Multivariate Frequency Analysis of Extreme Events and Applications to Uncertainty Quantification and Risk Assessment

Shih-Chieh Kao
GIST, CSED
Oak Ridge National Laboratory
kao@ornl.gov
http://www.ornl.gov/~5v1/

OAK RIDGE
National Laboratory

Hydrologic Extreme Events

Delphi, Indiana (Feb, 2008)
Flooding of Tippecanoe River

(AP Photo/Journal & Courier, Michael Heinz)

George Sparks Reservoir (Sept, 2007)
Lithia Springs, Georgia

(Barry Gillis, http://www.drought.unl.edu/gallery/2007/Georgia/Sparks1.htm)

- **Risk in hydraulic design - *Return Period***
- **Multivariate with spatio-temporal dependence structure**

Outline

- Background and motivation
- Correlation and dependence structure
- Copulas
- Application I: Extreme rainfall analysis
- Application II: Drought frequency analysis
- Application III: Climate extreme and impact
- Future works
- Summary and concluding remarks

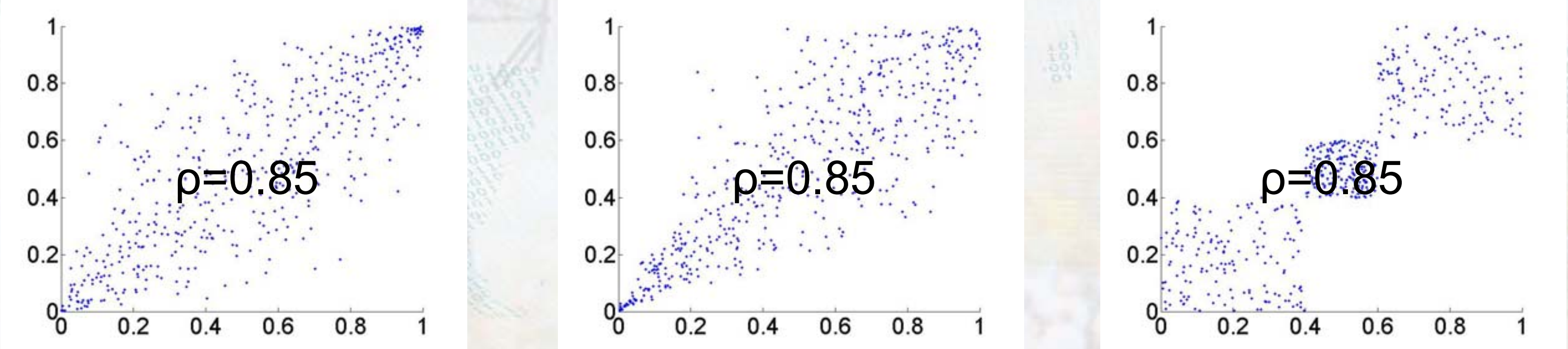
Background and Motivation

- **Uncertainty**
  - Central limit theorem and normal distribution
  - Sum of a sufficiently large number of independent random variables
- **Risk**
  - (probability of an event) * (losses)
- **How to compute the probability when variables are**
  - multidimensional and non-Gaussian
  - mixture of discrete and continuous variables
  - with complicated dependence structure
- **Need a flexible algorithm in constructing multivariate joint distribution**
  - Focus on dependence in this study

Correlation and Dependence

- **Classification**
  - **Temporal: autoregression model (AR), Markov chain**
  - **Spatial: geostatistics (Kriging method)**
  - **Inter-variable: Bayesian approach**
- **Conventionally quantified by the Pearson's linear correlation coefficient ρ**

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



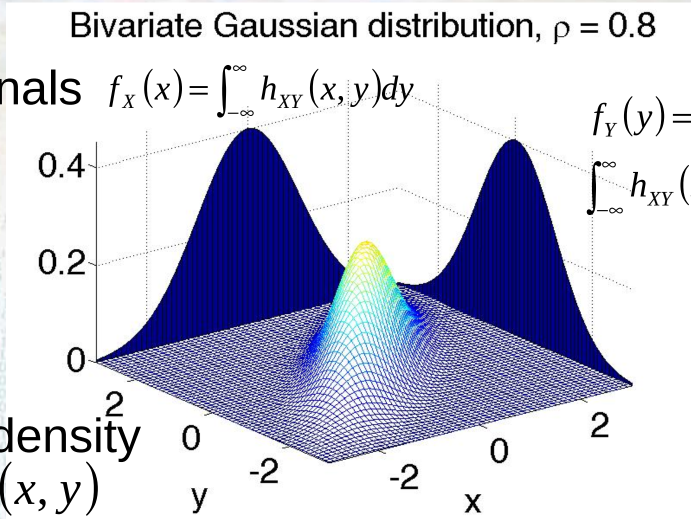
  - **Only valid for Gaussian (or elliptic) distributions**

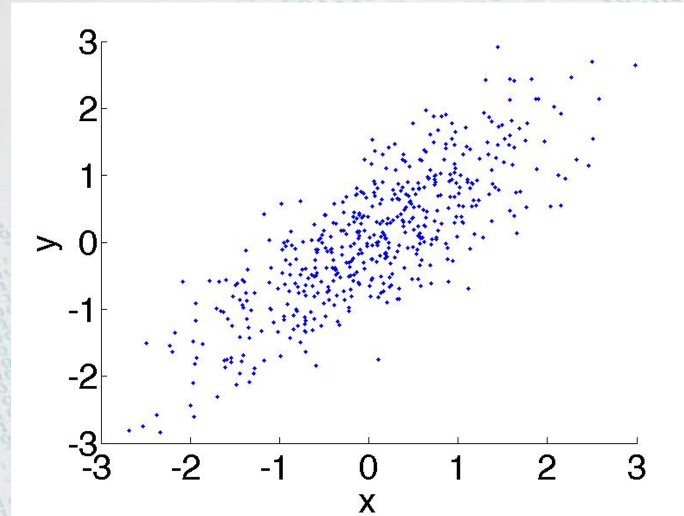
Example - Bivariate Distribution

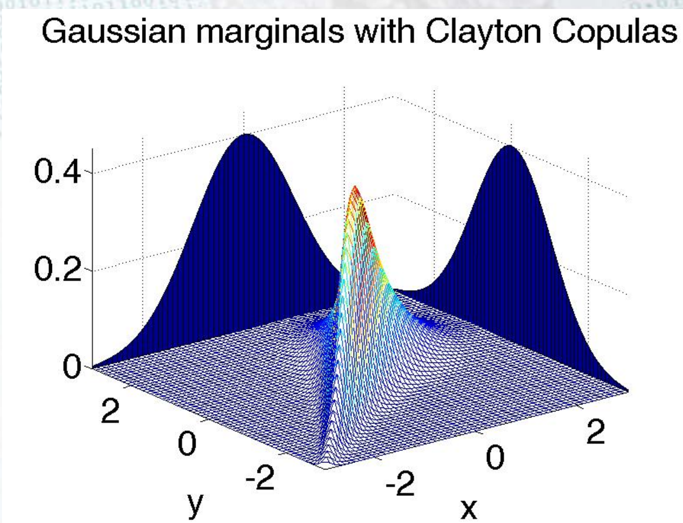
Marginals $f_X(x) = \int_{-\infty}^{\infty} h_{XY}(x,y)\,dy$

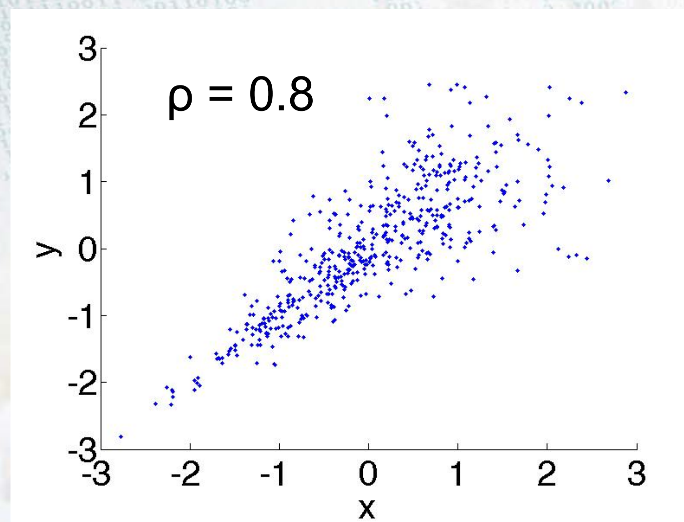
$f_Y(y) = \int_{-\infty}^{\infty} h_{XY}(x,y)\,dx$

Joint density $h_{XY}(x,y)$



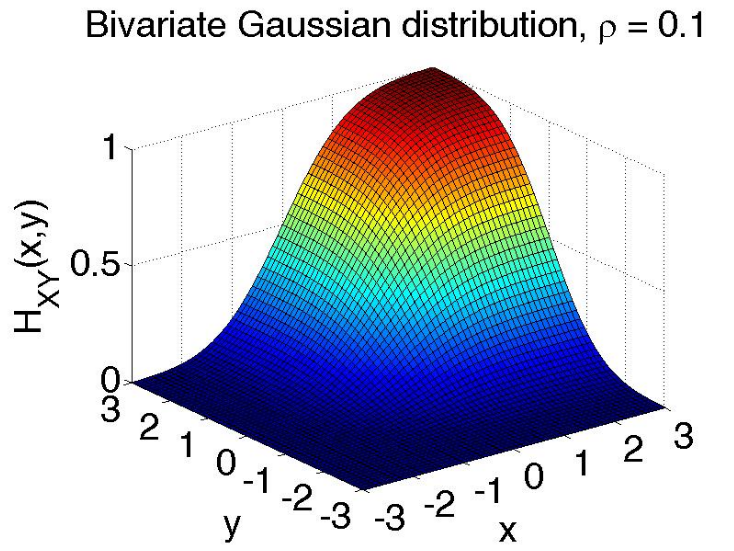


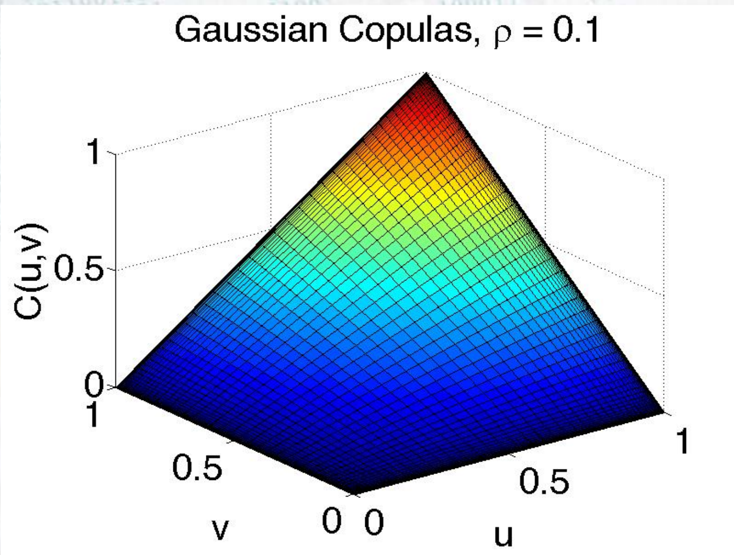




Copulas

- **Transformation of joint cumulative distribution**
  - **$H_{XY}(x,y) = C_{UV}(u,v)$ marginals: $u = F_X(x)$, $v = F_Y(y)$**
  - **Sklar (1959) proved that the transformation is *unique* for continuous r.v.s**
- **Use copulas to construct joint distributions**
  - **Marginal distributions => selecting suitable PDFs**
  - **Dependence structure => selecting suitable copulas**
  - **Together they form the joint distribution**

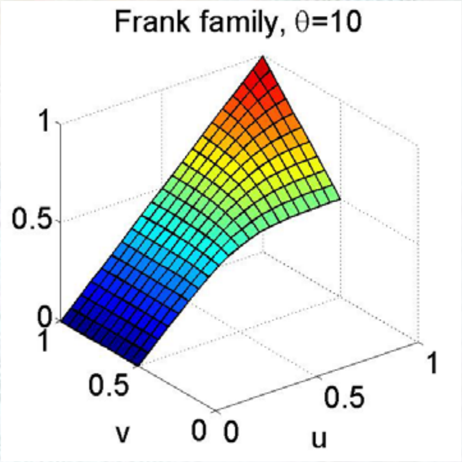


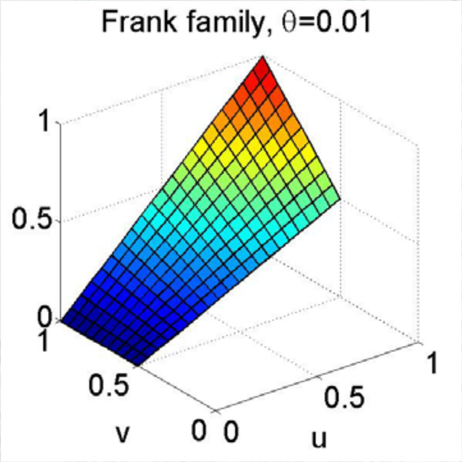


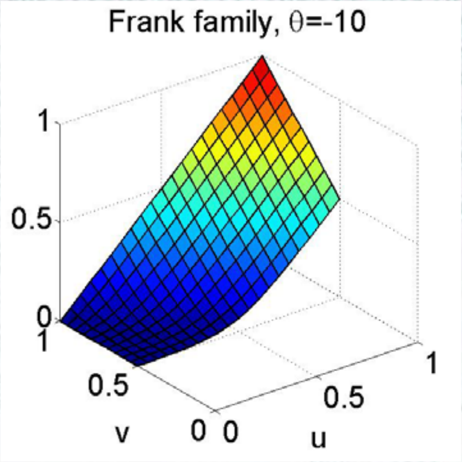
Example of Copulas – Frank Family

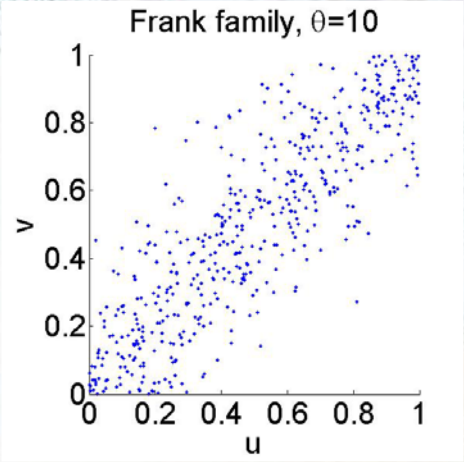
- **Frank family of Archimedean copulas**

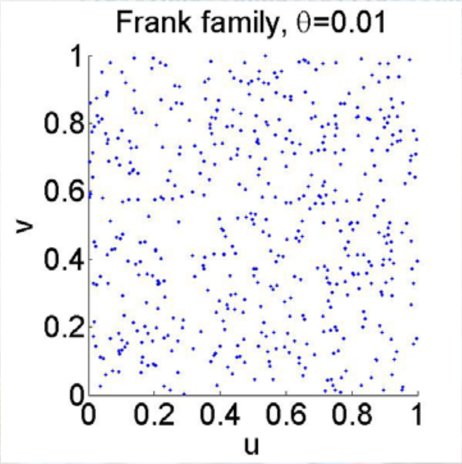
$$C_{Frank}(u,v) = -\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right)$$

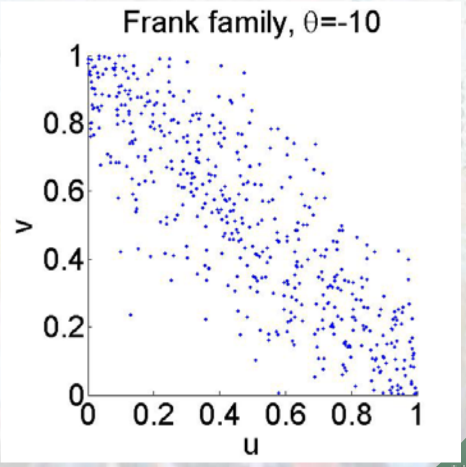






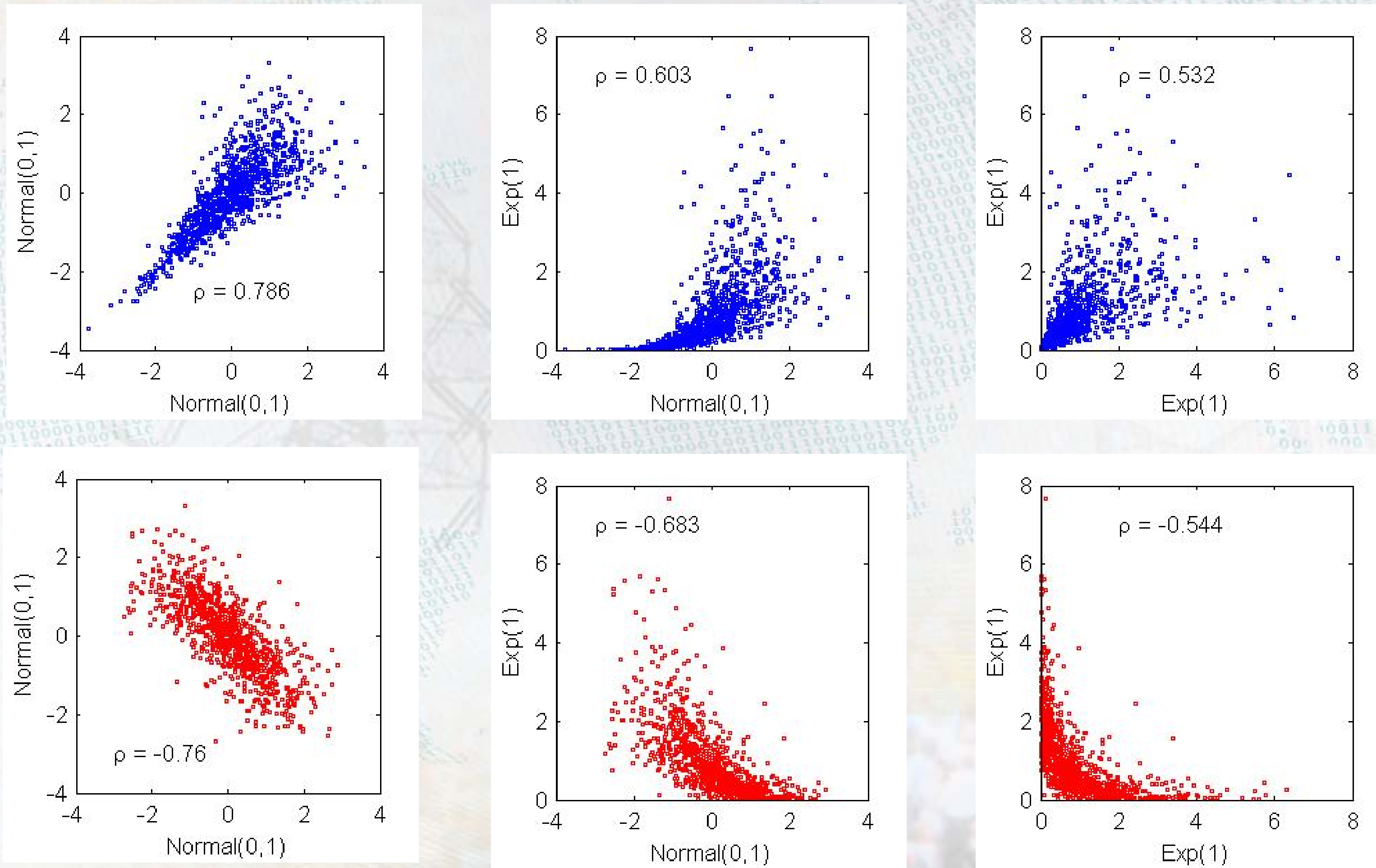






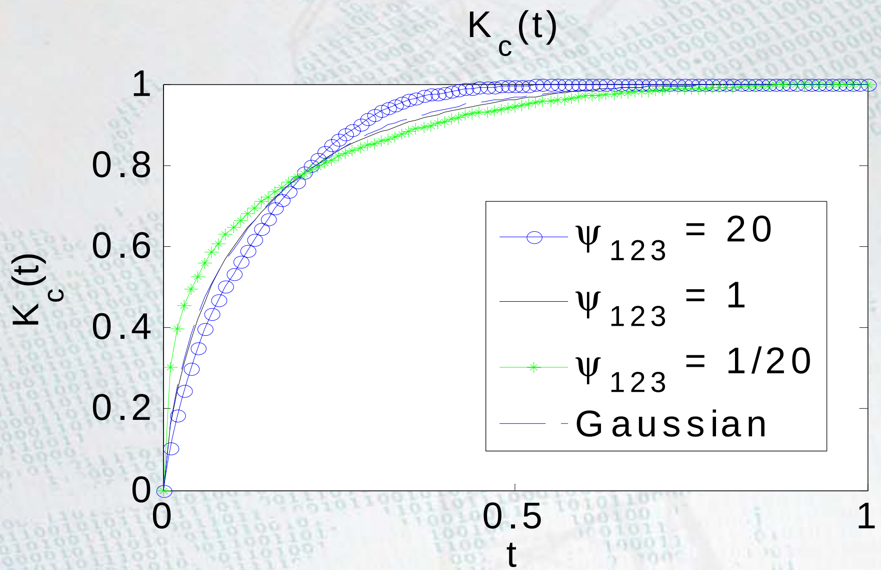
Monte Carlo Simulation

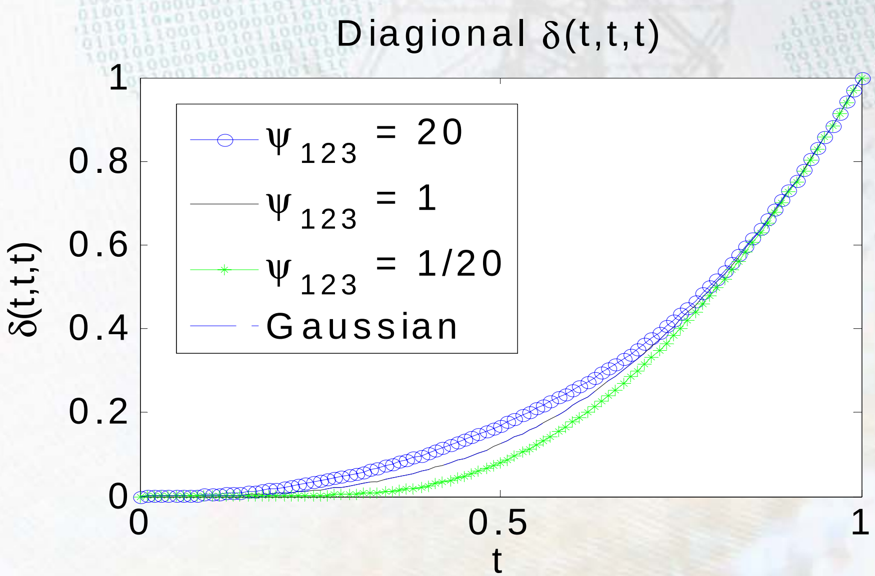
- **Clayton family (θ = 8.2), normal & exponential marginals**
- Frank family (θ = -8), normal & exponential marginals

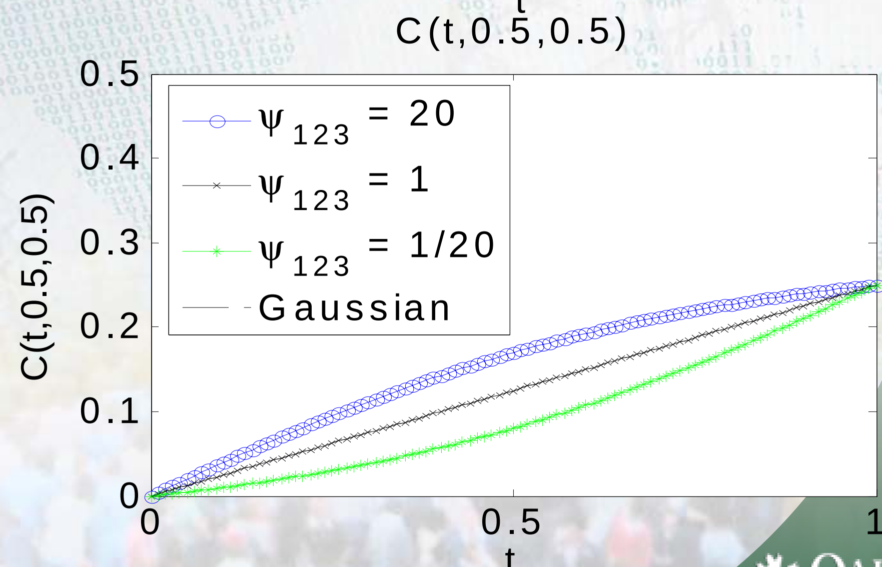


Beyond Bivariate Dependence

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



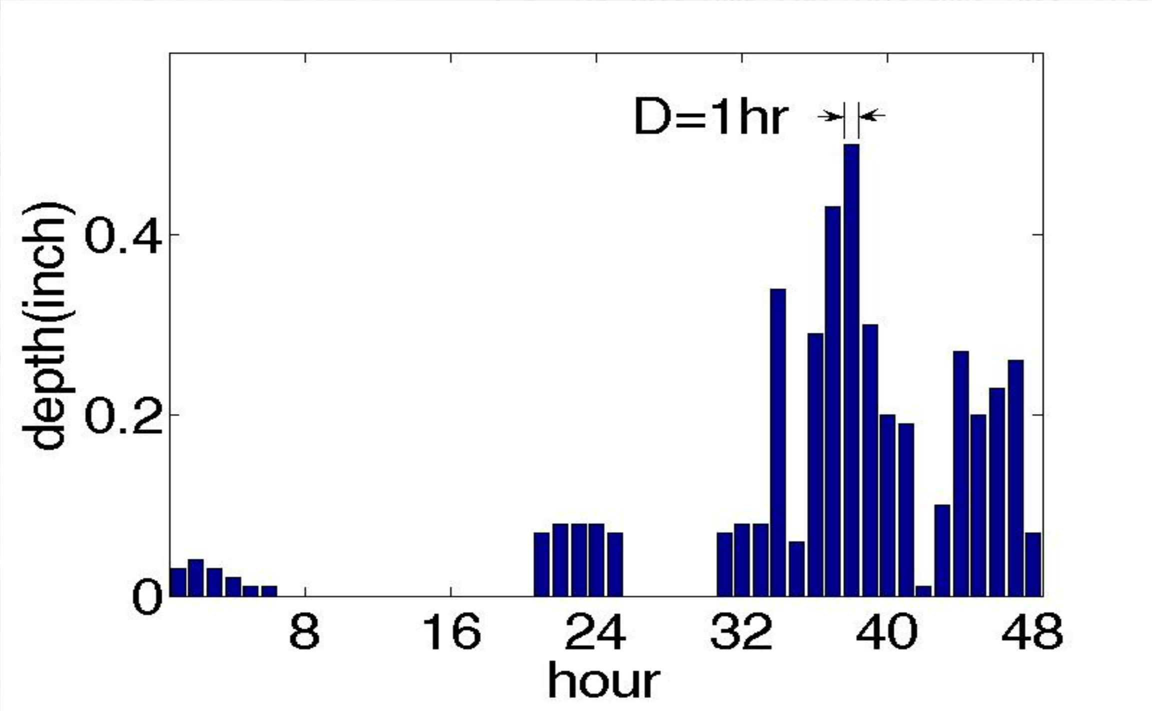




Extreme Rainfall - Univariate Approach

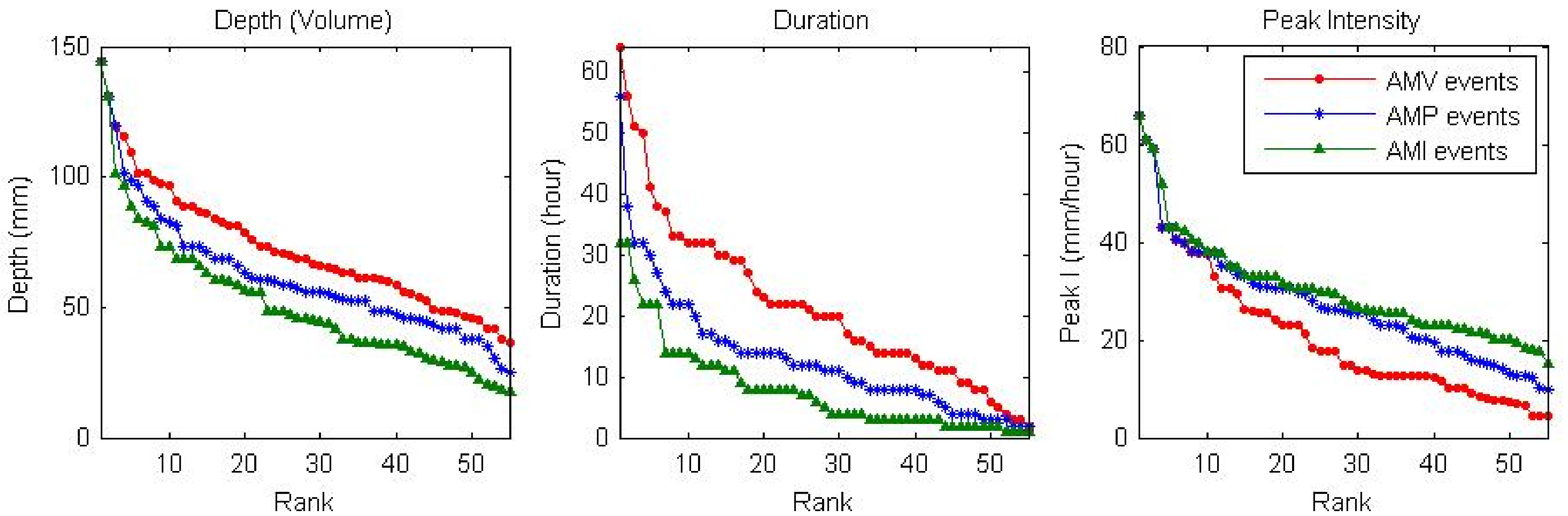
Geographic Information Science and Technology

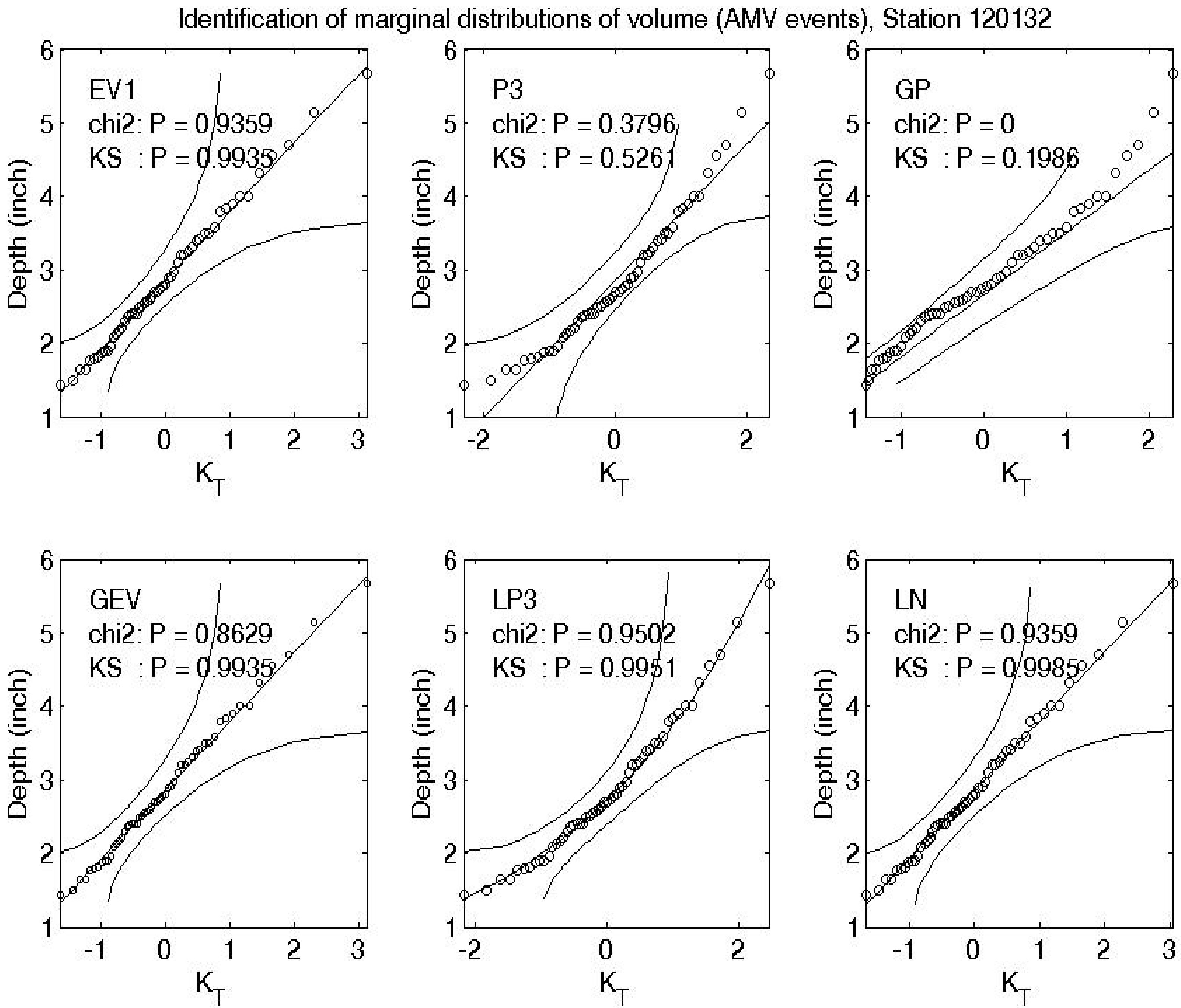
- **Selection of annual maximum precipitation**
  - ***Durations* are not the actual durations of rainfall events**
  - **Long-term maximum may cover multiple events**
  - **Short-term maximum encompasses only part of the extreme event**



Extreme Rainfall - Multivariate Approach

- **Definitions of Extreme Rainfall 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 between volume and peak intensity**
    - **Wide range of durations**




Identification of marginal distributions of volume (AMV events), Station 120132
EV1
chi2: P = 0.9359
KS : P = 0.9935
P3
chi2: P = 0.3796
KS : P = 0.5261
GP
chi2: P = 0
KS : P = 0.1986
GEV
chi2: P = 0.8629
KS : P = 0.9935
LP3
chi2: P = 0.9502
KS : P = 0.9951
LN
chi2: P = 0.9359
KS : P = 0.9985
Depth (inch)
K_T

Extreme Rainfall Frequency Analysis

Geographic Information Science and Technology

- **Bivariate distribution H_{PD}, H_{DI}, H_{PI}**
  - **Total precipitation (P), duration (D), and peak intensity (I)**
  - **Marginal: Extreme Value Type I (EV1), Log Normal (LN)**
  - **Dependence: Frank Family**

- **Applications**
  - **Estimate of depth for known duration**

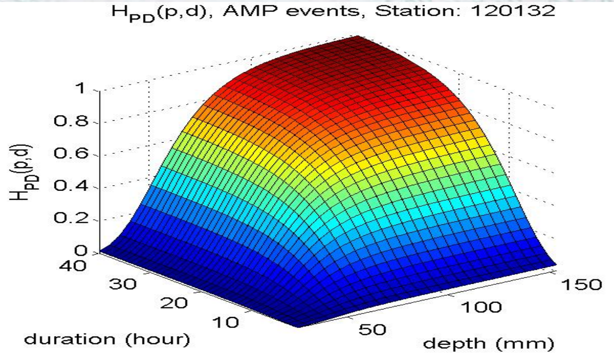
$$F_P\left(p_T \middle| d-1 < D \le d\right) = 1 - 1/T$$

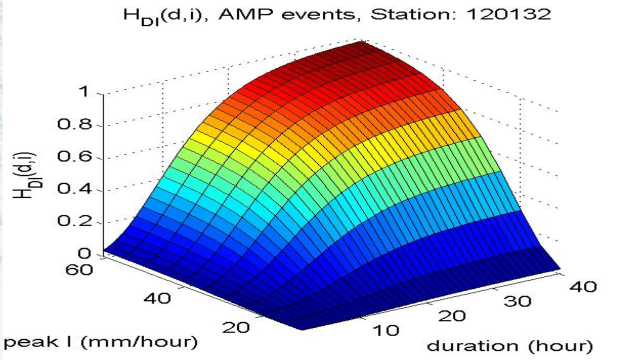
  - **Estimate of peak intensity for known duration**

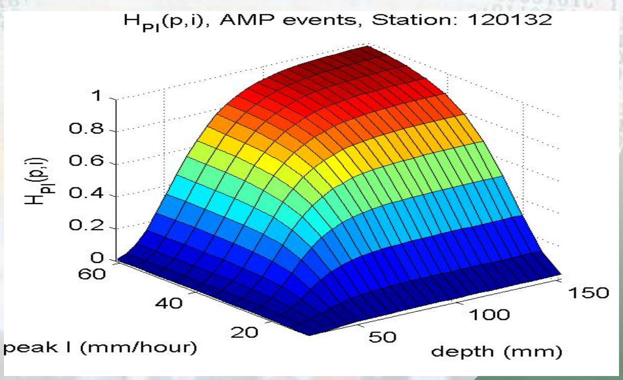
$$F_I\left(i_T \middle| d-1 < D \le d\right) = 1 - 1/T$$

  - **Estimate of peak intensity for known depth**

$$E\left[I \mid P > p\right]$$





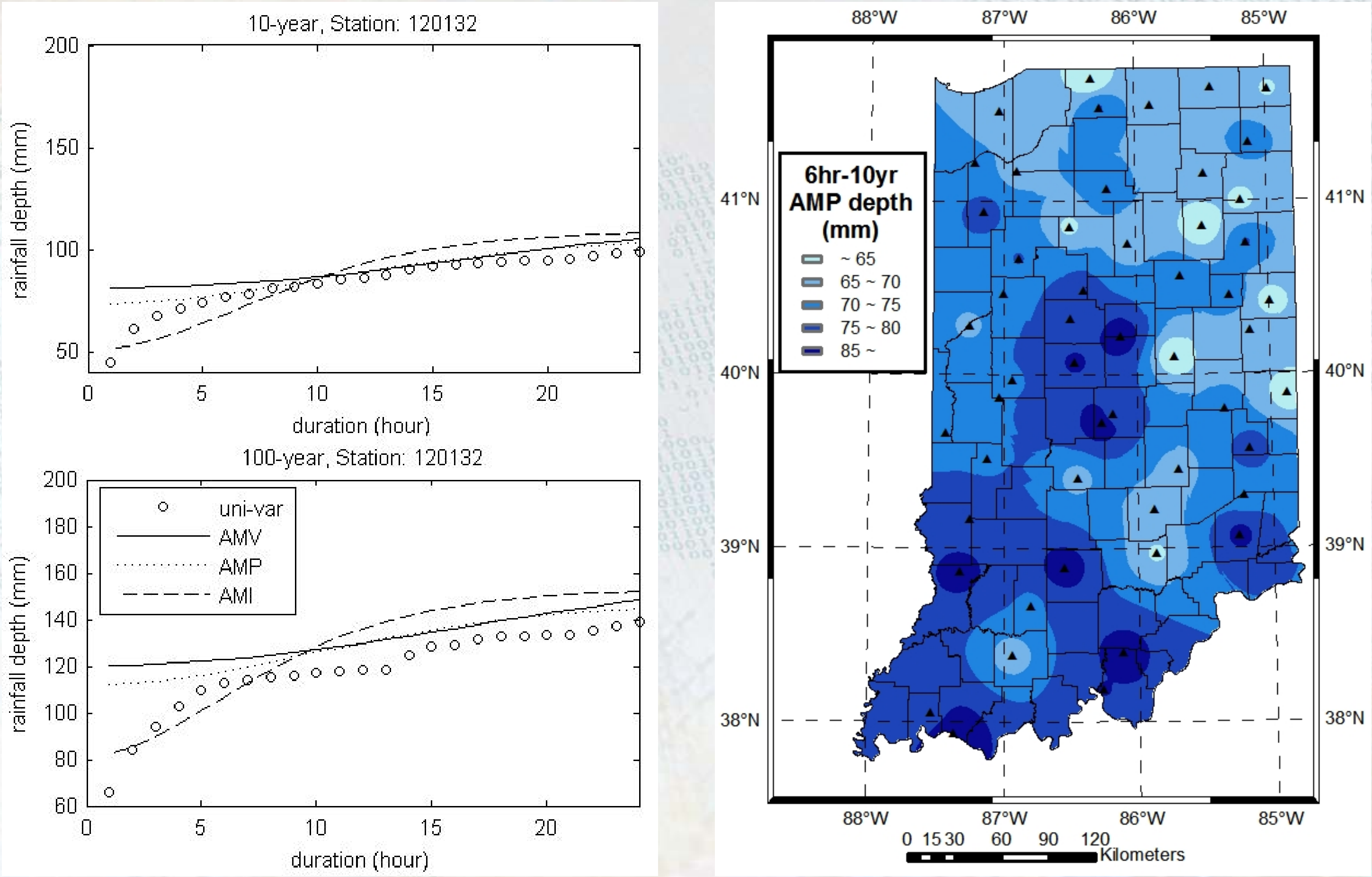


OAK RIDGE National Laboratory

Estimate of depth for known duration

Geographic Information Science and Technology

T-year depth p_T given duration d: $F_P(p_T|d-1<D<d)=1-1/T$

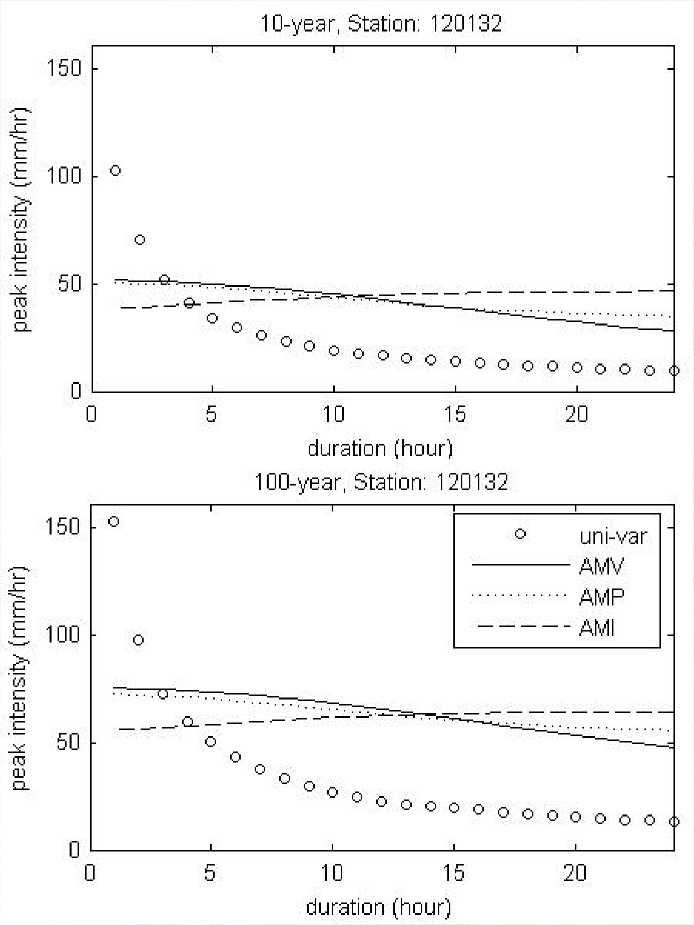


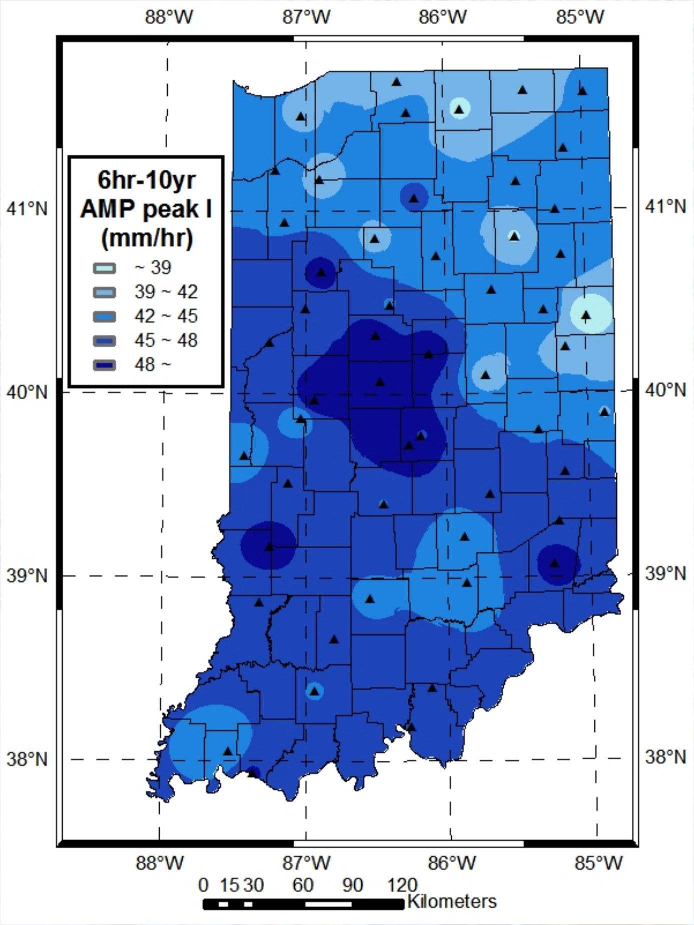
- **AMP definition seems to be an appropriate indicator for defining extreme events**

Oak Ridge National Laboratory

Estimate of peak intensity for known duration

T-year peak intensity i_T given duration d: $F_I(i_T|d-1<D<d)=1-1/T$



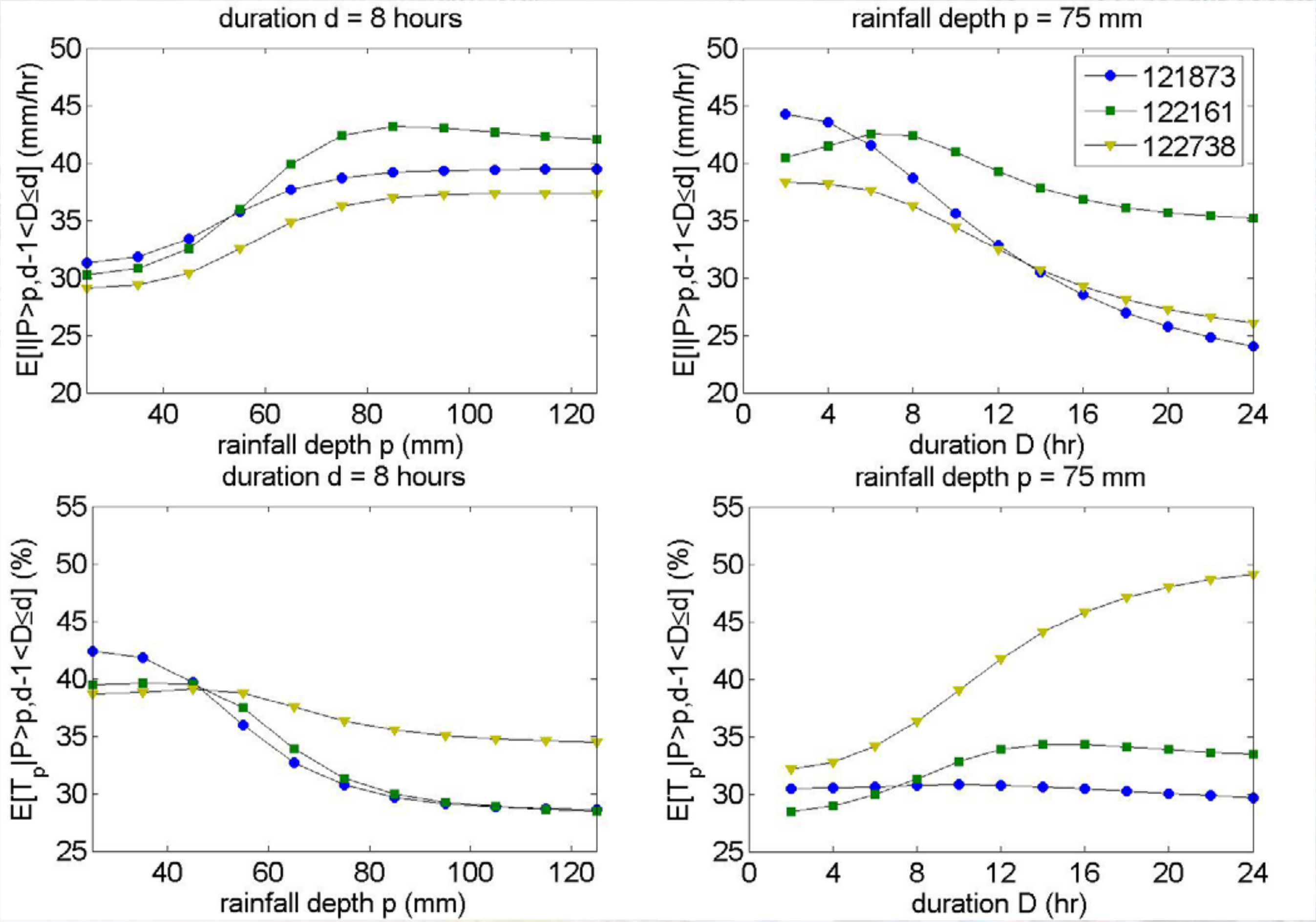


- **Conventional approach fails to capture the peak intensity**

Rainfall Peak Attributes

Geographic Information Science and Technology

- **Given depth (P) and duration (D), compute the conditional expectation of peak intensity (I) and percentage time to peak (T_p)**



Expectation of peak intensity given P & D

Expectation of time to peak (%) given P & D

OAK RIDGE National Laboratory

Drought Frequency Analysis

- **Challenges in characterizing droughts**
  - **No clear (scientific) definition: deficit of water for prolonged time**
  - **Phenomenon dependent in time, space, and between various variables such as precipitation, streamflow, and soil moisture**

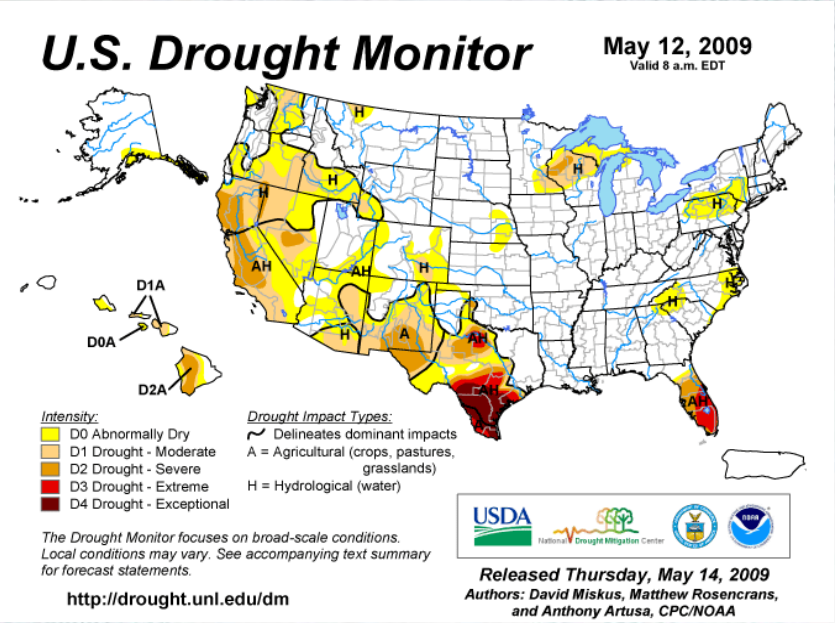
- **Classification of droughts**
  - **Meteorological drought: precipitation deficit**
  - **Hydrologic drought: streamflow deficit**
  - **Agricultural drought: soil moisture deficit**

- **Various drought indices**
  - **Palmer Drought Severity Index (PDSI), Crop Moisture Index (CMI), Surface Water Supply Index (SWSI), Vegetation Condition Index (VCI), CPC Soil Moisture, Standardized precipitation index (SPI)**

US Drought Monitor

Geographic Information Science and Technology

- **Overall drought status (D0 ~ D4) determined based on various indices together (Svobada *et al.*, 2002)**
  - **PDSI**
  - **CPC Soil moisture**
  - **USGS weekly**
  - **Percentage of normal**
  - **SPI**
  - **VCI**



http://drought.unl.edu/DM/MONITOR.html

- **Linear combination of selected indices (OBDI, objective blend of drought indicator) was adopted as the preliminary overall drought status**
- **The decision of final drought status relies on subjective judgment**

OAK RIDGE National Laboratory

Standardized Index Method

- **Proposed by McKee *et al.* (1993)**
- **Generalizable to various types of observations**
  - **For precipitation: SPI**
- **For a given window size, the observed precipitation is transformed to a probability measure using Gamma distribution, then expressed in standard normal variable**

Probabilities of Occurrence (%)	SI Values	Drought Monitor Category	Drought Condition
20 ~ 30	-0.84 ~ -0.52	D0	Abnormally dry
10 ~ 20	-1.28 ~ -0.84	D1	Drought - moderate
5 ~ 10	-1.64 ~ -1.28	D2	Drought - severe
2 ~ 5	-2.05 ~ -1.64	D3	Drought - extreme
< 2	< -2.05	D4	Drought - exceptional

- **Though SIs for different windows are dependent, no representative window can be determined**

Co-occurrence of Droughts

- **Precipitation SIs $\{u_1, u_2, \ldots, u_{12}\}$ and streamflow SIs $\{v_1, v_2, \ldots, v_{12}\}$ are selected**
  - **Annual cycle accounts for the seasonal effect naturally**
  - **Allow for a month-by-month assessment for future conditions**
- **Dependence structure**

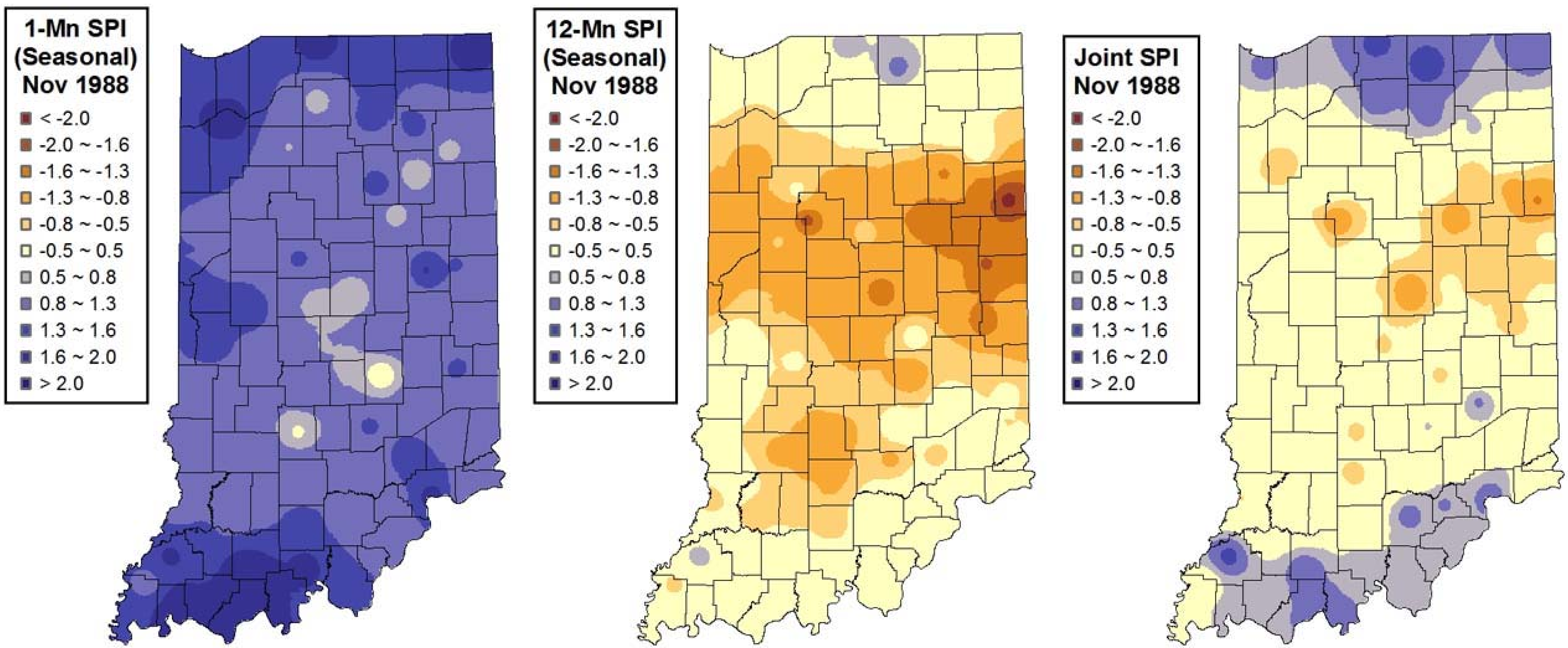
Spearman's $r_{i,j}$ between u_i and u_j (upper triangle); Spearman's $r_{i,j}$ between v_i and v_j (lower triangle)

j \ i	1	2	3	4	5	6	7	8	9	10	11	12
1		0.71	0.57	0.48	0.41	0.38	0.37	0.36	0.35	0.33	0.31	0.30
2	0.89		0.82	0.70	0.61	0.55	0.53	0.51	0.49	0.47	0.44	0.42
3	0.80	0.93		0.87	0.76	0.69	0.64	0.61	0.59	0.56	0.54	0.51
4	0.73	0.85	0.94		0.90	0.81	0.75	0.70	0.67	0.65	0.62	0.60
5	0.67	0.78	0.87	0.95		0.92	0.85	0.79	0.75	0.72	0.69	0.67
6	0.63	0.72	0.81	0.89	0.96		0.93	0.87	0.82	0.78	0.75	0.73
7	0.59	0.68	0.75	0.83	0.90	0.96		0.94	0.89	0.85	0.81	0.78
8	0.57	0.64	0.72	0.79	0.85	0.91	0.97		0.95	0.90	0.86	0.83
9	0.55	0.62	0.69	0.75	0.81	0.87	0.93	0.97		0.96	0.91	0.88
10	0.53	0.60	0.66	0.72	0.78	0.83	0.89	0.94	0.98		0.96	0.92
11	0.51	0.58	0.64	0.70	0.75	0.81	0.85	0.90	0.94	0.98		0.96
12	0.50	0.56	0.62	0.68	0.73	0.78	0.83	0.87	0.91	0.95	0.98	

Joint Deficit Index

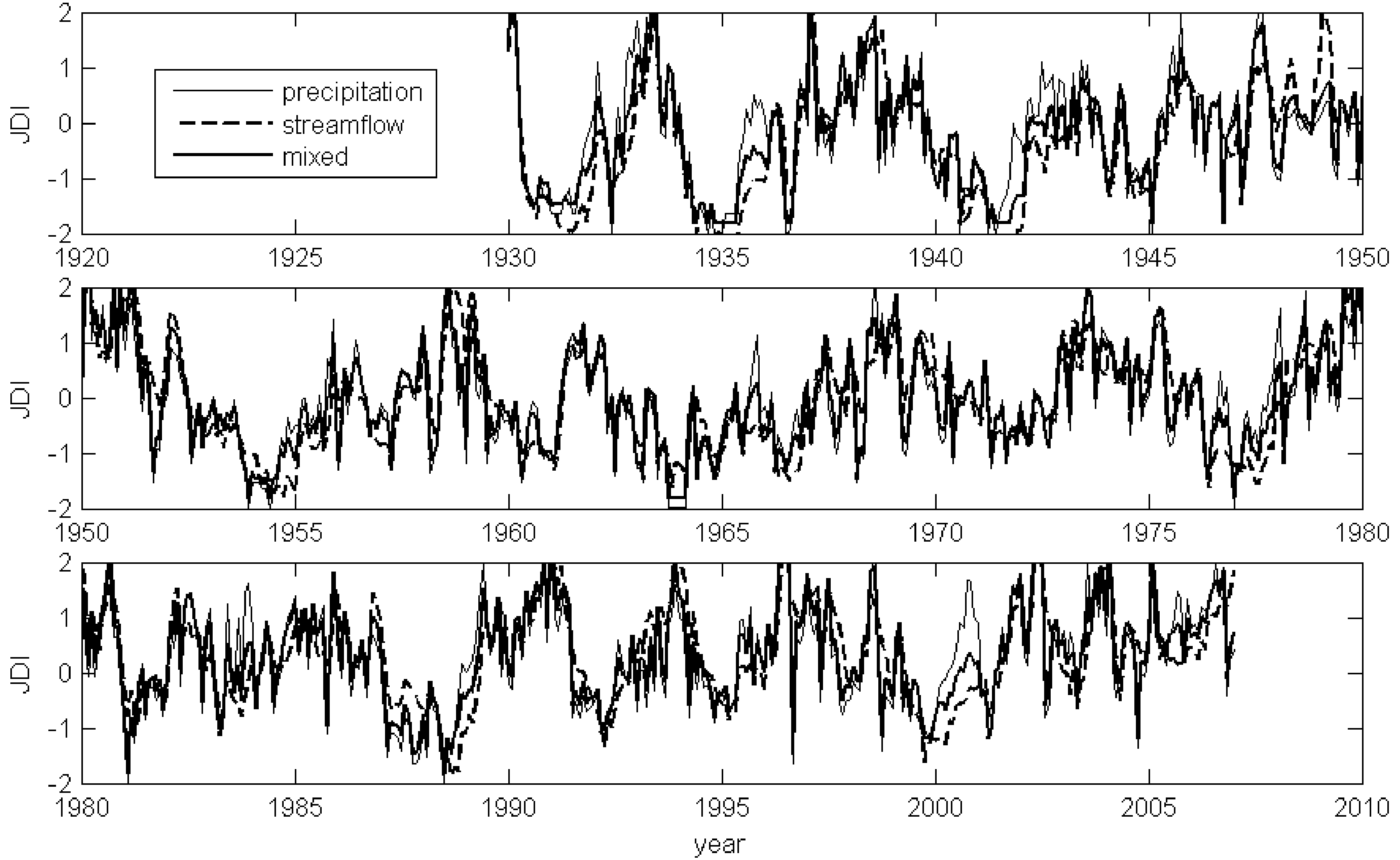
Geographic Information Science and Technology

- **Comparison between 1-mn SPI, 12-mn SPI, and JDI**
  - **12-Mn SPI changes slowly, weak in reflecting emerging drought**
  - **1-Mn SPI changes rapidly, weak in reflecting accumulative deficit**
  - **JDI reflects joint deficit**



Precipitation vs. Streamflow

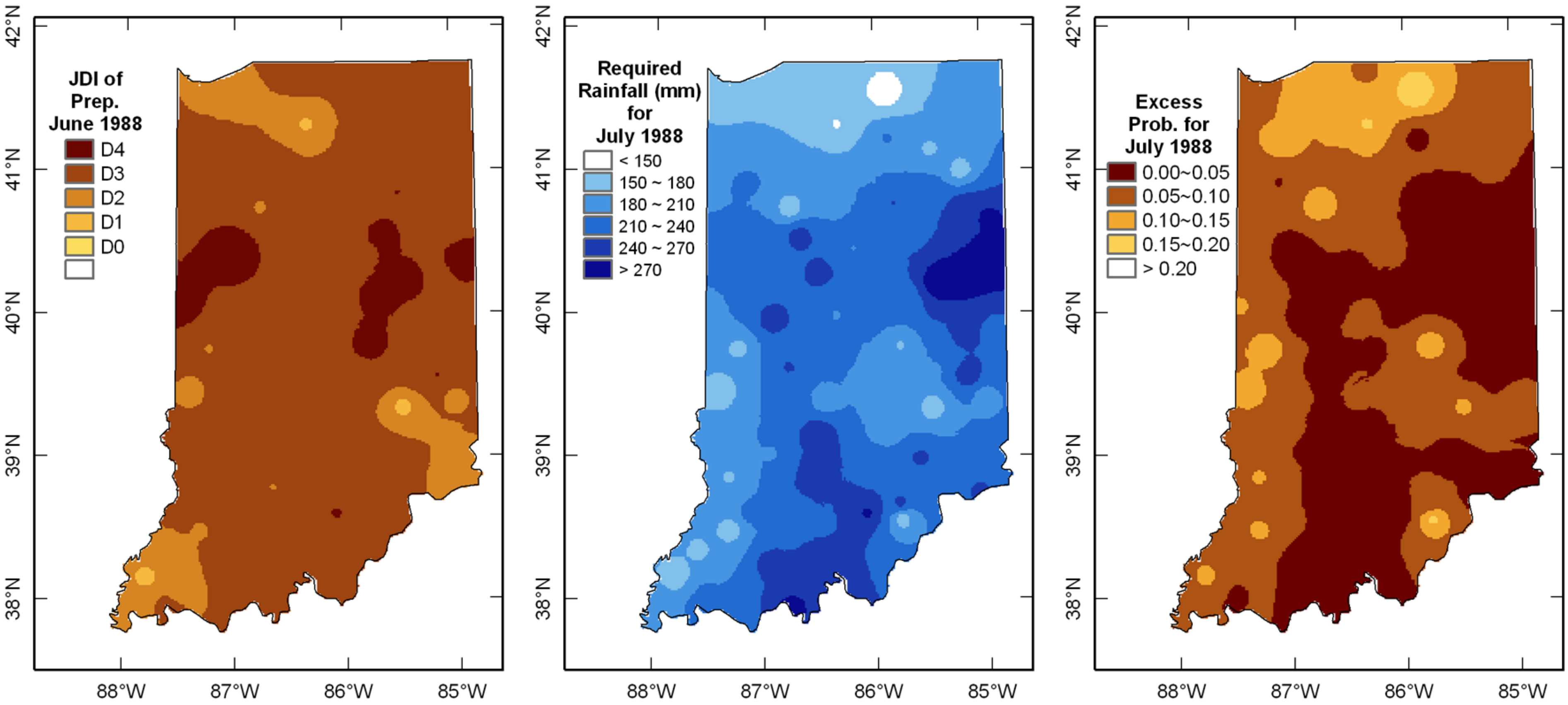
Geographic Information Science and Technology



Potential of Future Droughts

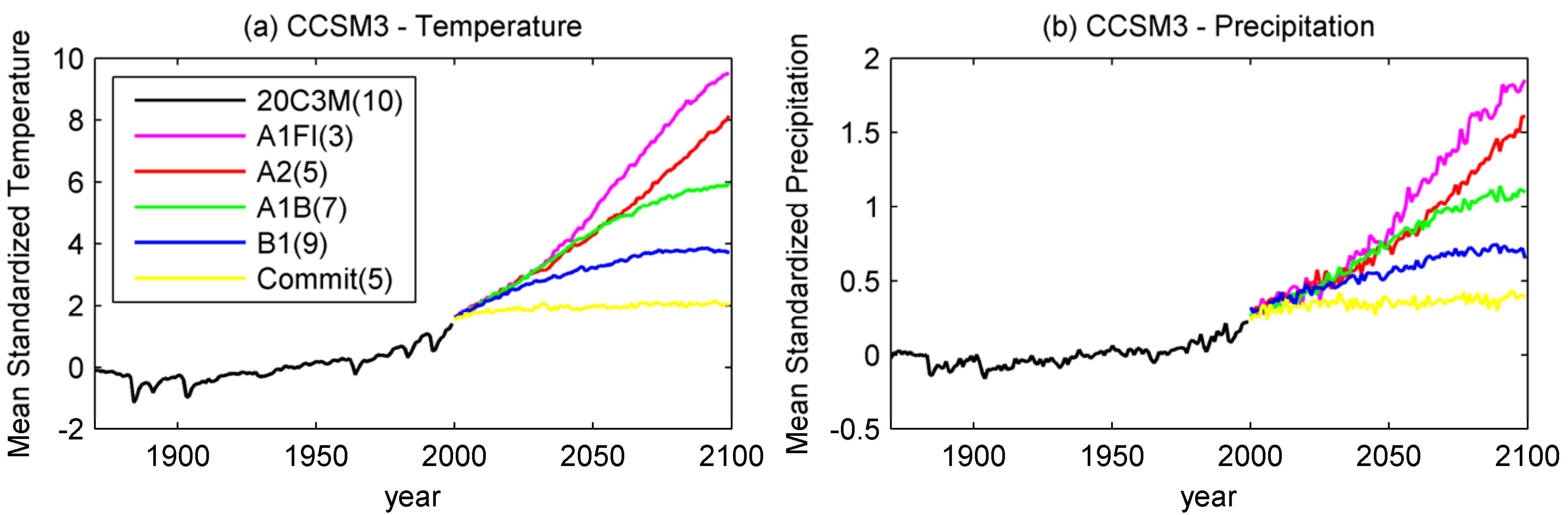
Geographic Information Science and Technology

- **Required precipitation for reaching joint normal status (K_C = 0.5) in the future**
- **Probability of drought recovery**

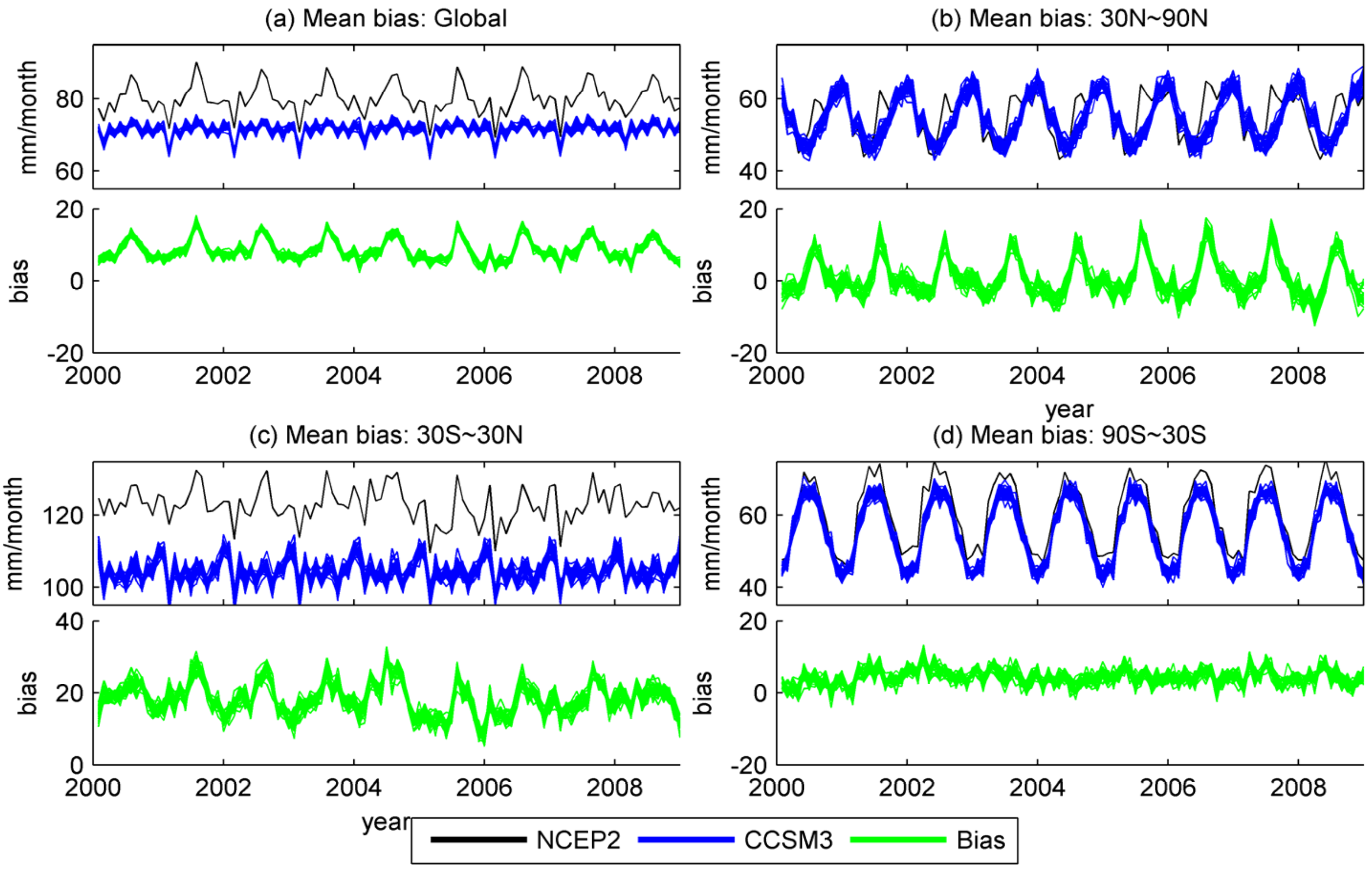


Rainfall in Climate Projections

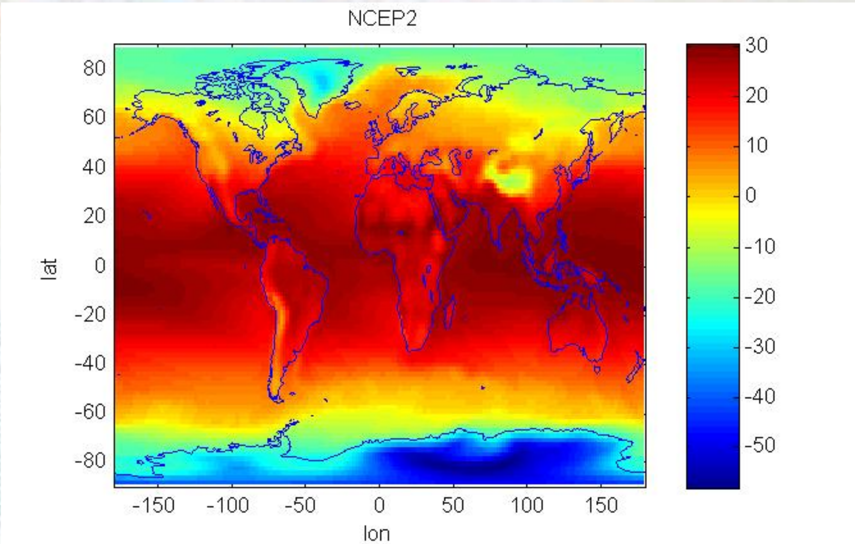
- **Temperature vs precipitation**
  - **Clausius-Clapeyron relationship**
  - **temperature => humidity => precipitable water => precipitation => *Surface Hydrology***
- **Model bias and uncertainty, spatio-temporal variability, extreme rainfall, drought potential, …**
- **Multivariate frequency analysis**
  - **Not so fast!**

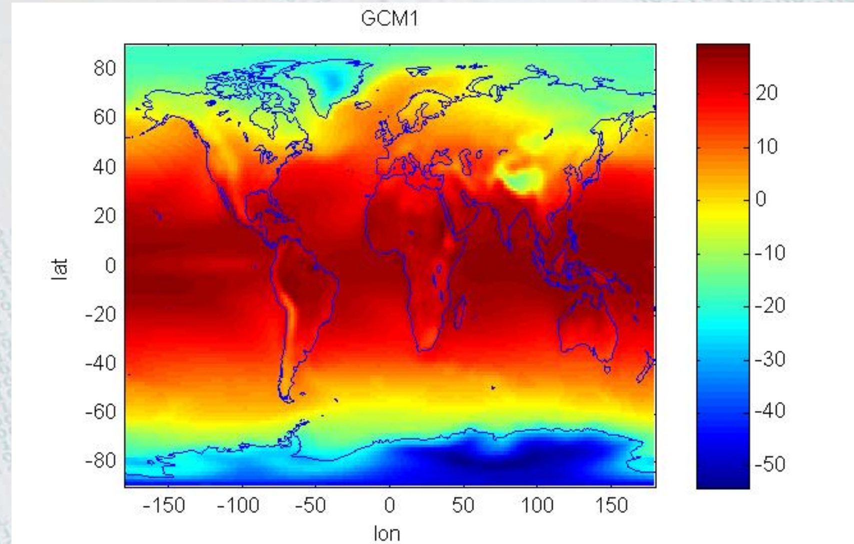


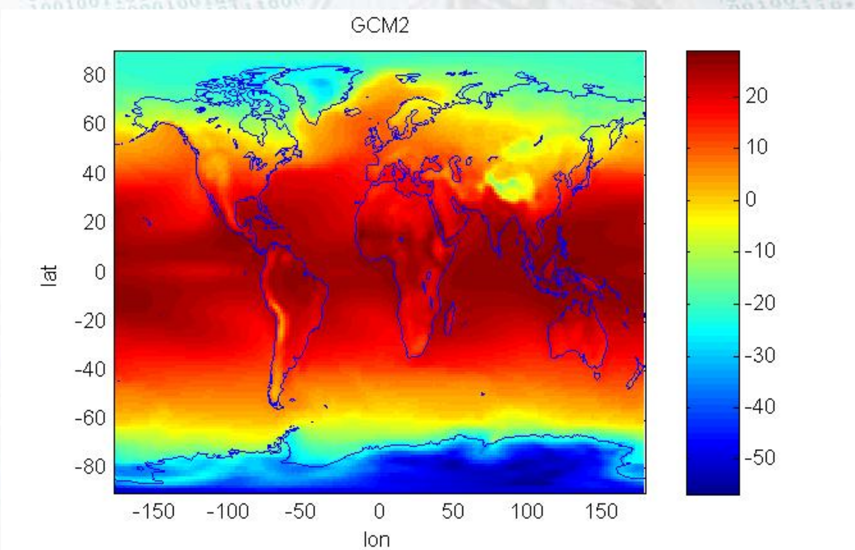
Model Bias

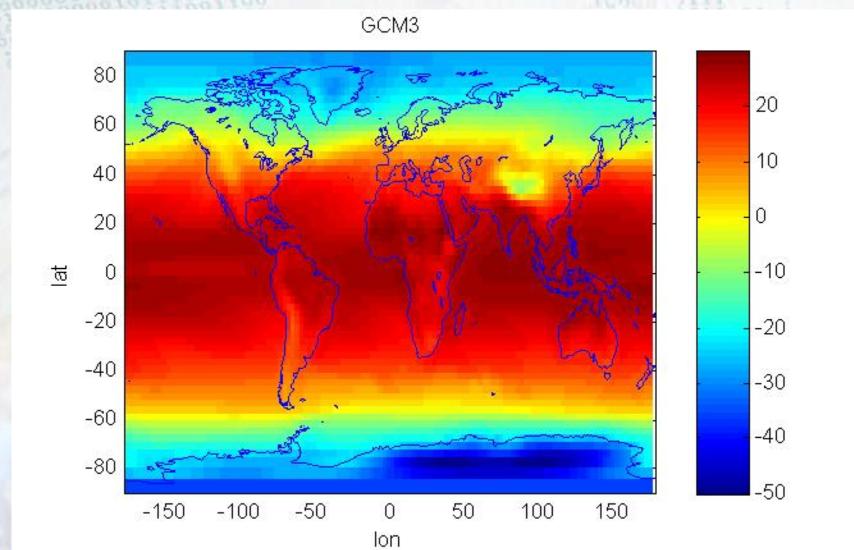

(a) Mean bias: Global
(b) Mean bias: 30N~90N
(c) Mean bias: 30S~30N
(d) Mean bias: 90S~30S
mm/month
bias
year
2000
2002
2004
2006
2008
NCEP2
CCSM3
Bias

Between GCMs

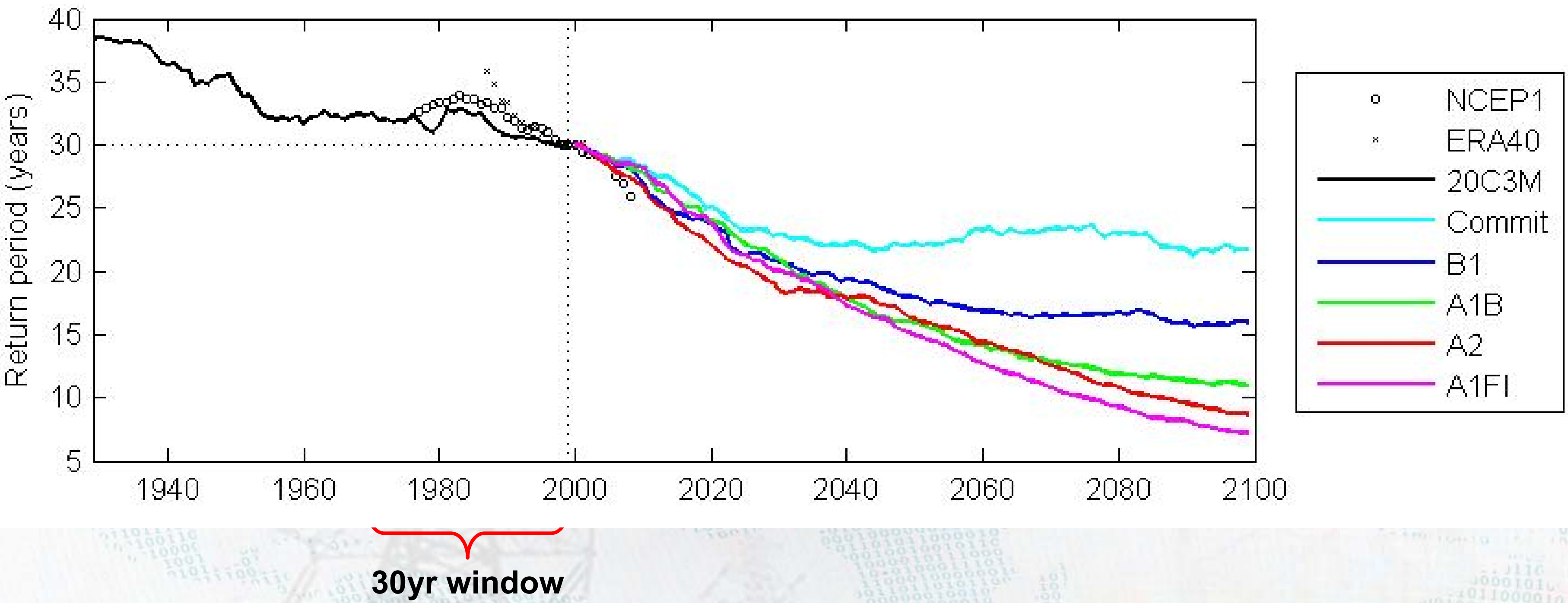








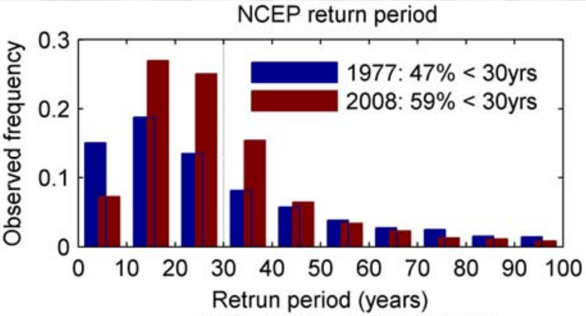
Return Period in the Changing Climate (I)

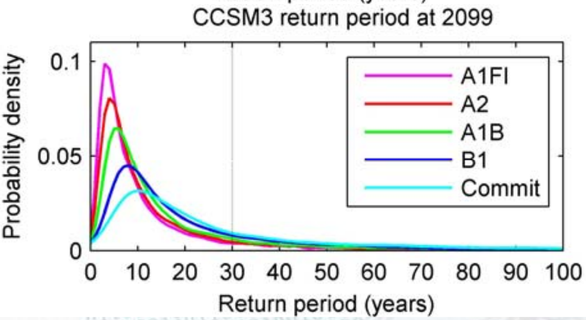


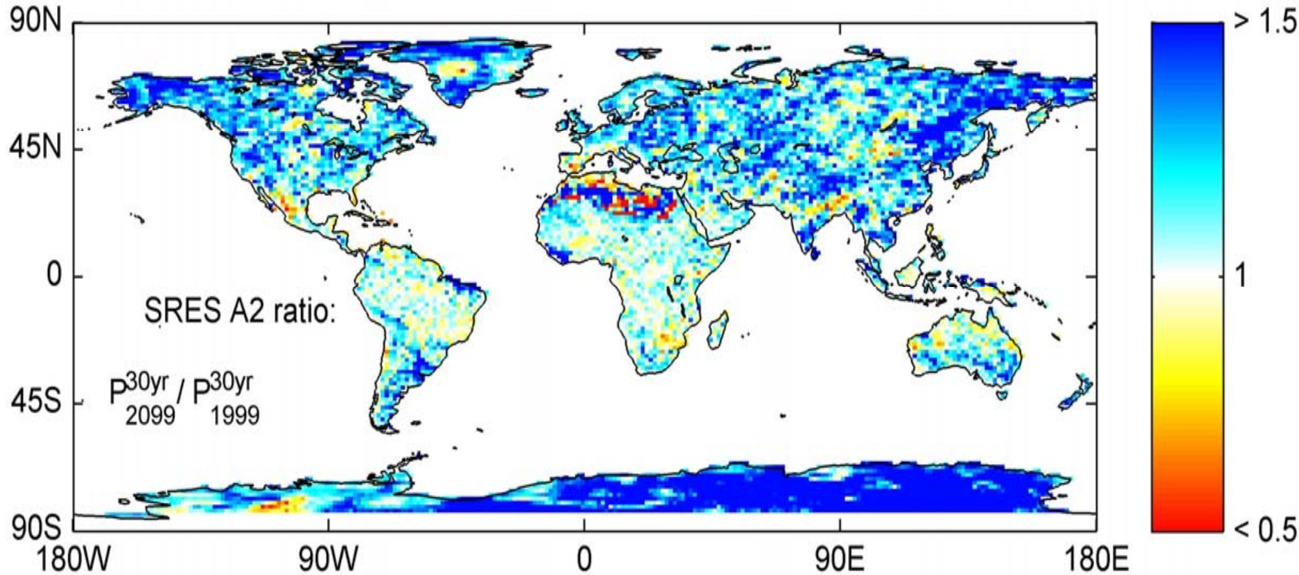
- **Annual maximum precipitation in a 6-hr interval**
- **Generalized extreme values (GEV) dist. with block maximum theory**
- **Median of global return period corresponding to year-1999 estimates**
- **Goodness-of-fit tests at 5% significant level:**
  - **NCEP: 2.56%, ERA40: 1.24%, CCSM3: 0.02%**

Return Period in the Changing Climate (II)

Geographic Information Science and Technology







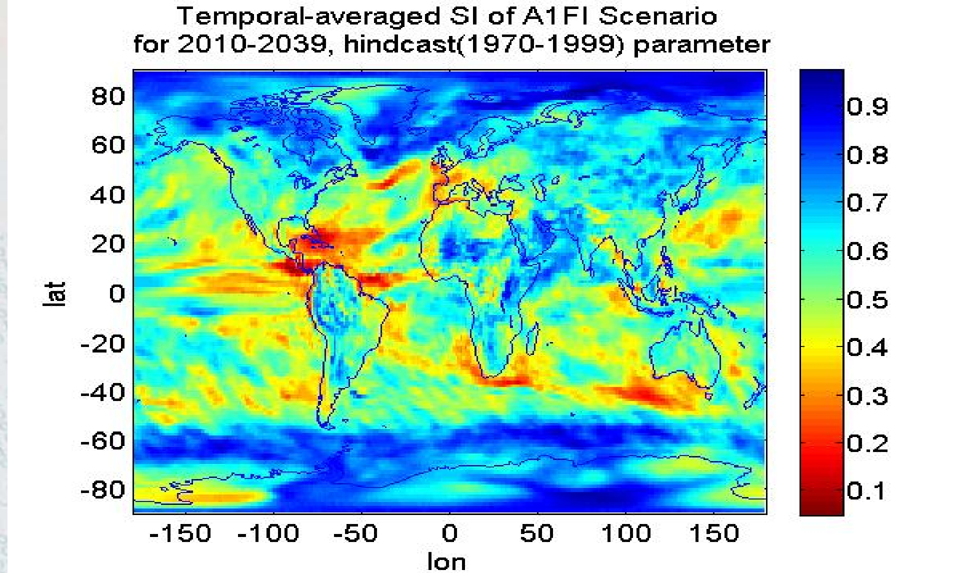
- **Spatial variability**
- **Computational challenges**
  - **Around 33GB outputs, 800 CPU-hour computation time**
  - **Parallel computing environment**
- **Uncertainty quantification**
  - **Bootstrapping => rapid increase in computation time**
- **Multivariate storm events analysis**

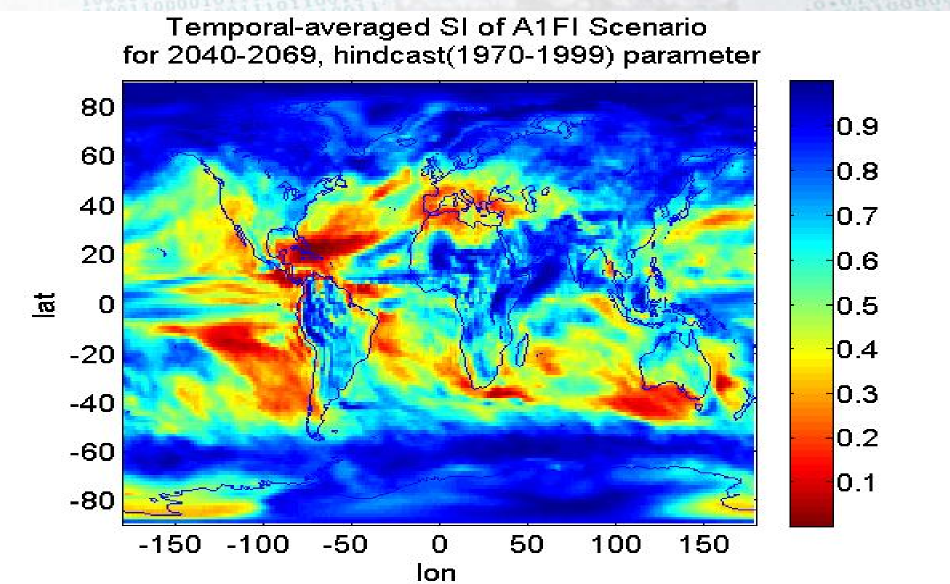
Oak Ridge National Laboratory

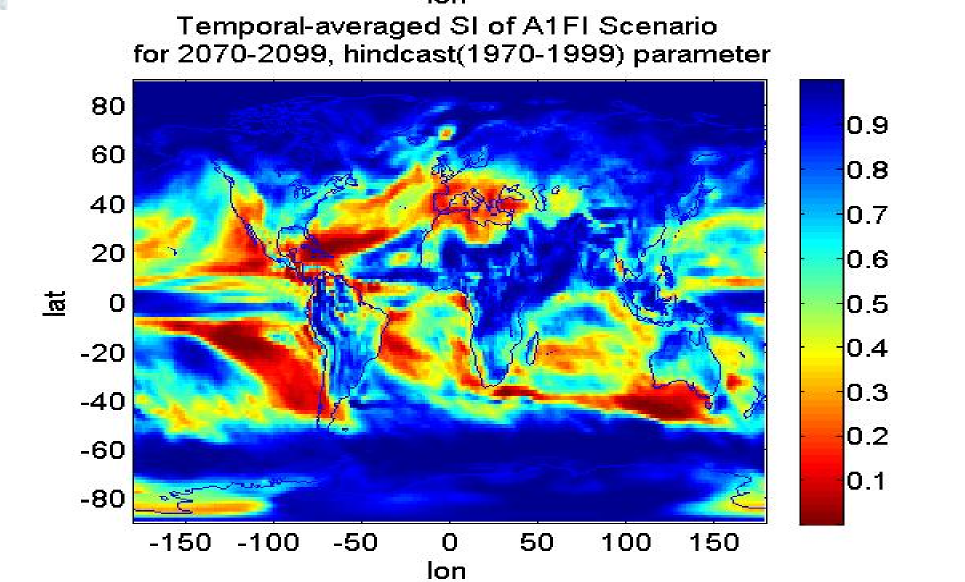
Droughts in the CCSM3 Projections

Geographic Information Science and Technology

- **12-month SPI comparing to the current (1970-1999) moisture status**
- **Assess of water availability**
- **Regions of interest**
- **Co-occurrence of droughts/natural disasters**







Future Works

- **More analysis of hydro-meteorologic components in the climate projections**
  - Specific humidity, wind speed, evapotranspiration, surface flow
  - Extreme, uncertainty, and potential impact
  - Multivariate frequency analysis
- **Multi-model inter-comparison**
  - Multi-model super-ensemble
  - Reanalysis data (NCEP1, NCEP2 and ERA40), and local observation (NOAA and USGS)
- **Co-occurrence of natural disaster**
  - Spatio-temporal and inter-variable dependence structure
- **Statistical/physical downscaling**
- **Prepare for the coming AR5**

Potential Applications in Other Fields

- **Mutual information and non-linear correlation (Auroop)**
- **Complex/social networks (Karsten)**
- **Simulation of household data (Cheng)**
- **Remote sensing data processing (Raju)**
- **Probabilistic decision making in the agent-base modeling (Xiaohui)**
- **Capabilities of copula-based approach**
  - **Median regression**
  - **Markov process**
  - **Copula-based geostatistics**
  - **Monte Carlo simulation**
  - **Conditional distribution and risk**

Concluding Remarks

- **Copulas are found to be flexible for constructing joint distributions**
  - **Toward better quantification of uncertainty and risk**
- **The dependence structure can be faithfully preserved**
- **Caution when using copulas**
  - **Need reliable data**
  - **Difficulties arise in higher dimensions**
    - **Mathematical complexity**
    - **Hard to preserve all lower level mutual dependencies**
    - **Compatibility problem**
    - **Limited choice of parametric models**

Is it the copula's fault?

WIRED MAGAZINE: 17.03

Recipe for Disaster: The Formula That Killed Wall Street

By Felix Salmon 02.23.09

$$\Pr[T_A<1, T_B<1] = \phi_2(\phi^{-1}(F_A(1)), \phi^{-1}(F_B(1)), \gamma)$$

Here's what killed your 401(k) *David X. Li's Gaussian copula function as first published in 2000. Investors exploited it as a quick—and fatally flawed—way to assess risk. A shorter version appears on this month's cover of* Wired.

Probability
Specifically, this is a joint default probability—the likelihood that any two members of the pool (A and B) will both default. It's what investors are looking for, and the rest of the formula provides the answer.

Survival times
The amount of time between now and when A and B can be expected to default. Li took the idea from a concept in actuarial science that charts what happens to someone's life expectancy when their spouse dies.

Equality
A dangerously precise concept, since it leaves no room for error. Clean equations help both quants and their managers forget that the real world contains a surprising amount of uncertainty, fuzziness, and precariousness.

Copula
This couples (hence the Latinate term copula) the individual probabilities associated with A and B to come up with a single number. Errors here massively increase the risk of the whole equation blowing up.

Distribution functions
The probabilities of how long A and B are likely to survive. Since these are not certainties, they can be dangerous: Small miscalculations may leave you facing much more risk than the formula indicates.

Gamma
The all-powerful correlation parameter, which reduces correlation to a single constant—something that should be highly improbable, if not impossible. This is the magic number that made Li's copula function irresistible.

Thank you
Questions?

Shih-Chieh Kao
kaos@ornl.gov; http://www.ornl.gov/~5v1/