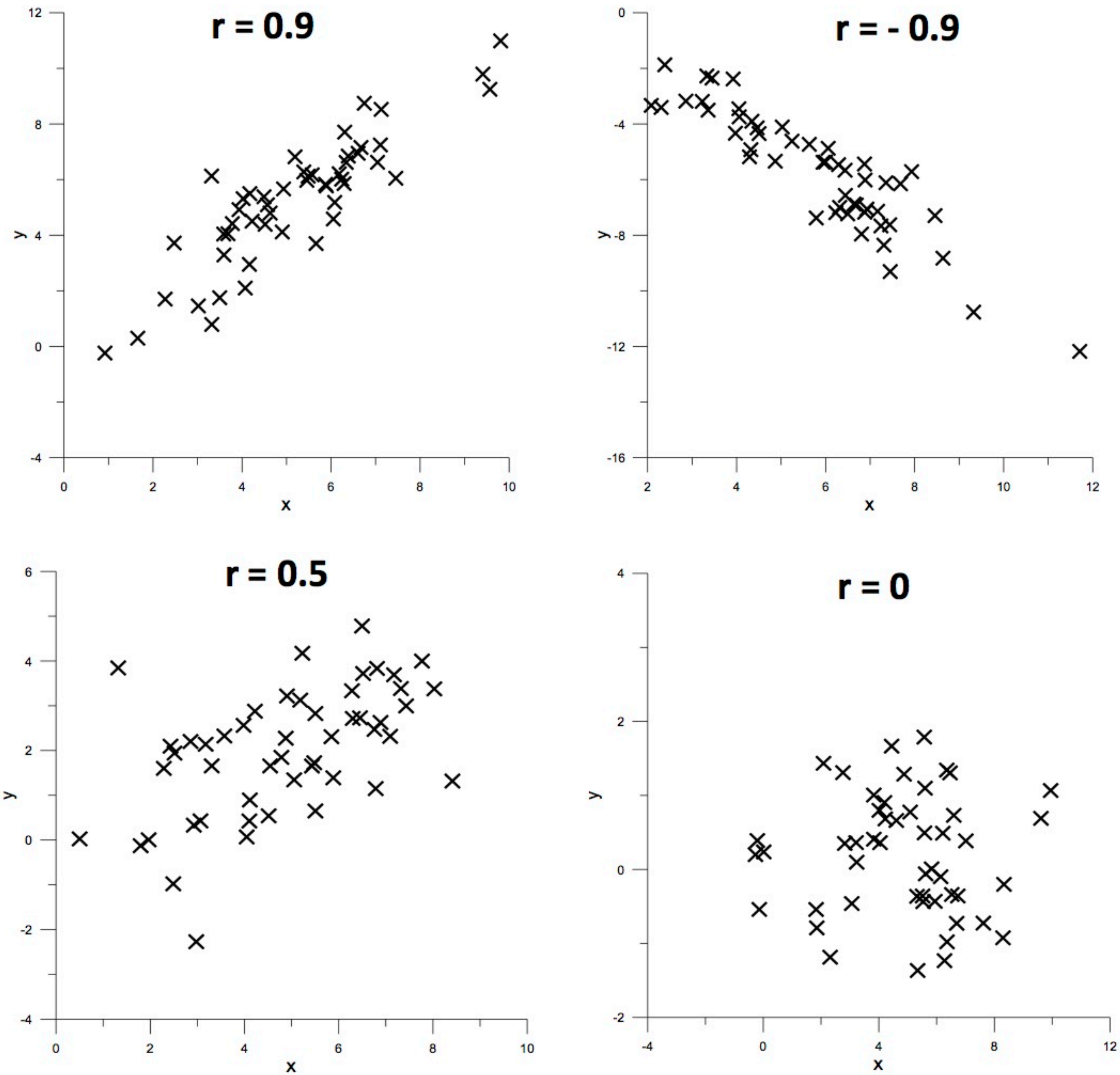
Correlation and Covariance



Correlation and Covariance



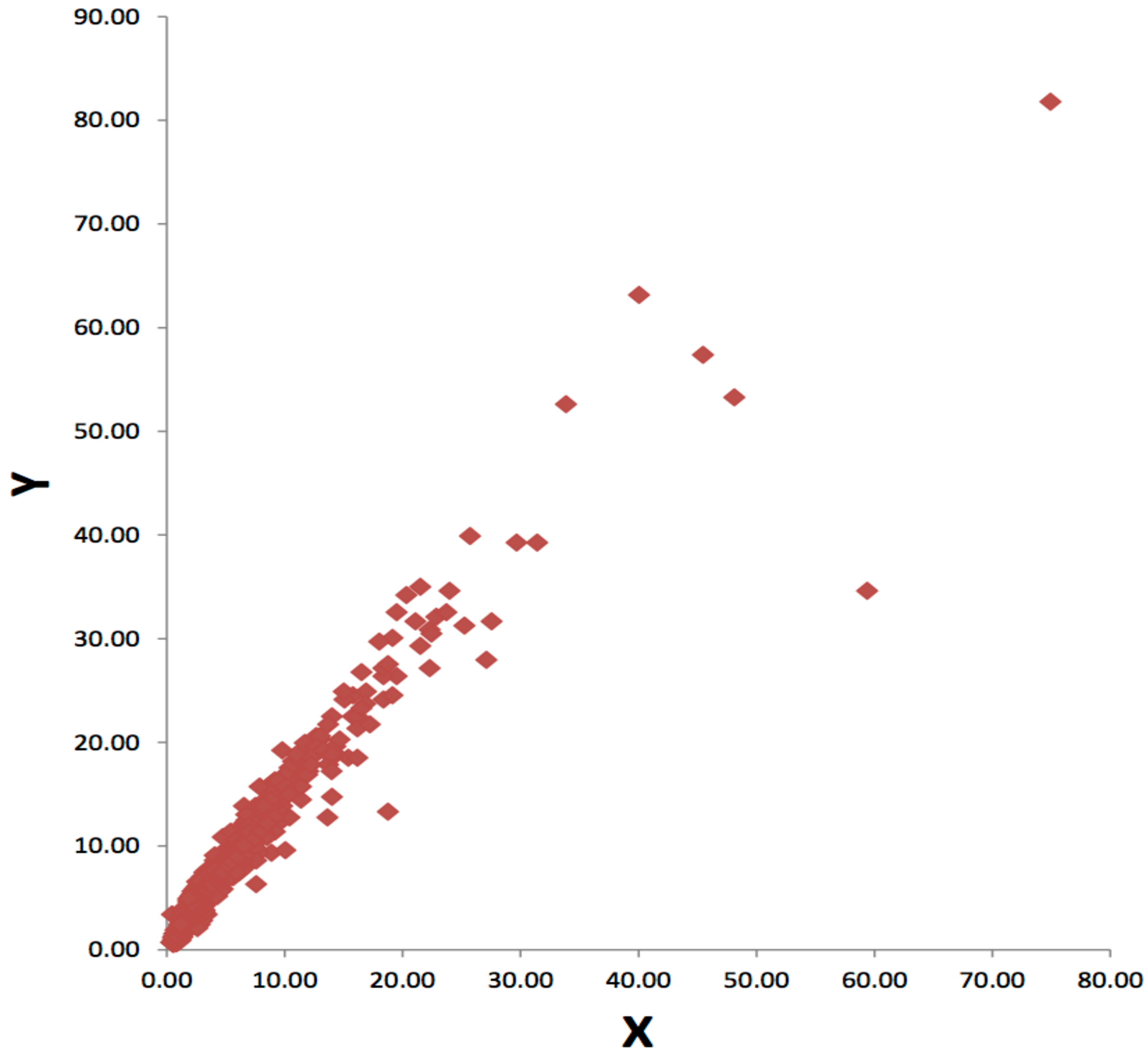
Correlation and Covariance



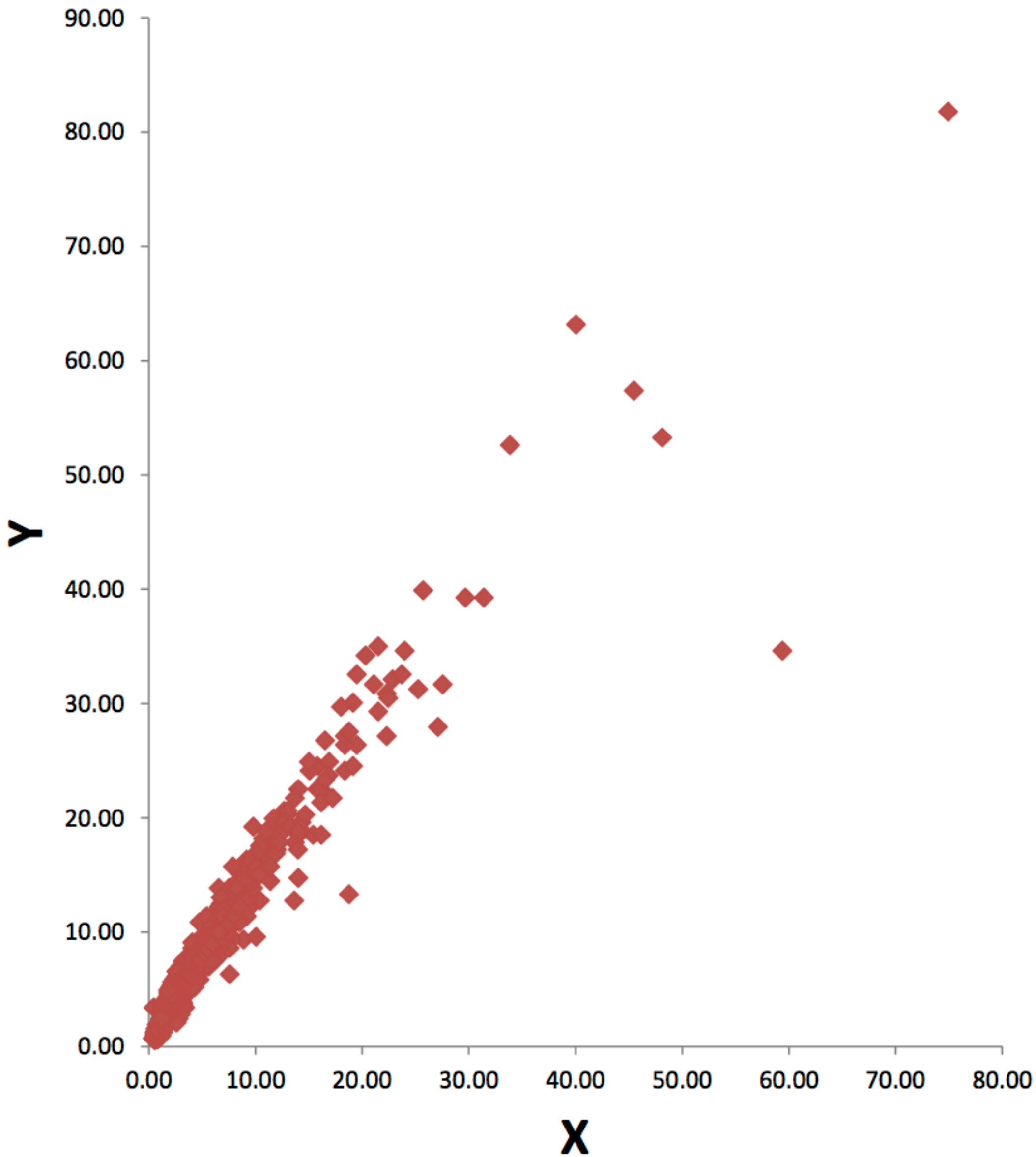
Correlation only describes the linear dependence between variables

A correlation coefficient of 0 correlation does not mean that there is no dependence

Dependence Concepts



Dependence Concepts

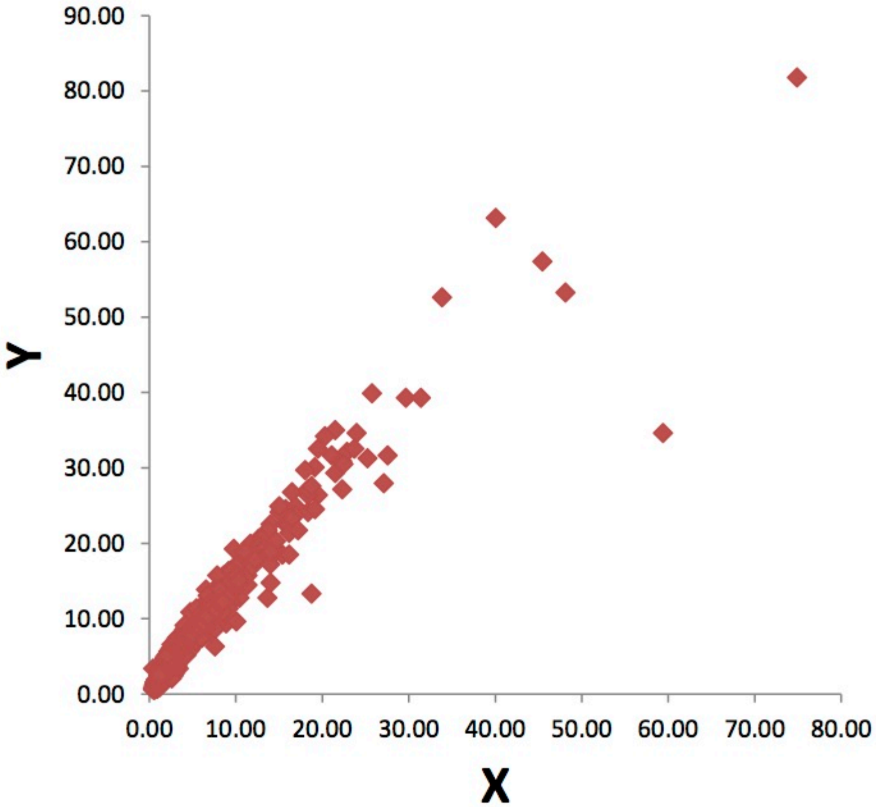


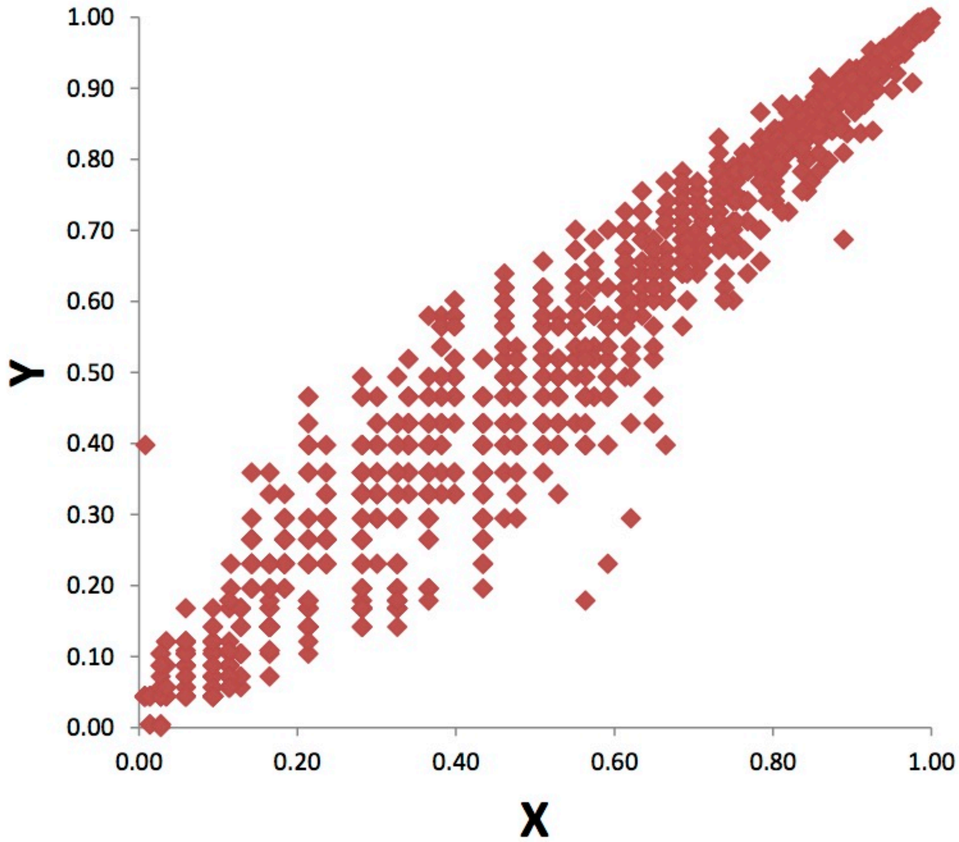
What is the dependence between large values of X and Y?

How large values are different than small values?

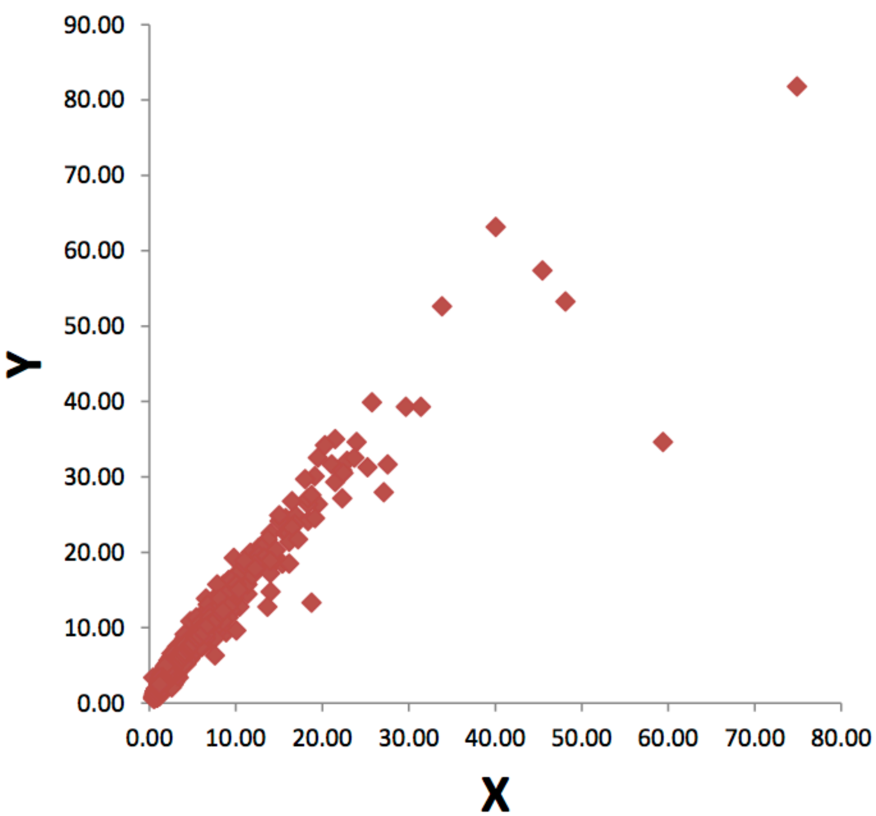
Where in the distribution there is a stronger relationship?

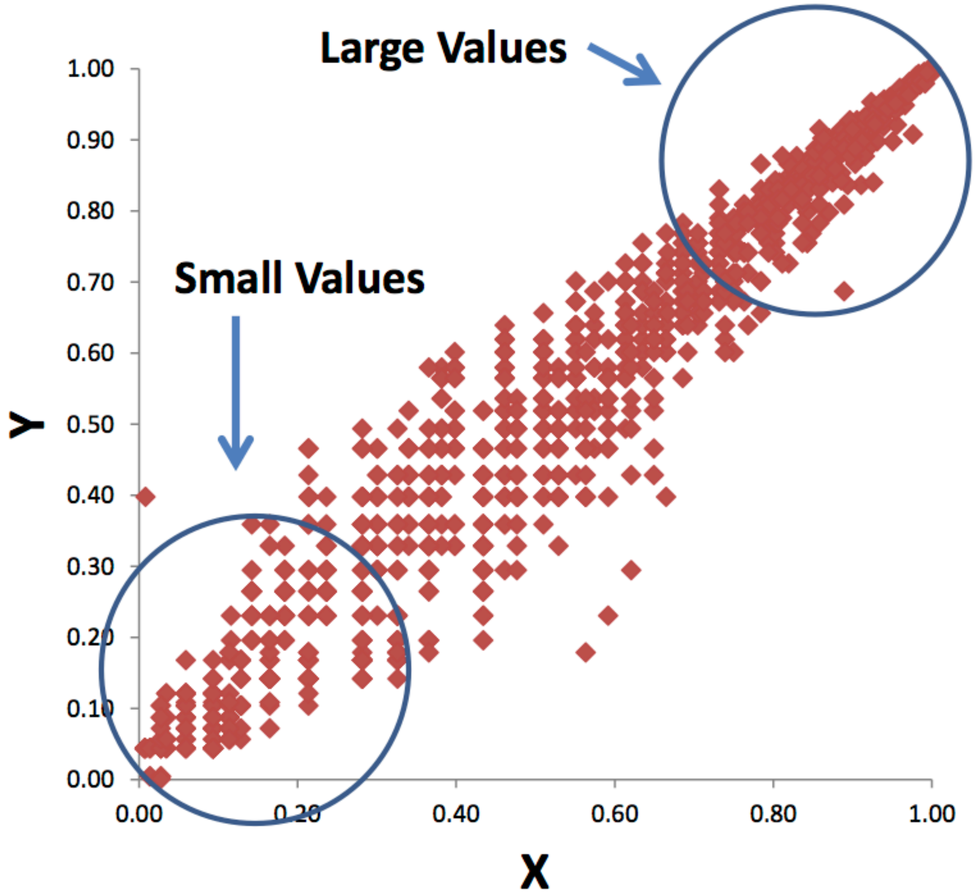
Dependence Concepts





Dependence Concepts

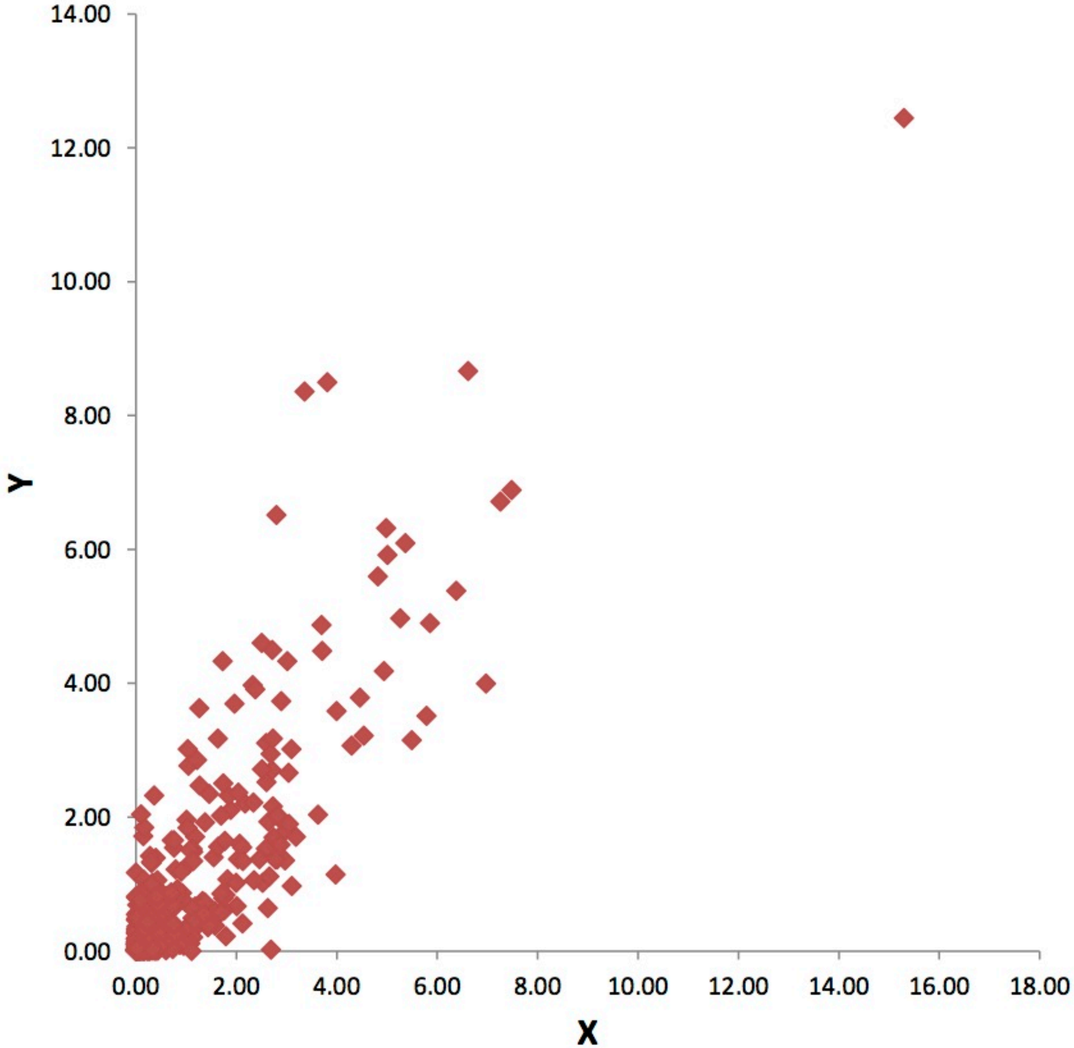




Large values of X and Y are strongly associated with each other

The dependence between large values is stronger than small values

Dependence Concepts

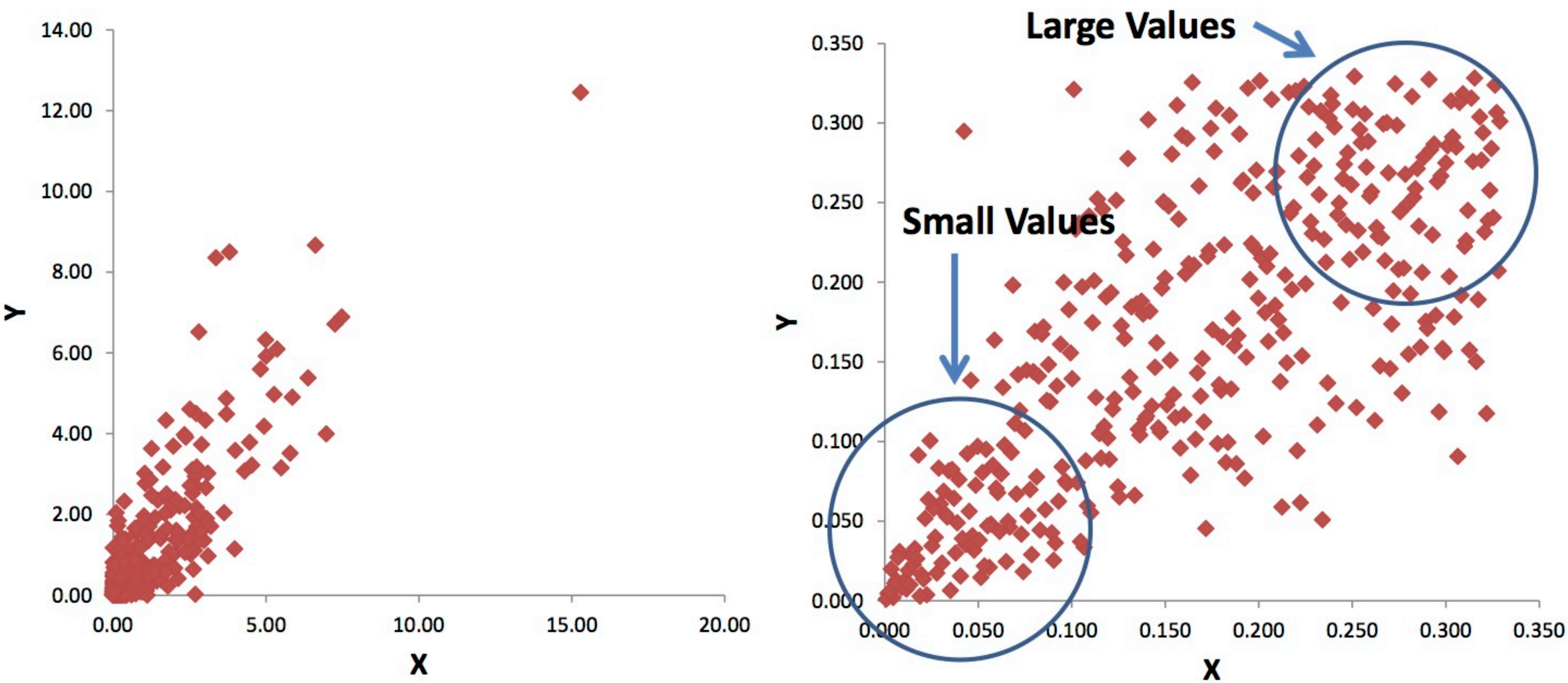


What is the dependence between large values of X and Y?

How large values are different than small values?

Where in the distribution there is a stronger relationship?

Dependence Concepts



The dependence between small values is stronger than large values

Transformation to uniform marginals

$$(x_i, y_i) \qquad i = 1, \ldots, n$$

$$\left(\frac{n - R(x_i) + \frac{1}{2}}{n}, \frac{n - R(y_i) + \frac{1}{2}}{n} \right) \qquad i = 1, \ldots, n$$

$$R(x_i) = \text{the rank of } x_i \text{ in the set } \{x_1, \ldots, x_n\}$$

Rank Correlation

The above transformation dissociates the correlation structure between variables from their marginal distributions.

Rank Correlation Methods:

Spearman's Rank Correlation Coefficient ρ_s

Kendall's Rank Correlation Coefficient τ

Rank correlation methods measure the degree of monotone (increasing or decreasing) dependence (or association) between two variables.

Spearman's Rank Correlation Coefficient ρs

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}$$

Here, d_i denotes differences between the ranks of two variables

Spearman's Rank Correlation Coefficient ρs

X	Y	Rank (X_i)	Rank (Y_i)	d_i
9	28.4	1	1	0
15	29.3	2	2	0
24	37.6	3	7	-4
30	36.2	4	4.5	-0.5
38	36.5	5	6	-1
46	35.3	6	3	3
53	36.2	7	4.5	2.5
60	44.1	8	8	0
64	44.8	9	9	0
76	47.2	10	10	0
				$\Sigma d^2=32.5$

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$$

$$\rho_s = 1 - \frac{6(32.5)}{10(99)}$$

$$= 0.80$$

Spearman's Rank Correlation Coefficient ρs

X	Y	Rank (X_i)	Rank (Y_i)	d_i
9	28.4	1	1	0
15	29.3	2	2	0
24	37.6	3	7	-4
30	36.2	4	4.5	-0.5
38	36.5	5	6	-1
46	35.3	6	3	3
53	36.2	7	4.5	2.5
60	44.1	8	8	0
64	44.8	9	9	0
76	47.2	10	10	0
				Σd^2=32.5

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

$$\rho_s = 1 - \frac{6(32.5)}{10(99)}$$

= **0.80**

For $n > 10$

Z= $\rho_s\sqrt{n-1}$

= 0.80 * $\sqrt{9}$

=**2.4**

p-value = 1- $\emptyset(z)$

= 1- normcdf(2.4)

=**0.0082**

Spearman's Rank Correlation Coefficient ρs

X	Y	Rank (X_i)	Rank (Y_i)	d_i
9	28.4	1	1	0
15	29.3	2	2	0
24	37.6	3	7	-4
30	36.2	4	4.5	-0.5
38	36.5	5	6	-1
46	35.3	6	3	3
53	36.2	7	4.5	2.5
60	44.1	8	8	0
64	44.8	9	9	0
76	47.2	10	10	0
				Σd^2=32.5

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

$$\rho_s = 1 - \frac{6(32.5)}{10(99)}$$

= 0.80

strong positive dependence

Dependence Concepts

Spearman's rho is the linear correlation between $F_1(X)$ and $F_2(Y)$, which are integral transforms of X and Y. In this sense it is a measure of rank correlation. Both $\rho_S(X,Y)$ and $\rho_\tau(X,Y)$ are measures of monotonic dependence between (X,Y). Both measures are based on the concept of **concordance**, which refers to the property that large values of one random variable are associated with large values of another, whereas discordance refers to large values of one being associated with small values of the other.

Copulas and Dependence

- A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables.

$$p = P(X \leq x, Y \leq y)$$

$$p = C[F(X), G(Y)]$$

C is the copula and *F(X)* and *G(Y)* are the marginal cumulative distribution functions of precipitation (*X*) and soil moisture (*Y*), respectively

Copulas and Dependence

- Sklar's Theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables.

$$p = P(X \leq x, Y \leq y)$$

$$p = C\left[F(X), G(Y)\right]$$

Copulas and Dependence

- Sklar's Theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables.

$$p = P(X \leq x, Y \leq y)$$

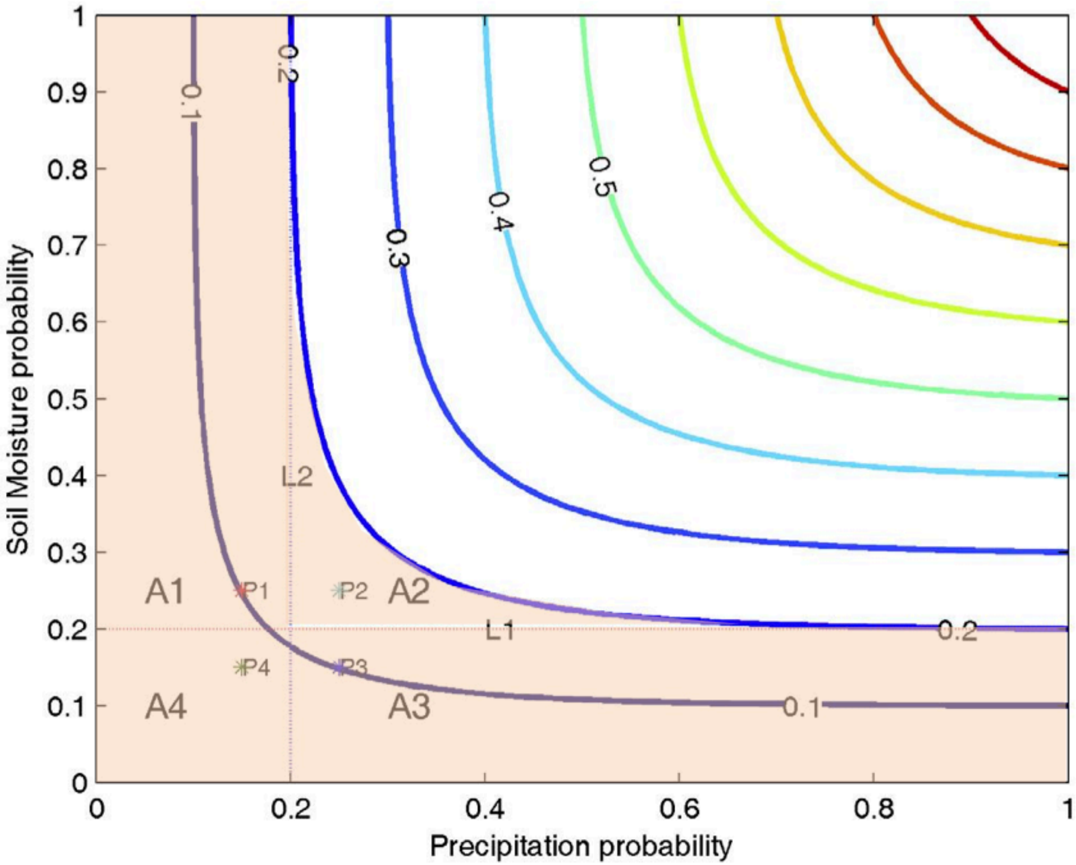
$$p = C[F(X), G(Y)]$$

1) Choice of arbitrary marginal distributions:

They could take different forms;
They could involve covariates.

2) Choice of an arbitrary copula function (dependence structure).

Copulas and Dependence



Precipitation

$$p_p = P(X \leq x)$$

Soil moisture

$$p_{sm} = P(Y \leq y)$$

$$p_{p-sm} = P(X \leq x, Y \leq y)$$

Where: *X*: accumulated precipitation;
Y: accumulated soil moisture;

Copulas and Dependence

- Copulas are popular in high-dimensional statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating marginals and copulae separately.

- There are many parametric copula families available, which usually have parameters that control the strength of dependence.

Copula type	Function $C(u_1, u_2)$
Product	$u_1 u_2$
FGM	$u_1 u_2(1 + \theta(1 - u_1)(1 - u_2))$
Gaussian	$\Phi_G[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta]$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$
Frank	$-\frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right)$
Ali-Mikhail-Haq	$u_1 u_2(1 - \theta(1 - u_1)(1 - u_2))]^{-1}$

- Relationship with Spearman's correlation coefficient and Kendall's correlation coefficient

Both $\rho_S(X,Y)$ and $\rho_\tau(X,Y)$ can be expressed in terms of copulas as follows:

$$\rho_S(X,Y) = 12\int_0^1\int_0^1 \{C(u_1,u_2) - u_1u_2\}\,du_1du_2,$$

$$\rho_\tau(X,Y) = 4\int_0^1\int_0^1 C(u_1,u_2)\,dC(u_1,u_2) - 1$$

- Relationship with Spearman's correlation coefficient and Kendall's correlation coefficient

Copula type	Function $C(u_1, u_2)$	θ-domain	Kendall's τ	Spearman's ρ
Product	$u_1 u_2$	N.A.	0	0
FGM	$u_1 u_2(1+\theta(1-u_1)(1-u_2))$	$-1 \leq \theta \leq +1$	$\frac{2}{9}\theta$	$\frac{1}{3}\theta$
Gaussian	$\Phi_G[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta]$	$-1 < \theta < +1$	$\frac{2}{\pi}\arcsin(\theta)$	$\frac{6}{\pi}\arcsin(\frac{\theta}{2})$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$	$\frac{\theta}{\theta+2}$	*
Frank	$-\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right)$	$\theta \in (-\infty, \infty)$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$	$1 - \frac{12}{\theta}[D_1(\theta) - D_2(\theta)]$
Ali-Mikhail-Haq	$u_1 u_2(1-\theta(1-u_1)(1-u_2))]^{-1}$	$-1 \leq \theta \leq 1$	$(\frac{3\theta-2}{\theta}\ )$ $-\frac{2}{3}(1-\frac{1}{\theta}\ )^2 \ln(1-\theta)$	*

$D_k(x)$ denotes the "Debye" function $k/x^k \int_0^x \frac{t^k}{(e^t-1)} dt$, $k = 1, 2$

Copulas and Dependence

Copula type	Function $C(u_1, u_2)$	θ-domain	Kendall's τ	Spearman's ρ
Product	$u_1 u_2$	N.A.	0	0
FGM	$u_1 u_2(1+\theta(1-u_1)(1-u_2))$	$-1 \leq \theta \leq +1$	$\frac{2}{9}\theta$	$\frac{1}{3}\theta$
Gaussian	$\Phi_G[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta]$	$-1 < \theta < +1$	$\frac{2}{\pi}\arcsin(\theta)$	$\frac{6}{\pi}\arcsin(\frac{\theta}{2})$
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$	$\frac{\theta}{\theta+2}$	*
Frank	$-\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right)$	$\theta \in (-\infty, \infty)$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$	$1 - \frac{12}{\theta}[D_1(\theta) - D_2(\theta)]$
Ali-Mikhail-Haq	$u_1 u_2(1-\theta(1-u_1)(1-u_2))]^{-1}$	$-1 \leq \theta \leq 1$	$(\frac{3\theta-2}{\theta}\ )$ $-\frac{2}{3}(1-\frac{1}{\theta}\ )^2 \ln(1-\theta)$	*

$D_k(x)$ denotes the "Debye" function $k/x^k \int_0^x \frac{t^k}{(e^t-1)}dt,\ k = 1, 2$

Theoretical Rank Correlation

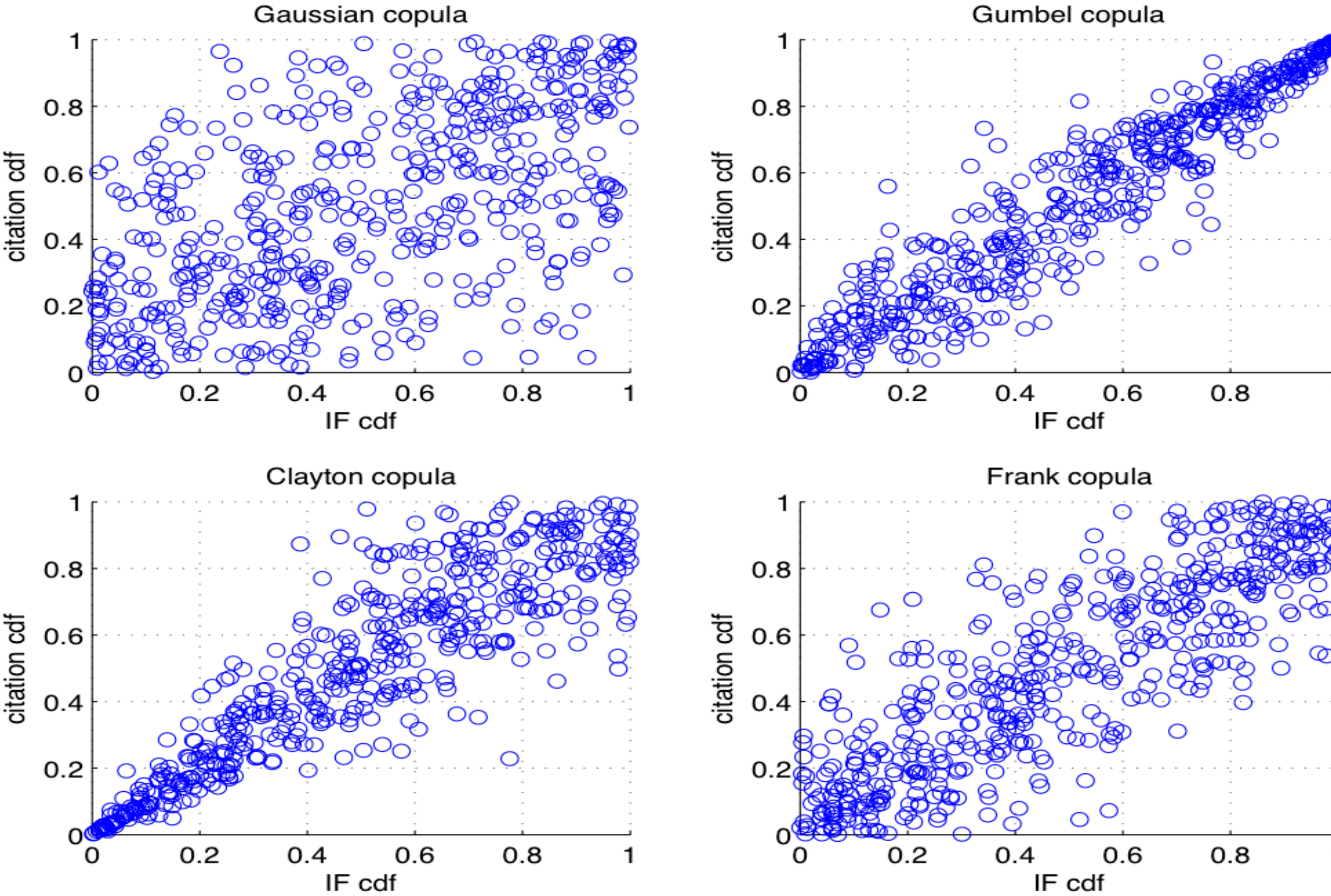
$$\rho_s = \frac{1}{3}\theta$$

Empirical Rank Correlation

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$$

* Some notes are from Climate Data Analysis course

Copulas and Dependence



From Google

- Goodness-of-fit test:

  1) Graphical comparison: Theoretical vs. Empirical

  2) Compute Maximum log-likelihood

  3) p-value test

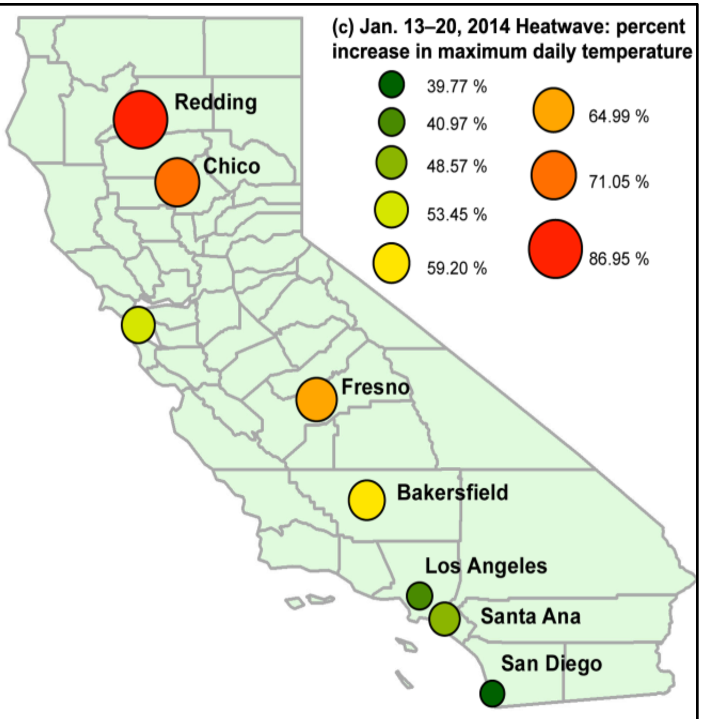
- Package: copula

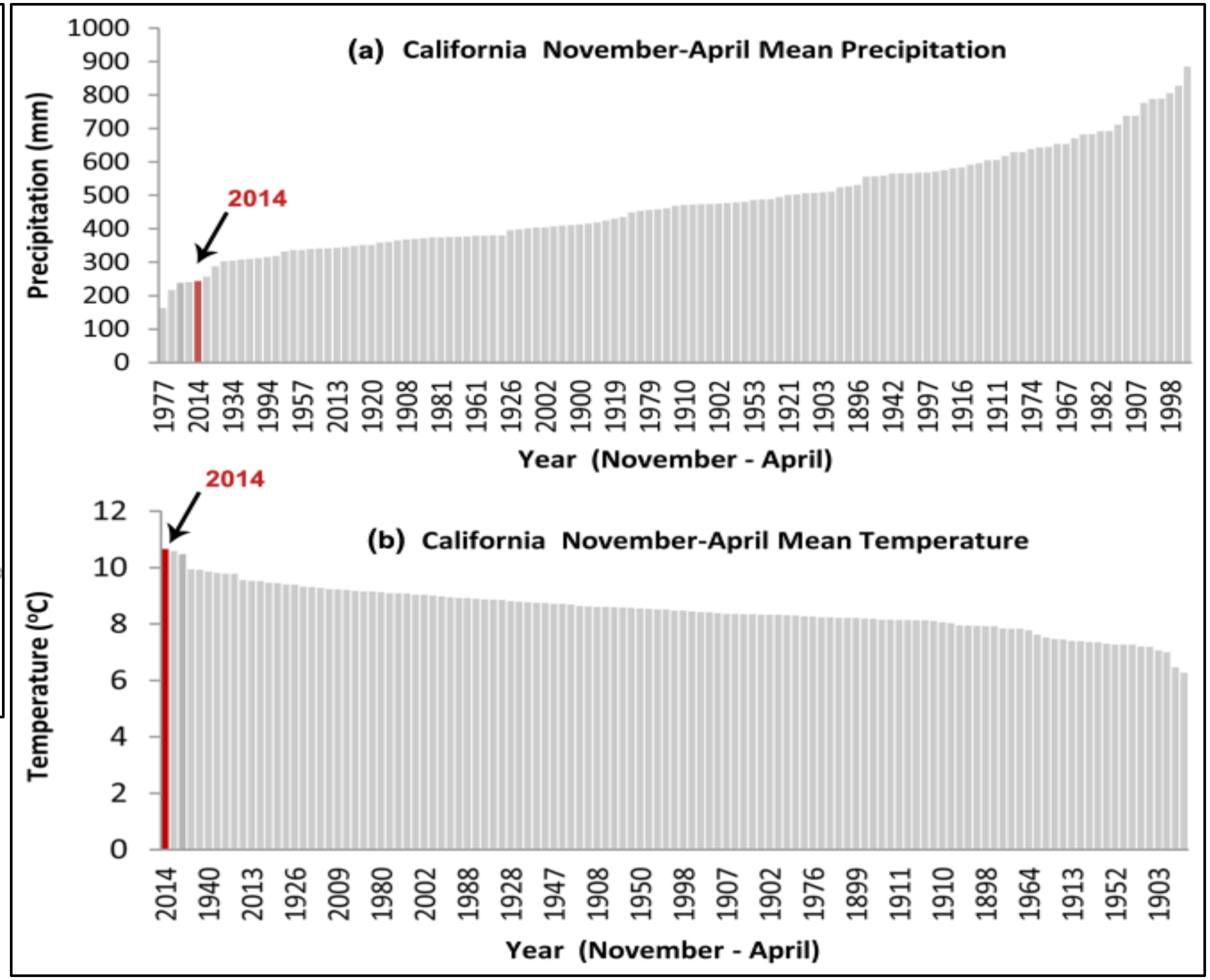
  Author: Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan

R code:

```
setwd("C:/Users/HRG/Desktop")
library(copula)
da90<-read.delim("marxy.txt",header=FALSE, sep="\t", dec=".")
names(da90)<- c("Prcp", "Temp")
attach(da90)
u<-pobs(da90[,1:2])

fc<-frankCopula(dim=2)
ffc<-fitCopula(fc,u)

nc<-normalCopula(dim=2, dispstr="un")
fnc<-fitCopula(nc,u)

fgc@loglik; fcc@loglik; ffc@loglik; fnc@loglik; ftc@loglik; fpc@loglik; fjc@loglik;
```

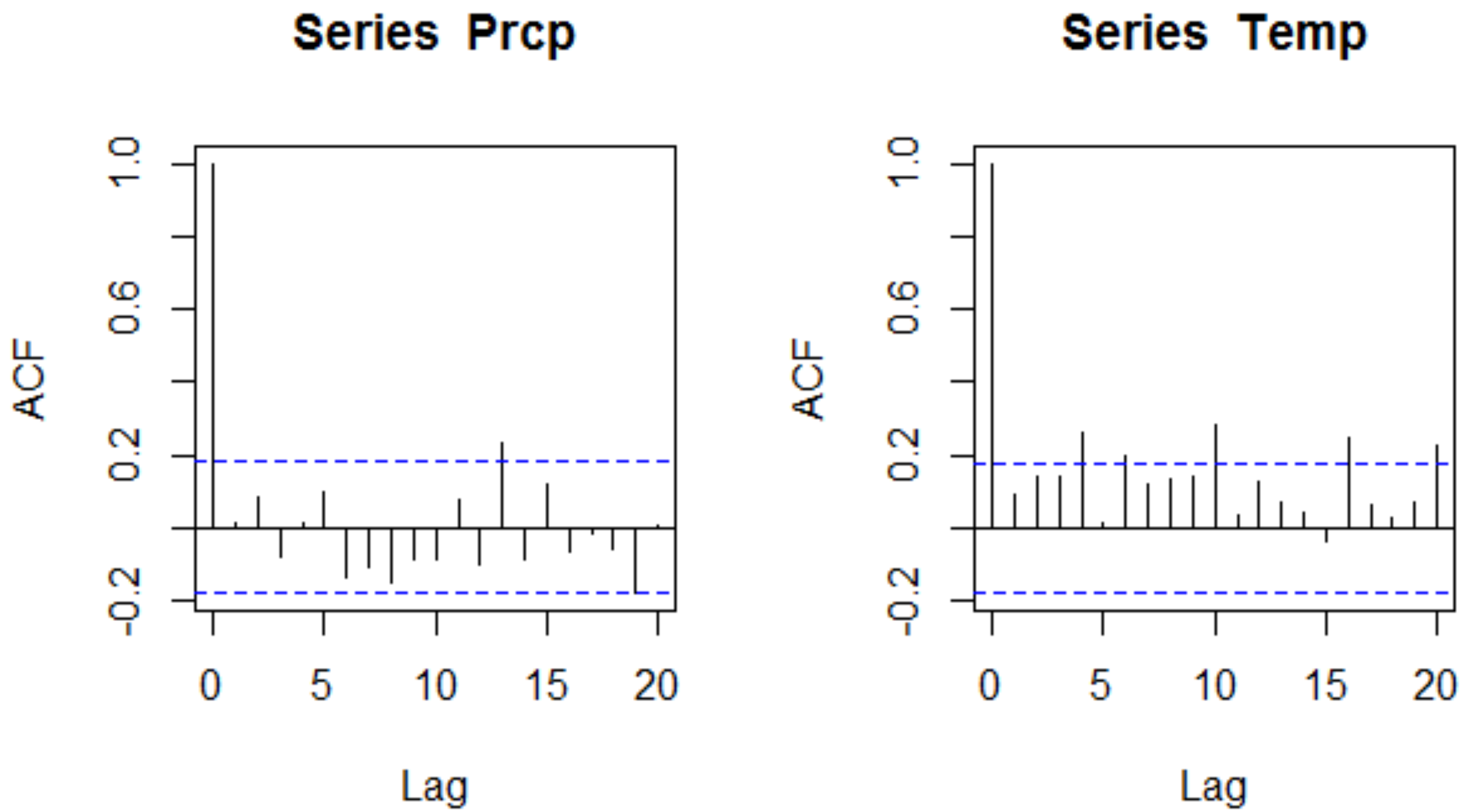
Example 1: CA 2014 temperature and precipitation





- AghaKouchak A., Cheng L., Mazdiyasni O., Farahmand A., 2014, Global Warming and Changes in Risk of Concurrent Climate Extremes: Insights from the 2014 California Drought, ***Geophysical Research Letters***

Autocorrelation:

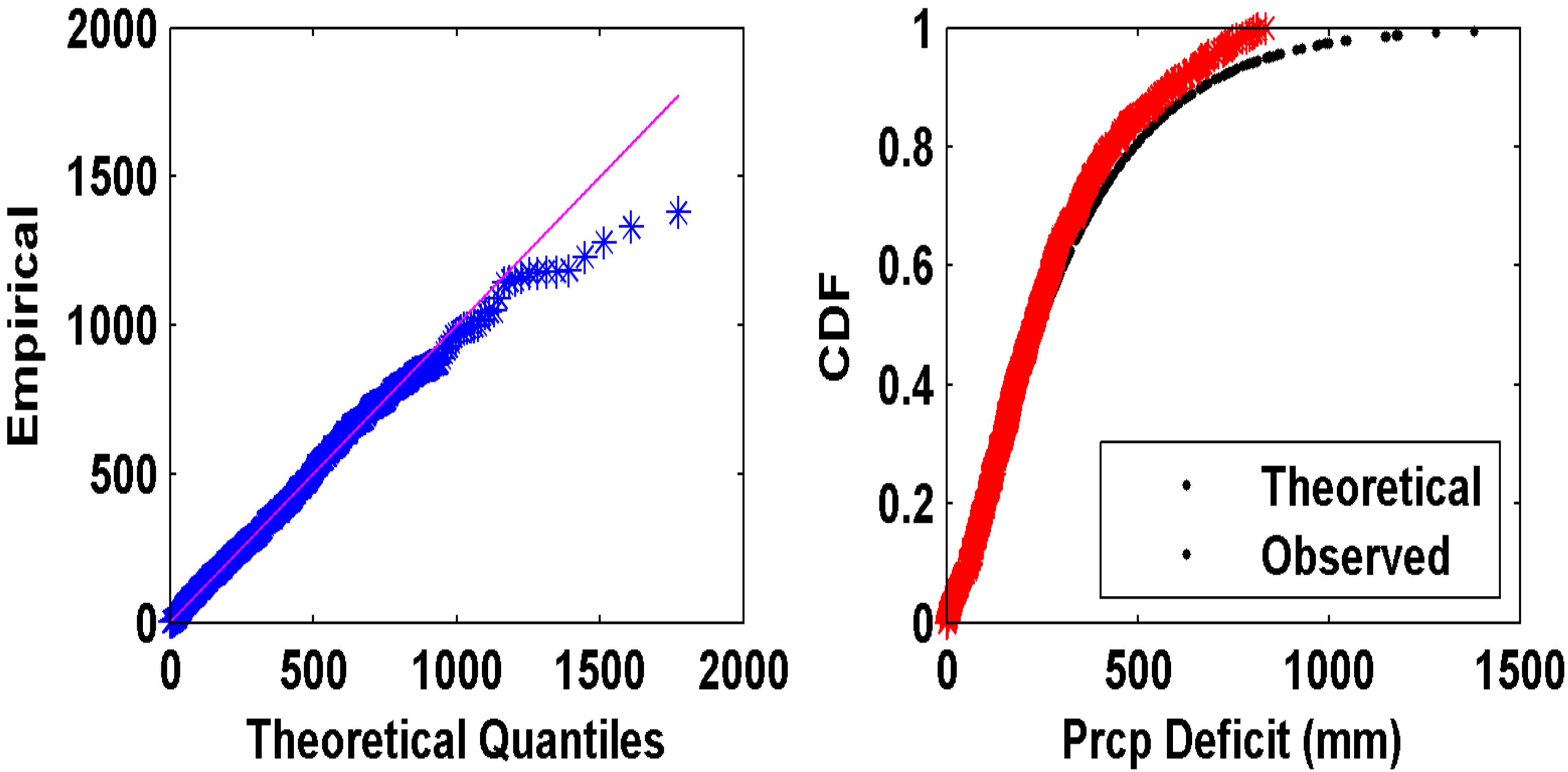


Applications using Copulas

Goodness-of-fit test:

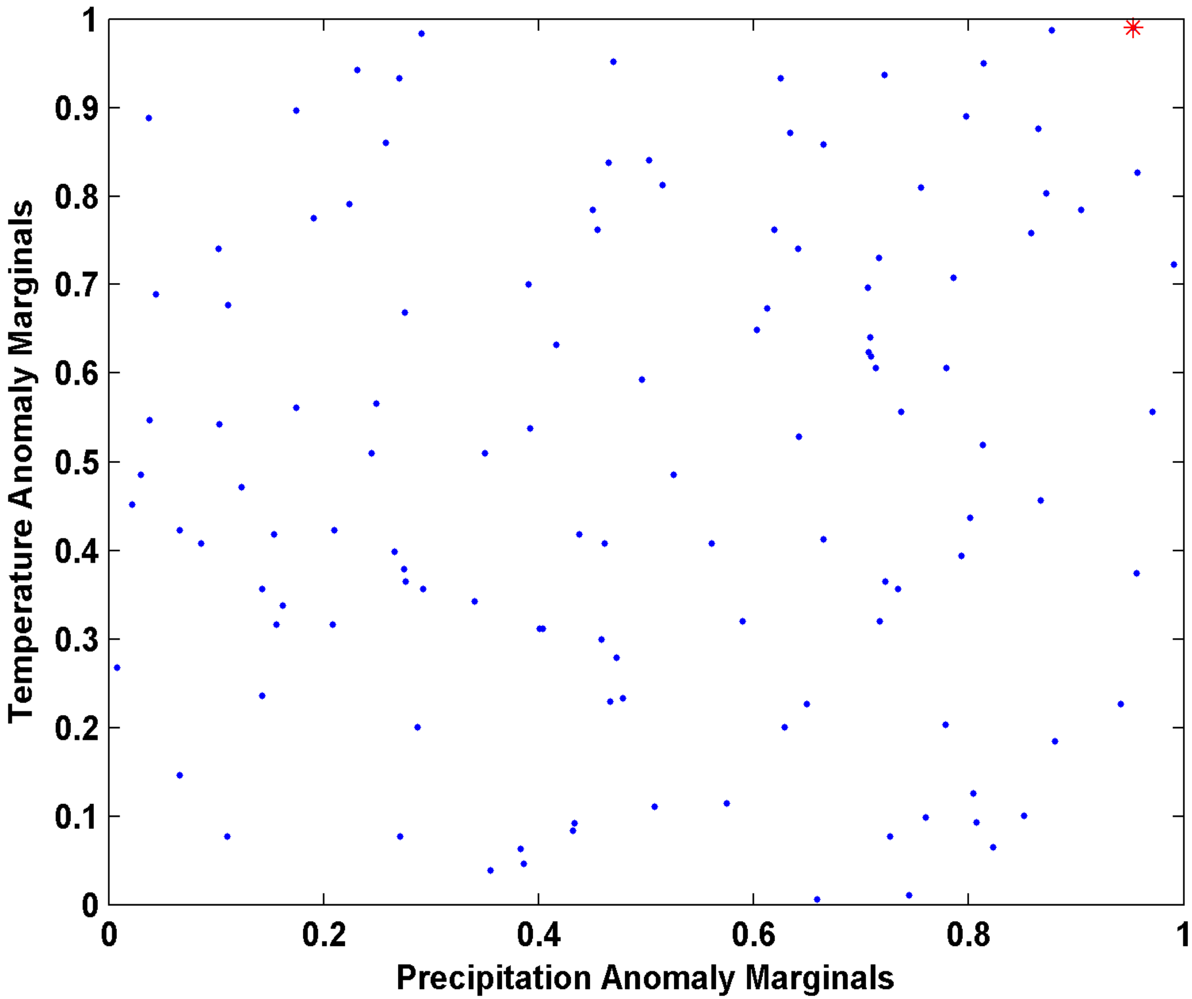
RMSE	GEV	Lognorm	GP	Gamma	Exp	Well
Temp	0.085	0.099	0.093	0.105	0.125	0.108
Pcpn	0.014	0.048	0.029	0.012	0.052	0.012

Goodness-of-fit test:



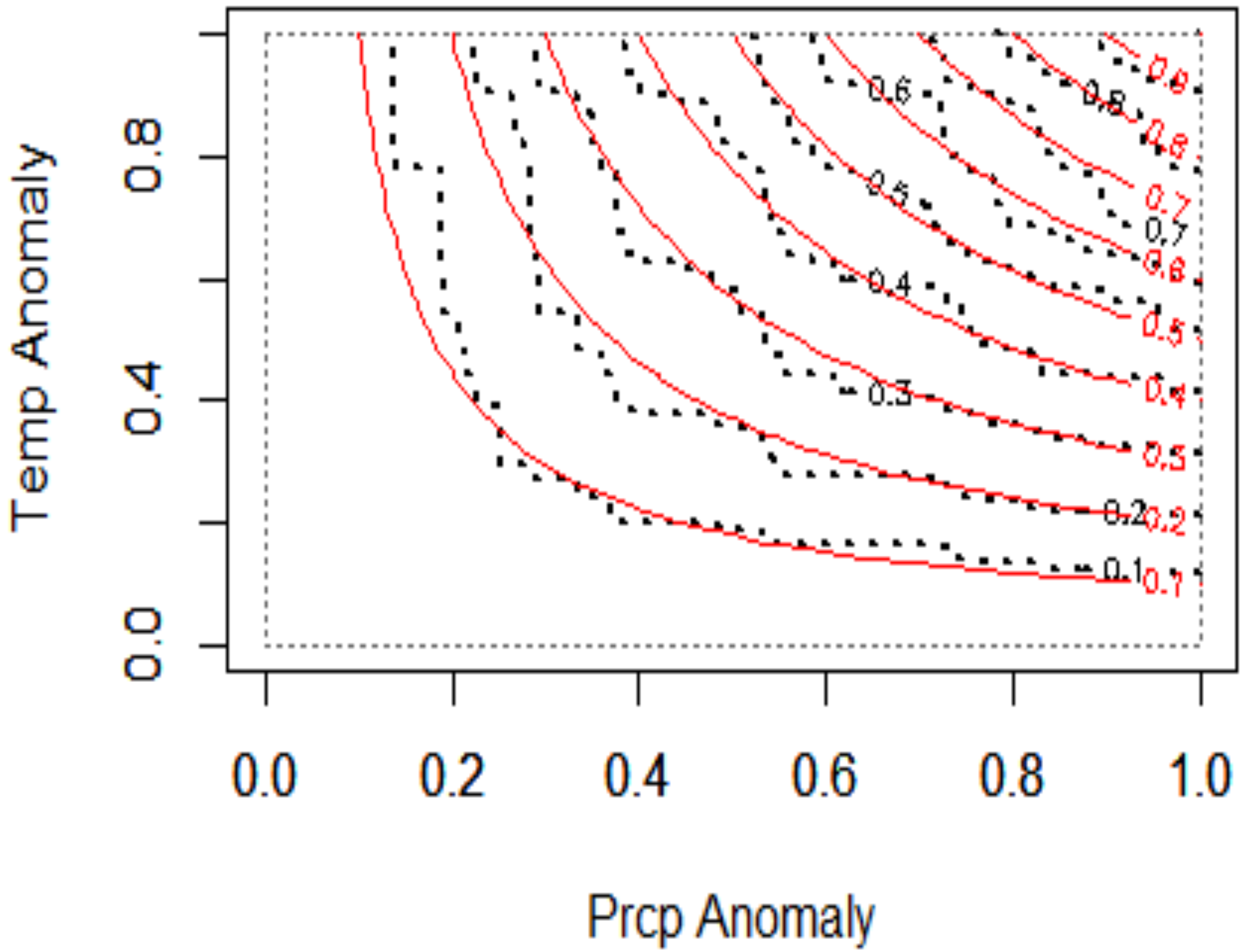
Applications using Copulas

Marginal:

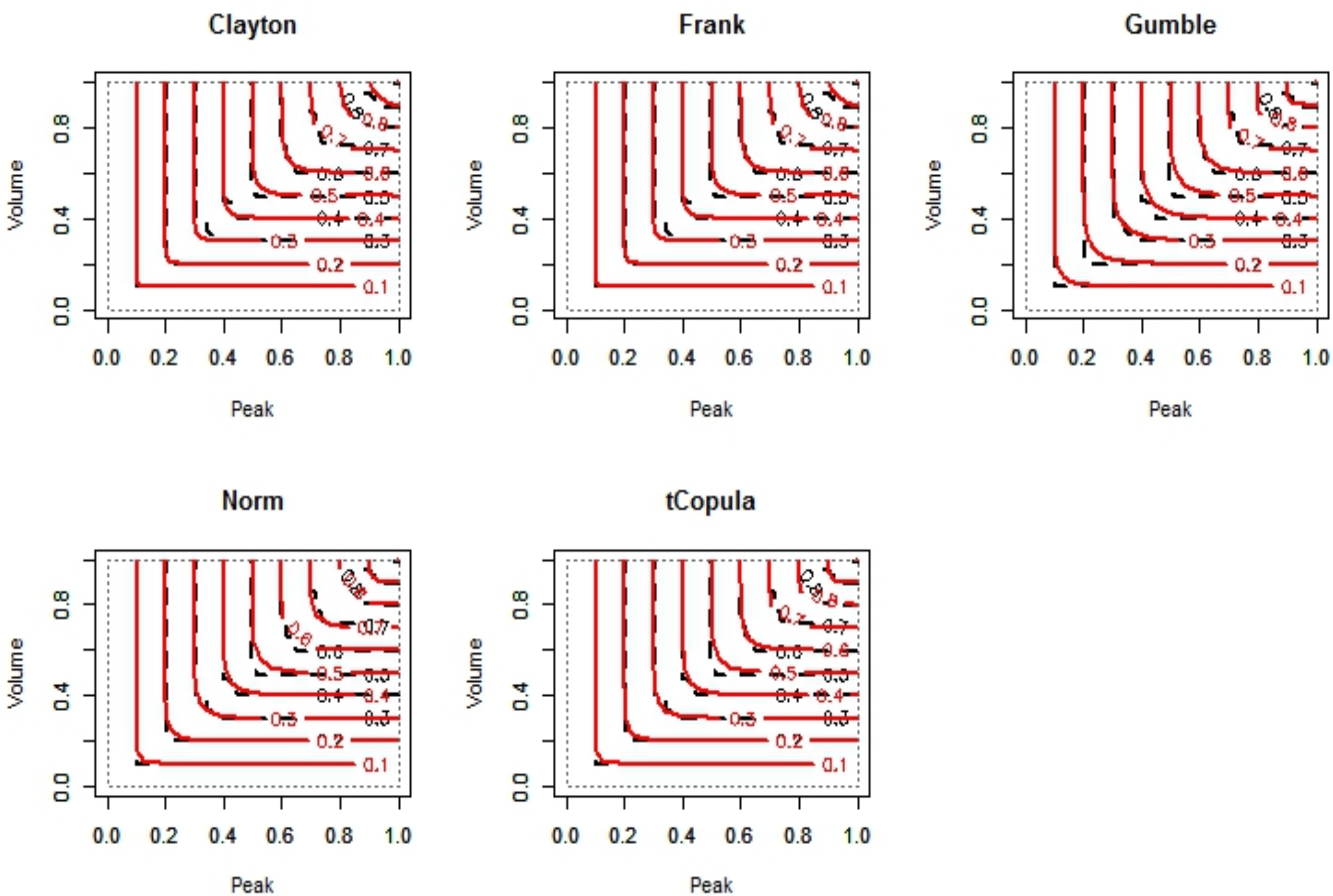


Applications using Copulas

Goodness-of-fit test:
t copula fit (red curve) against empirical (black dashed lines)



Goodness-of-fit test:



Applications using Copulas

Goodness-of-fit test:

	Parameter	loglikelihood	p-value
Gumbel	NA	NA	NA
Clayton	0.198	1.53	0.6479
Frank	0.469	0.36	0.1833
Normal	0.096	0.42	0.2672
t	0.096	0.42	0.2772

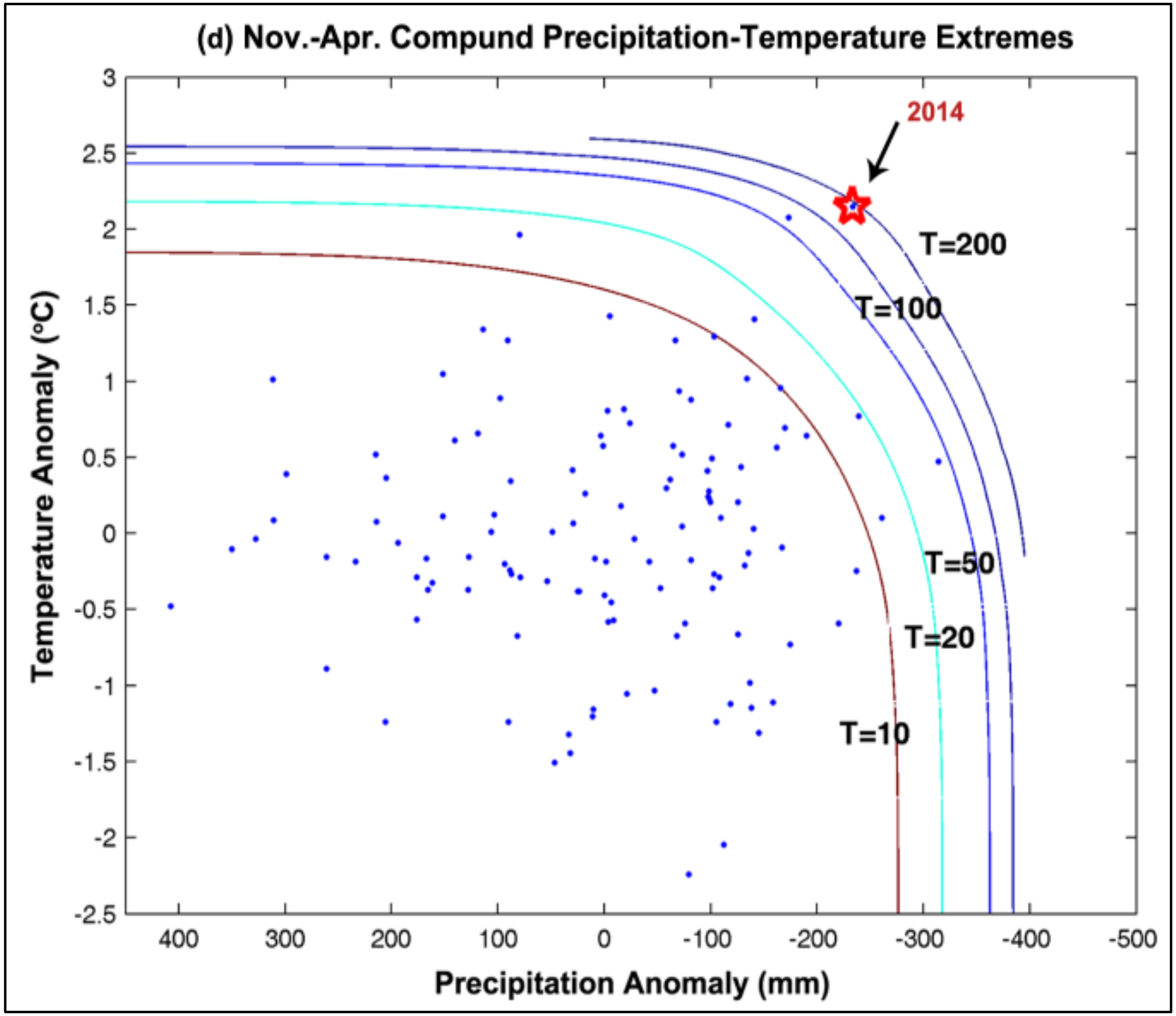
Applications using Copulas

Goodness-of-fit test:

	Parameter	loglikelihood	p-value
Gumbel	NA	NA	NA
Clayton	0.198	1.53	0.6479
Frank	0.469	0.36	0.1833
Normal	0.096	0.42	0.2672
t	0.096	0.42	0.2772

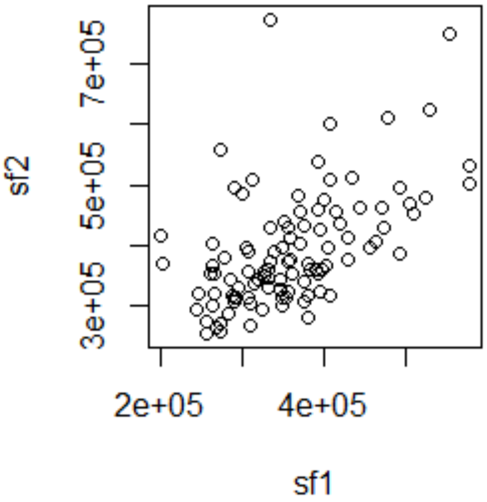
Applications using Copulas

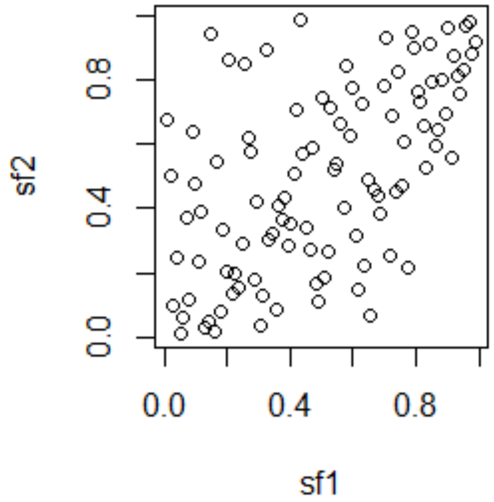
Example 1: **Bivariate Return Period**: joint analysis of temp and pcpn

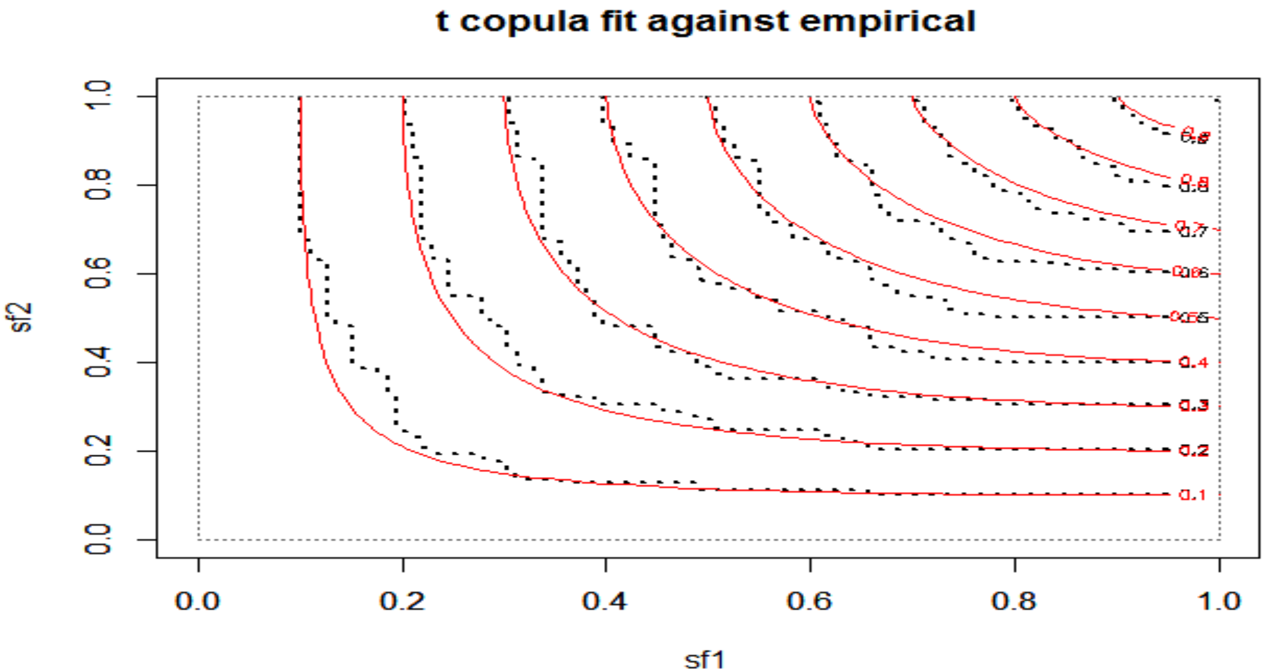


Applications using Copulas

Streamflow data:



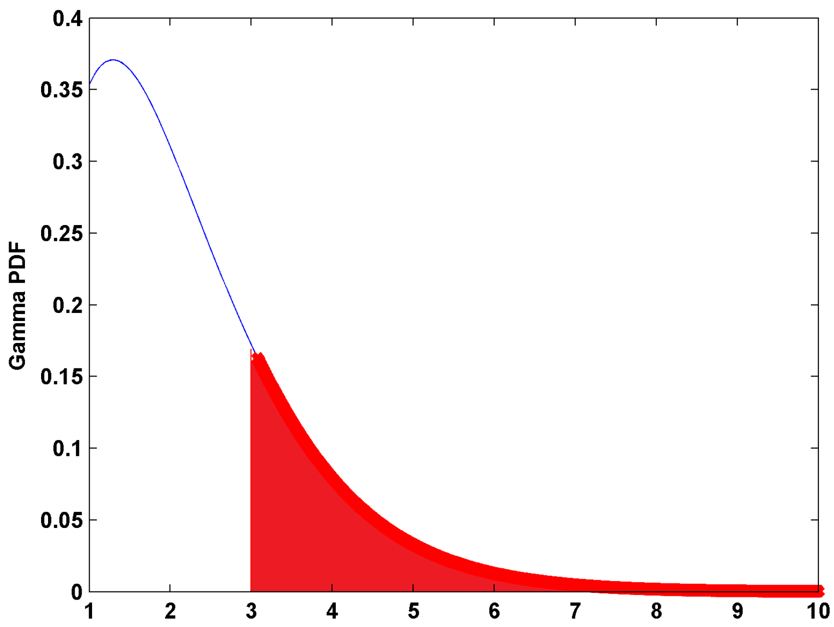




What is the return period T of the univariate of CA 2014 precipitation?

$$T = \frac{m}{1-p}$$

where $m > 0$ is the average interarrival time of two consecutive events; p is the non-exceedance probability.



Applications using Copulas

Example 2:

Bivariate Return Period: analysis of CA drought duration and severity

Applications using Copulas

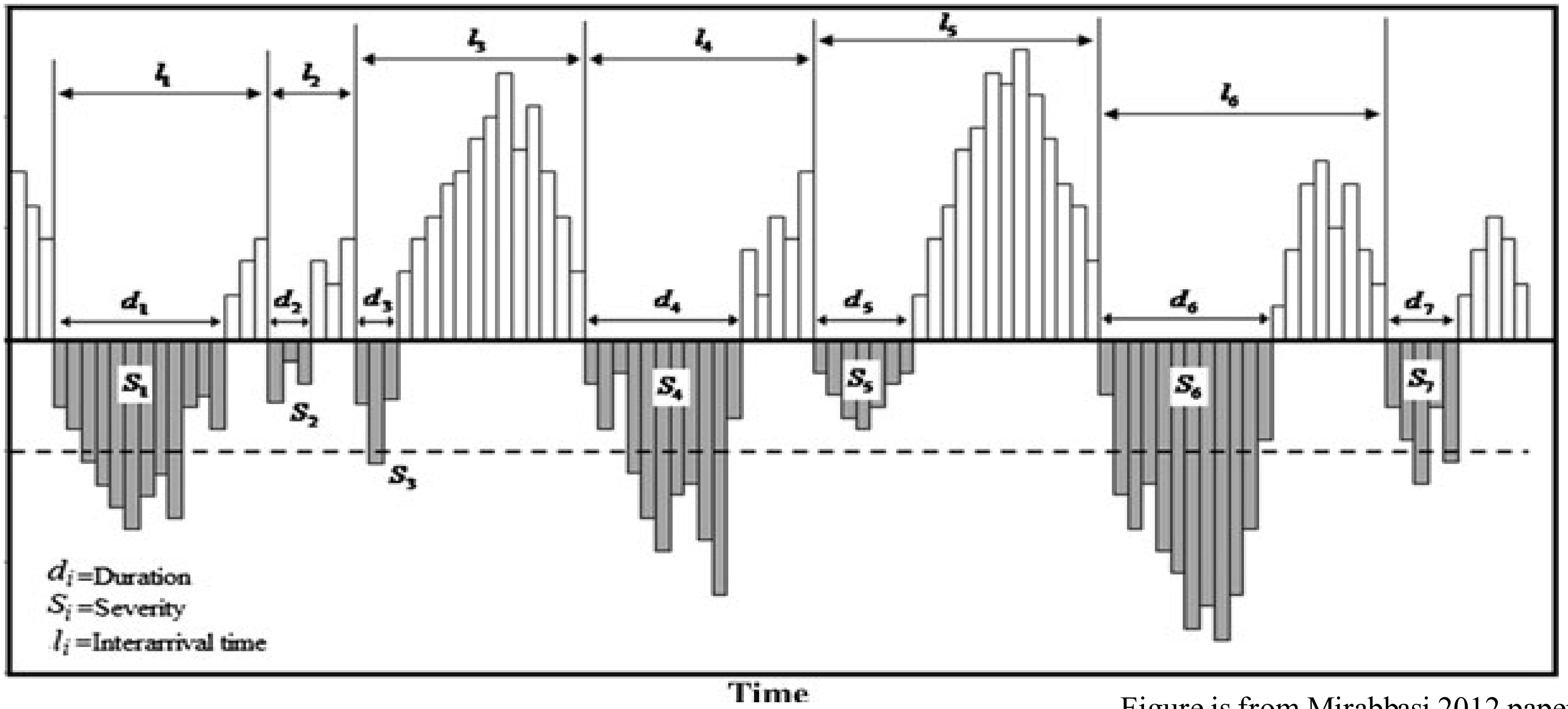


Figure is from Mirabbasi 2012 paper

Applications using Copulas

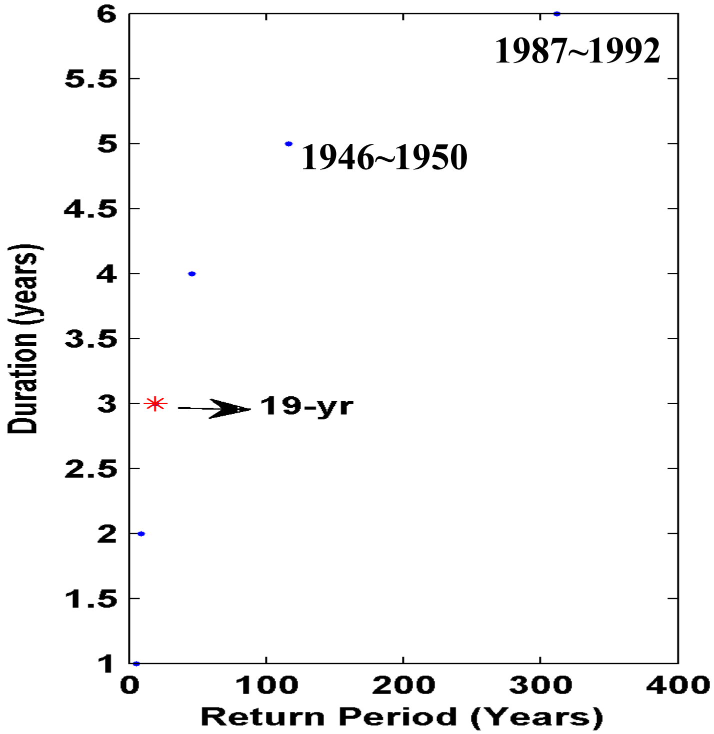
Example 2:

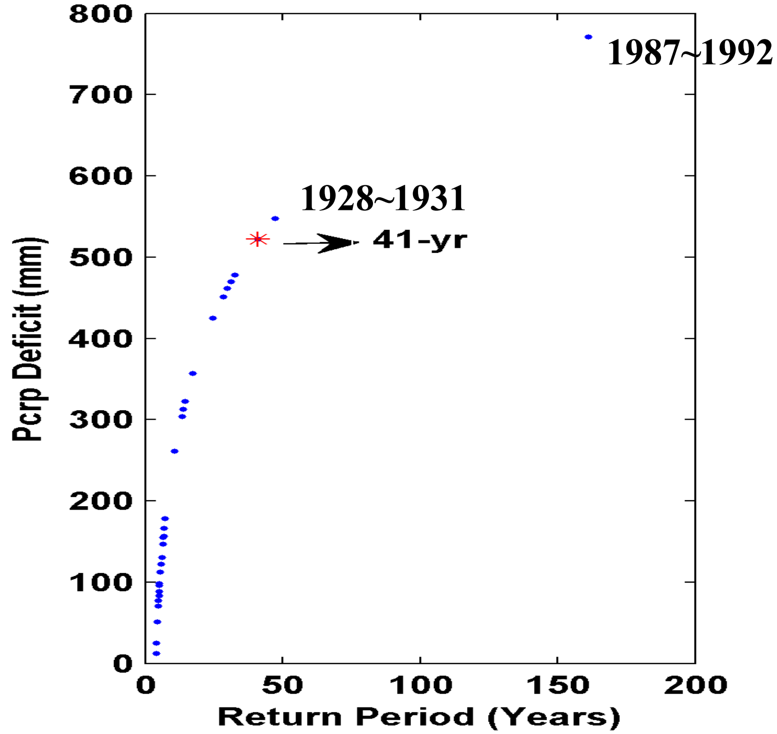
<u>Univariate</u>: the current CA drought duration is 3 years (ranked 7^{th})
the 3-year precipitation deficit is 522 mm (ranked 3^{rd})

Applications using Copulas

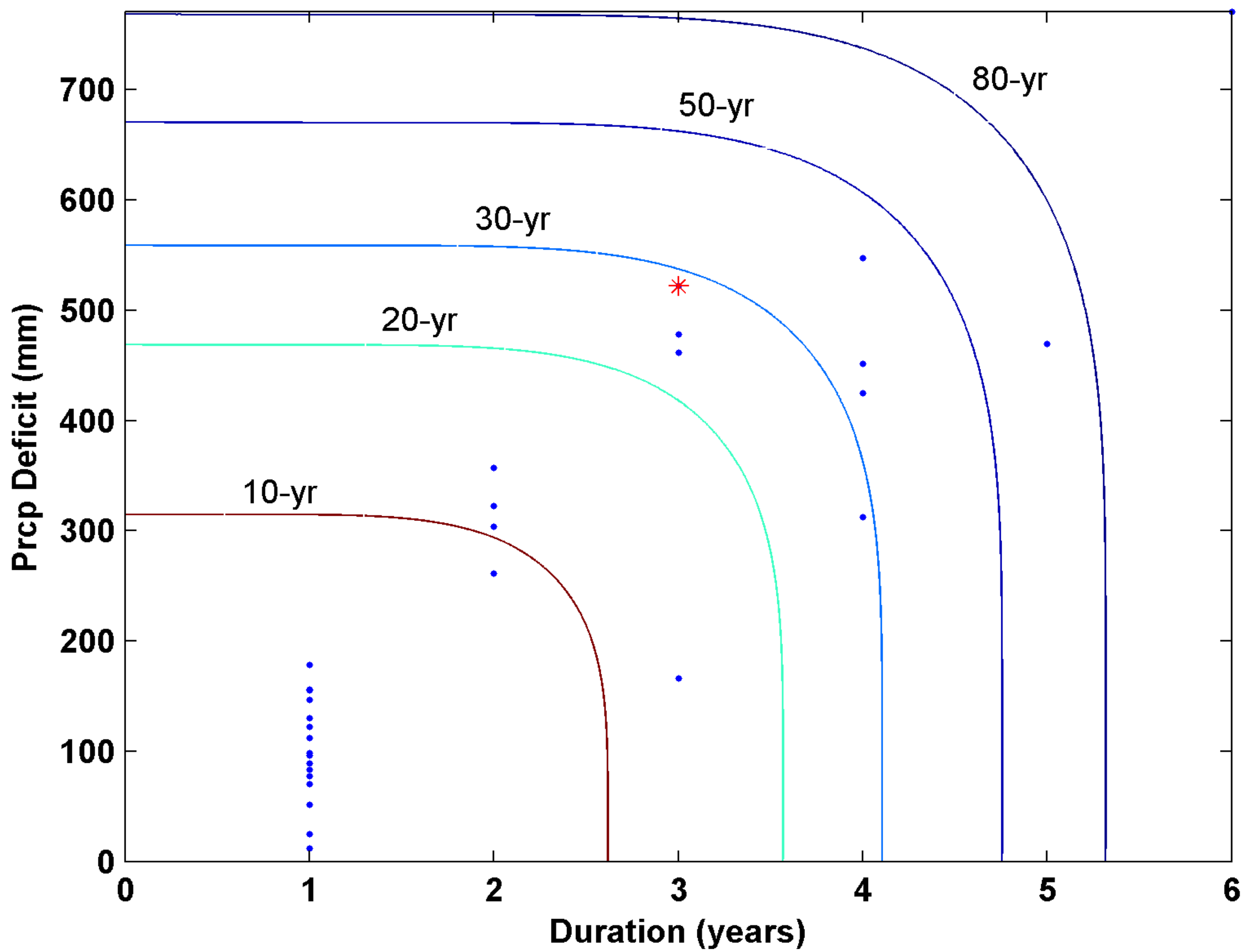
Example 2:

Univariate: the current CA drought duration is 3 years (ranked 7th)
the 3-year precipitation deficit is 522 mm (ranked 3rd)



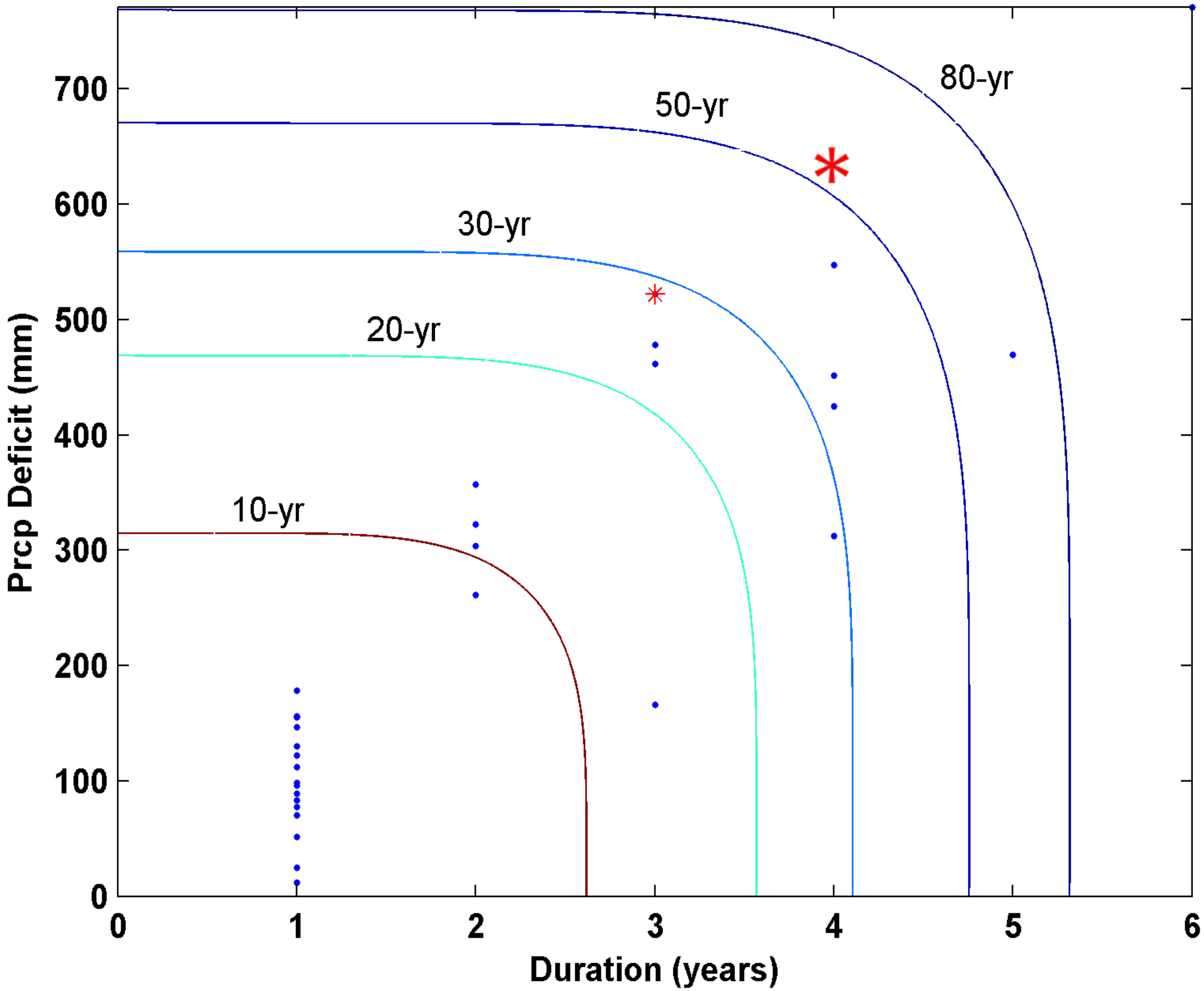


Example 2: **<u>Bivariate Return Period</u>**: Joint analysis of CA drought duration and severity



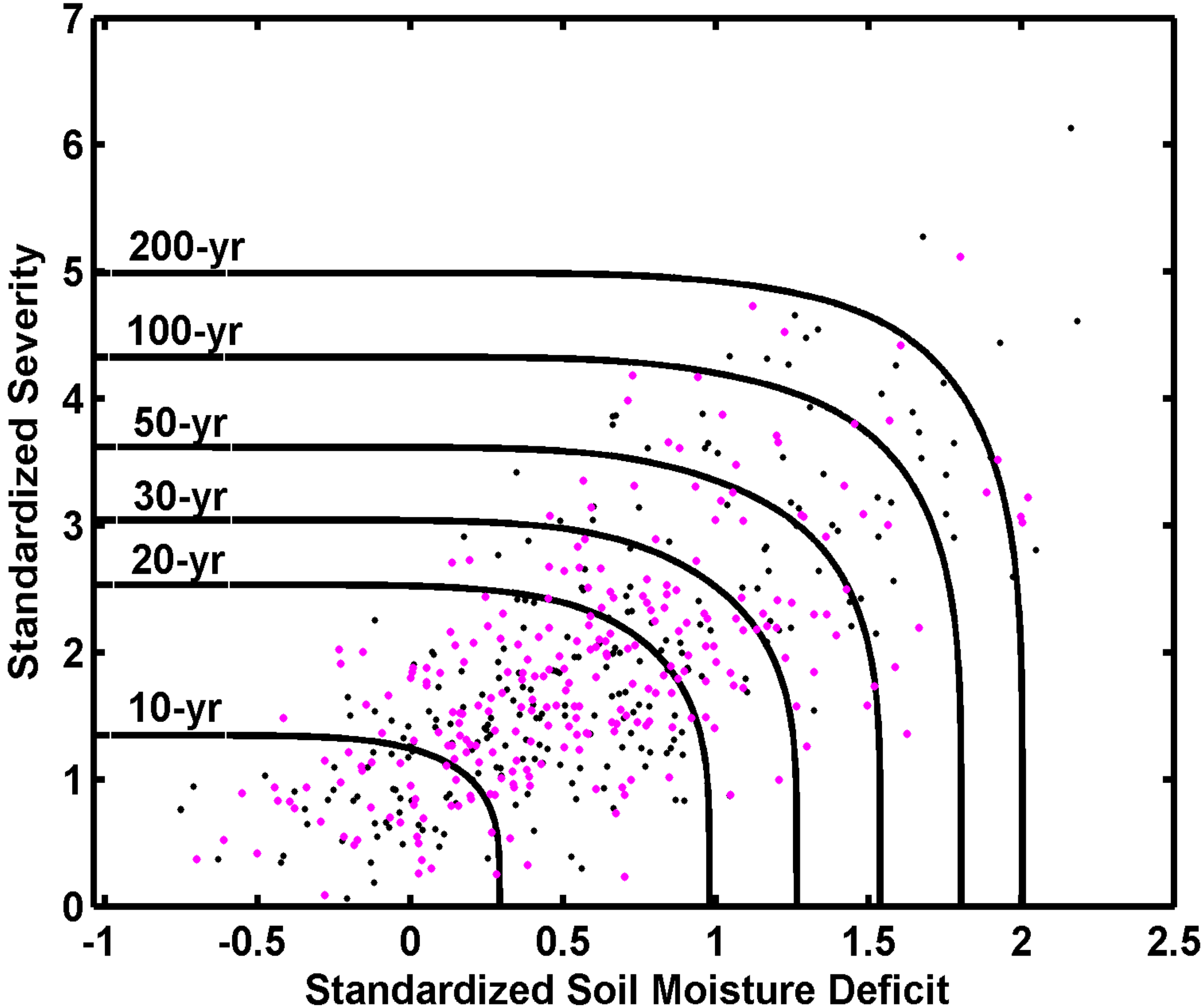
- Cheng L., Hoerling M., AghaKouchak A., Livneh B., Quan X., 2015, Current Effects of Human-induced Climate Change on California Drought, ***Journal of Climate*** (in press)

Example 2: **Bivariate Return Period**: Joint analysis of CA drought duration and severity



Applications using Copulas

Example 3: Precipitation and Soil Moisture (at 10cm) from preindustrial and industrial periods



More Applications

1. Medicine: Estimate the effect of an endogenous binary regressor (the "treatment") on a binary health outcome variable.

2. Finance: estimate the credit risk and the market risk.

3. Insurance

4. Biology

5. Health and environmental science

6. …

Research Topic

1. Nonstationarity of the dependence structure (change-point)

2. Conditional predictability (ungauged point or time step)

3. Spatial dependence:

   e.g. 10 out of 100 stations get flooding in a watershed. In a changing climate, the number of flooding stations increases to15.
   What are the plausible reasons?

4. Parameter uncertainty estimation (Bayesian inference)