

Statistical Analysis of Extreme Rainfall Events over Indiana, USA

Shih-Chieh Kao[1] and Rao S. Govindaraju[2]

[1] School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA; email: kao@purdue.edu
[2] School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA; email: govind@purdue.edu

Abstract

Analysis of extreme rainfall events is important for hydraulic and hydrologic studies, and has conventionally been performed by pre-specifying rainfall duration as a filter to abstract the information of annual maximum rainfall depths for further examination. However, this single-variate approach does not account for dependence between rainfall properties. To characterize extreme rainfall events, a multi-variate analysis is conducted in this study using hourly precipitation data from Indiana, USA. Samples of extreme rainfall events are chosen based on two different criteria: annual maximum volume, and annual maximum peak intensity. Rainfall properties, such as total depth, duration, and peak intensity are analyzed using copulas to describe the dependence structures between rainfall variables and to construct their joint distribution for extreme rainfall events. Results from the derived multivariate model are compared to those from conventional single-variate analysis by computing the corresponding conditional distributions. The proposed stochastic model for extreme rainfall is expected to provide better estimates of design rainfall.

Key words: Copulas, multivariate analysis, joint-distribution, extreme rainfall

1. Introduction

In order to prevent loss of property and human life, designs of hydraulic and hydrologic structures are based on extreme rainfall events. Since a deterministic relationship for extreme rainfall events in future cannot be established, statistical methods are adopted to quantify rainfall by probability of exceedance (for example, the use of return period and hydrologic risk) as in rainfall frequency analysis.

Rainfall frequency analysis is currently performed through single-variate approaches, i.e. by treating the total rainfall depth as the only variable (for instance, Bonnin *et al.*, 2004). To relate rainfall depth to duration, a pre-specified duration is declared as a filter to find the annual maxima as samples for analysis. In this sense, stochastic rainfall models are constructed for various "durations". However, it should

be noted that this "duration" is artificially prescribed and does not reflect the actual duration of rainfall events. When using a shorter prescribed duration (say 1-hour), the selected maximum event may be from a longer duration extreme rainfall event, and possibly represents the peak intensity part. On the other hand, when using a longer prescribed duration (say 48-hour), the selected maximum may cover several short-term events with periods of rainfall hiatus. Therefore, the current practice provides estimates for various artificial durations, but is not able to truly characterize the behavior of extreme rainfall events. Rainfall records reveal that rainfall events exhibit high variability in their properties such as total depth (volume), duration, and peak intensity. Clearly, there is a need to perform a muti-variate analysis to construct a more realistic stochastic model for extreme rainfall events.

However, multi-variate frequency analyses are much more complicated than single-variate procedures. The main challenge is due to the mathematical complexity of the joint probability distribution that encompasses knowledge of both marginal distributions and dependence structure. Over the last decade, copulas have emerged as a method for addressing multi-variate problems in several disciplines. Using Sklar's (1959) theorem, the analysis of joint distributions can be performed separately for the marginal distributions and for the dependence structure. Nelsen (2006) provides a theoretical background and description on the use of copulas. De Michele and Salvadori (2003) were perhaps the first to apply copulas in hydrology to analyze the joint behavior between rainfall duration and average intensity. Grimaldi and Serinaldi (2006) explored the multi-variate relationships among critical rainfall depths, peak intensity, and total depth. These studies explained the methodology for constructing multi-variate stochastic rainfall models. However, due to the small sample size adopted (both studies used 7-year data), the behavior of extreme rainfall events could not be studied in either case. Other examples of applications of copulas in hydrology are Favre *et al.* (2004), Salvadori and De Michele (2004), De Michele *et al.* (2005), Salvadori and De Michele (2006), and Zhang and Singh (2006).

In this study, three defining properties of rainfall: total depth (volume) P, duration D, and peak intensity I are utilized to perform multi-variate frequency analysis. Sufficiently long (over 50 years) hourly precipitation datasets are adopted to provide a more statistically reliable description of extreme rainfall behavior. Multi-variate frequency analysis based on copula technique is performed somewhat analogous to the conventional single-variate approach. This new stochastic model is expected to provide a better understanding of extreme rainfall.

2. Selection of Extreme Events

53 hourly raingauges from Hourly Precipitation Database (TD 3240) of National Climate Data Center (NCDC, http://www.ncdc.noaa.gov/oa/ncdc.html) in Indiana are selected in this study. Each selected station possessed 50 to 55 years of data, which should be sufficient for performing single-variate at-site frequency analysis (criterion taken from Bonnin *et al.*, 2004). A minimum rainfall hiatus of six hours between non-zero records was selected to abstract rainfall events (Huff, 1967). An average of about 4800 observed events are available for each station.

Unlike the definition of annual maximum precipitation series used in conventional analysis, the definition of annual maximal events for mutli-variate

problems is somewhat ambiguous. Depending on the problem at hand, rainfall intensity or volume may govern hydrologic design. Therefore, in this study, two sets of extreme events are selected for each station by using annual maximum volume (AMV) and annual maximum peak intensity (AMI) as the defining criteria. An example plot of the selected extreme events for station Alpine 2 NE (COOPID: 120132) is shown in Figure 1. It can be observed that these two criteria result in different distributions. AMV events seem to be long-term (half of them are over 20 hours), while AMI events seem to be short-term (half of them are less than 6 hours). As expected, the total volume of AMV events is generally higher than AMI events, while the average intensity of AMV events is less than AMI events. The selected sets of events are analyzed further.

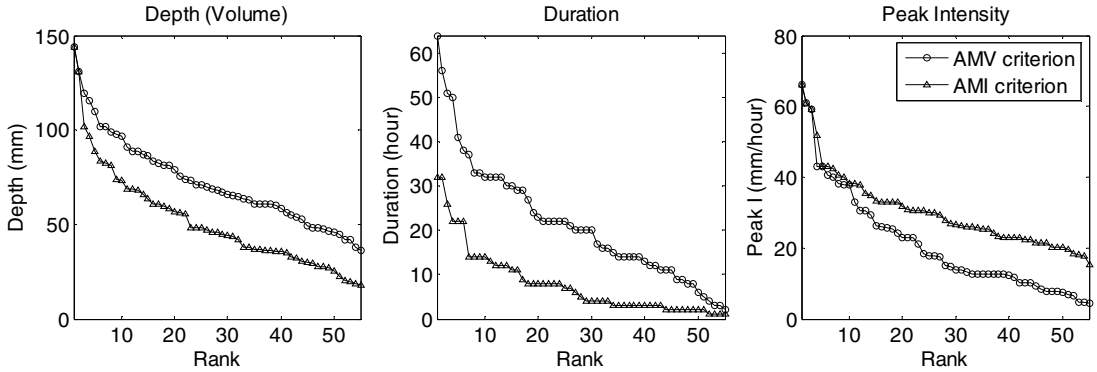

Figure 1 – AMV and AMI events of station Alpine 2 NE (COOPID: 120132)

3. Analysis of Marginal Distributions

The process of constructing joint distribution through copulas can be decomposed into two parts: marginal distributions, and dependence structure. Marginal distributions are analyzed through the conventional single-variate approach for each station. In this study, six probability density functions (PDFs) are applied and tested for their applicability to individual rainfall attributes. They are extreme value type I (EV1), generalized extreme value (GEV), Pearson type III (P3), log-Pearson type III (LP3), generalized Pareto (GP), and the log-normal (LN) distribution. The theoretical background for single-variate analysis can be obtained from Rao and Hamed (2000). Model parameters are estimated primarily by maximum likelihood (ML) method or by method of moments (MOM). Gringorton formula is chosen to estimate the empirical probabilities. Chi-square and Kolmogorov-Smirnov (KS) tests are applied for goodness-of-fit with 10% significance level. The summary of test results for the 53 selected stations in Indiana is shown in Table 1. EV1 fitting of station Alpine 2 NE is shown in Figure 2 as an example.

Based on the rejection rate, it can be observed that EV1, GEV, LP3, and LN provide better fit than P3 and GP. It should be noticed that though GP was reported supreme by De Michele and Salvadori (2003) for their model describing regular rainfall events, it is found to be the weakest distribution in this study. This suggests the different nature of extreme rainfall behavior when compared to regular rainfall. It is also observed that fitting for duration of AMI events can not yield good result. This may be due to the fact that most AMI events are short-term, and therefore the

Table 1 – Summary of Chi-square and KS test results for marginal distributions

AMV	Rejection rate (%) of Chi-square test						Rejection rate (%) of KS test					
events	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	Rejection rate (%) of Chi-square test						Rejection rate (%) of KS test					
events	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

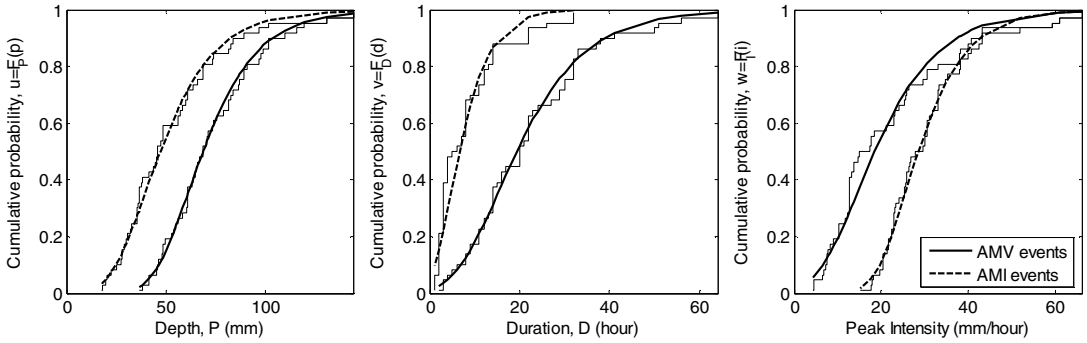

Figure 2 – EV1 fitting for marginal distributions of station Alpine 2 NE (COOPID: 120132)

recording unit (hour) used in this study is not fine enough. The marginals (or cumulative density functions (CDFs)) of depth P, duration D, and peak intensity I are expressed as $u = F_P(p)$, $v = F_D(d)$, and $w = F_I(i)$ in the following discussion.

4. Analysis of Dependence Structure Using Copulas

A copula C is a function composed of marginals. Sklar (1959) showed that for continuous random variable X and Y with marginals $F_X(x) = u$ and $F_Y(y) = v$, there exists one unique C_{UV} such that:

$$C_{UV}(u, v) = C_{UV}(F_X(x), F_Y(y)) = H_{XY}(x, y) \quad (1)$$

where H_{XY} is the joint distribution. Since probability measurements are absolutely increasing (for absolutely increasing continuous random variables) from 0 to 1, copulas C_{UV} can be regarded as a transformation of H_{XY} from $[-\infty, \infty]^2$ to $[0,1]^2$. In other words, it simplifies the joint distribution to a bounded domain, and therefore attentions can be focused on the dependence structure described by copulas.

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

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

In (2), the generator φ is an absolutely decreasing function defined in [0,1], and $\varphi(1) = 0$. A special case is the independent copula $\Pi(u,v) = uv$ with generator $\varphi(t) = -\ln t$. For Archimedean copulas, several statistical properties can be simply expressed in terms of φ, such as the distribution function K_C of copulas (i.e. $K_C(t) = P[C(u,v) \le t]$) and the concordance measure Kendall's τ:

$$K_C(t) = t - \frac{\varphi(t)}{\varphi'(t)}, \qquad t \in [0,1] \quad (3)$$

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

The distribution function K_C offers a cumulative probability measure for the set $\{(u,v) \in [0,1]^2 | C(u,v) \le t\}$ and therefore can be applied for examining the goodness-of-fit of copulas onto one single dimension (along t). By using (3), the theoretical Kendall's τ can be derived as (4). Apart from being a better measurement of dependence than the traditional correlation coefficient ρ, Kendall's τ has also been extensively used for obtaining a non-parametric estimator for dependence parameter θ by equating sample $\hat{\tau}$ to theoretical τ (for example, De Michele and Salvadori (2003), Favre *et al.* (2004), Zhang and Singh (2006)). This estimator does not rely on prior information of marginal distributions, and hence provides a more objective measure of dependence structure.

In this study, sample $\hat{\tau}_{PD}$, $\hat{\tau}_{DI}$, and $\hat{\tau}_{PI}$ for each pair of variables P, D, and I are computed for all selected stations. Mean and standard deviation are computed and tabulated in Table 2.

Table 2 – Statistics of Kendall's $\hat{\tau}$ between extreme rainfall properties in Indiana

	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

It can be observed that depth and duration are positively correlated, duration and peak intensity are negatively correlated, and depth and peak intensity are positively correlated. It is important to notice that the dependence level are not high (close to ± 1) in each case, and are neither low (close to 0) except for duration and peak intensity for AMI events. When dependence level is close to ± 1, number of variables can be reasonably reduced and replaced by a reliable regression formula. On the contrary, low dependence validates the assumption of independence and hence the joint distribution decays to a simple product of marginals. These two limiting

approximations are common in engineering applications, but seem not appropriate for mid-dependence as in this study. When analyzing important problems like extreme rainfall behavior, the construction of dependent joint distribution is inevitable. It can also be observed that the dependence levels for AMV and AMI events are not similar. Not surprisingly, stochastic models based on events selected by different criteria would lead to different models, and the choice of which one to adopt should be based on the nature of the problem at hand. For example, rainfall models constructed from AMV events are likely more suitable for longer durations and larger watersheds, while AMI events provide better models for shorter durations and smaller watersheds.

The choice of a copula function depends on the range of dependence level it can describe. Numerous families of Archimedean copulas are available for positive dependence structure. In this study, four commonly used families of one-parameter Archimedean copulas are adopted and examined, including: Frank, Clayton, Genest-Ghoudi, and Ali-Mikhail-Haq. All of these are valid both for positive and negative dependence (note that Ali-Mikhail-Haq is valid only for -0.1807 < τ < 0.3333). The parameters are estimated by non-parametric procedure using Kendall's τ. For sample size n, empirical copulas C_n described by Nelsen (2006) are computed for examining goodness-of-fit:

$$C_n\left(\frac{i}{n},\frac{j}{n}\right)=\frac{a}{n} \tag{5}$$

where a is the number of pairs (x,y) in the sample with $x \le x_{(i)}$ and $y \le y_{(j)}$, and $x_{(i)}$, $y_{(j)}$, $1 \le i,j \le n$, is the order statistics from the sample. Similarly, empirical distribution function K_{C_n} can be written as:

$$K_{C_n}\left(\frac{k}{n}\right)=\frac{b}{n} \tag{6}$$

where b is the number of pairs (x,y) in the sample with $C_n(i/n, j/n) \le k/n$. K_C and K_{C_n} are plotted for visual examination, and an example of AMV events for station Alpine 2 NE is shown in Figure 3.

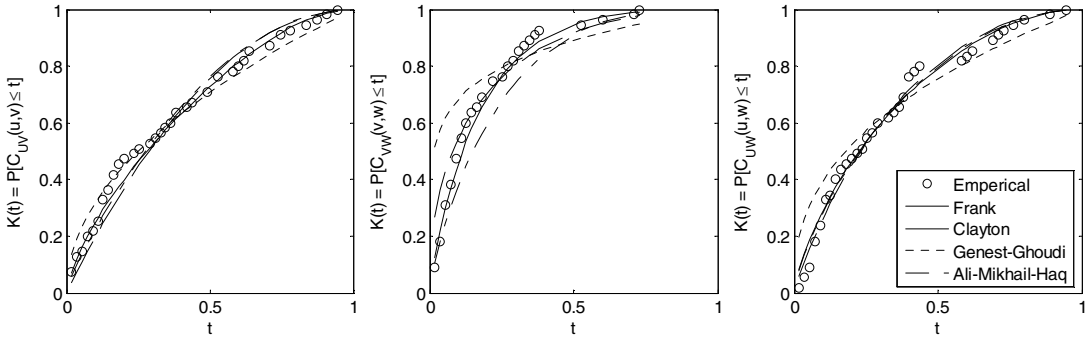

Figure 3 – Visual examination using K_C for station Alpine 2 NE (COOPID: 120132)

Generally, it is observed that Clayton and Ali-Mikhail-Haq families perform well for positive dependence cases (C_{UV} and C_{UW}), and Frank family performs well for both positive and negative dependence. In fact, Frank family is the only Archimedean copula which satisfies radial symmetry, and is suitable for the entire range of dependence. It makes Frank family a popular choice for constructing dependence structure.

5. Joint Distribution and Applications

The bivariate joint distribution can be constructed by merging marginal distribution and dependence structure obtained in sections 3 and 4. This bivariate model can be applied for many purposes, such as risk assessment, flood frequency derivation, and expectation computation for rainfall-related properties. An application of conditional distribution is presented here. For a known (measured) d-hour rainfall event, the conditional cumulative distribution for depth P can be written as:

$$F_P\left(p \middle| d-1 < D \le d\right) = \frac{H_{PD}(p,d) - H_{PD}(p,d-1)}{F_D(d) - F_D(d-1)} \qquad (7)$$

where the joint distribution is constructed by GEV marginals and Frank family of Archimedean copulas. For given return period T, the T-year, d-hour rainfall estimate p_T will satisfy $F_P\left(p_T \middle| d-1 < D < d\right) = 1 - 1/T$. An example for station Alpine 2 NE is shown in Figure 4 along with the corresponding conventional single-variate rainfall estimates using GEV distribution fitted separately for various durations. It is found that AMV estimates are smaller than AMI estimates in the longer duration range, and larger than AMI estimates in the shorter duration range. The conventional single-variate GEV estimates are closer to AMV estimates for longer duration and closer to AMI estimates for shorter duration. It may be recalled that the abstracted AMV samples generally correspond to longer durations, and hence AMV estimates should be more reliable in this range. Similarly, AMI estimates are prone to be better for shorter durations. The similarity between AMV estimates and single-variate estimates for longer durations in Fig. 4 further supports the applicability of AMV estimates in longer duration range. It is interesting to note that the samples of volume for AMI are smaller than AMV, but the resulting conditional estimates show the opposite trend. Because the bivariate model has more parameters than its single-variate counterpart, this proposed model is more flexible and is expected to provide better estimates for characterizing extreme rainfall behavior. More study is required on this important topic, and copulas offer a promising approach.

6. Conclusions

The following conclusions are presented on the basis of this study.

(1) Samples of annual extreme rainfall are selected by AMV and AMI criteria in this study. It is found that the duration of AMV events is generally longer than AMI events. The model based on AMV events is expected to perform better for long-term rainfall and larger watersheds, while AMI should be better for short-

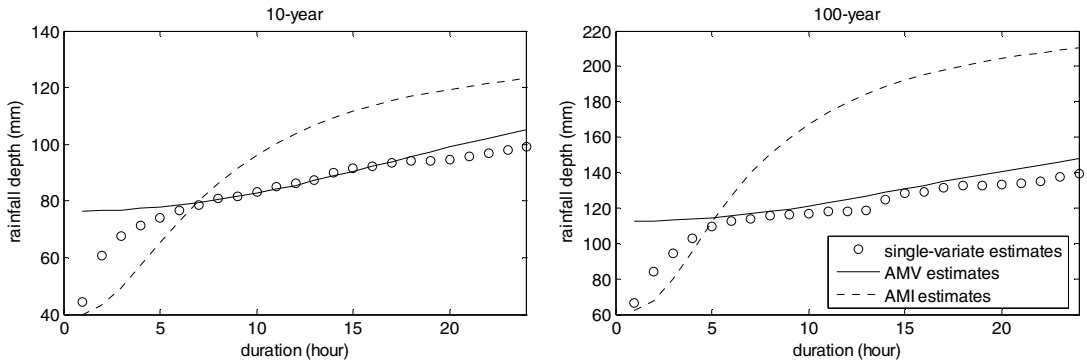

Figure 4 – Rainfall estimates for various durations for station Alpine 2 NE (COOPID: 120132)

term rainfall and smaller watersheds where the effect of peak intensity is likely to be more prominent.

(2) The total volume (depth), duration, and peak intensity are selected as variables of interest in this study. EV1, GEV, LP3, and LN are found to be appropriate marginal models for extreme rainfall. While GP was found to perform well for regular rainfall models in previous studies, it is found to be the weakest in this study for extreme events.

(3) The dependence between volume and duration is found to be positively correlated, between duration and peak intensity to be negatively correlated, and between volume and peak intensity to be positively correlated. The Frank family of Archimedean copulas was shown to be an appropriate model for characterizing these dependence structures.

(4) The bivariate joint distribution can be constructed by merging marginal distribution and dependence structure. The application of conditional distribution of depth given a known measured duration yields rainfall estimates that are qualitatively similar to what is obtained through the conventional single-variate approach. This proposed multi-variate stochastic rainfall model is expected to provide a better characterization for extreme rainfall behavior in Indiana.

References

Bonnin, G. M., Martin, D., Lin, B., Parzybok, T., Yekta, M., and Riley, D. (2004). "Precipitation-Frequency Atlas of the United States", NOAA Atlas 14, Volume 2, U. S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Silver Spring, Maryland.

De Michele, C., and Salvadori, G. (2003). "A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas", *Journal of Geophysical Research*, 108(D2), ACL 15-1-11.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., and Rosso, R. (2005). "Bivariate Statistical Approach to Check Adequacy of Dam Spillway", *Journal of Hydrologic Engineering*, *10*(1), 50-57.

Favre, A.-C., El Adlouni, S., Perreault, L., Thiémonge, N. and Bobée, B. (2004). "Multivariate hydrological frequency analysis using copulas", *Water Resour. Res.*, *40*, W01101.

Grimaldi, S. and Serinaldi, F. (2006). "Design hyetograph analysis with 3-copula

function", *Hydrological Sciences Journal*, *51*(2), 223-238.

Huff, Floyd A. (1967). "Time Distribution of Rainfall in Heavy Storms", *Water Resources Research*, 3, 1007-1019.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd edition, Springer-Verlag New York, Inc., New York, NY.

Rao, A. R., and Hamed, K. H. (2000). *Flood Frequency Analysis*, CRC Press LLC, FL.

Salvadori, G., and De Michele, C. (2004). "Frequency analysis via copulas: Theoretical aspects and applications to hydrological events", *Water Resour. Resear.*, *40*, W12511.

Salvadori, G., and C. De Michele (2006), Statistical characterization of temporal structure of storms, *Advances in Water Resources*, *29*(2006), 827-842.

Sklar, A. (1959). "Fonctions de repartition a n dimensions et leurs marges", *Publ. Inst. Statist. Univ. Paris* 8, 229-231.

Zhang, L., and Singh, V. P. (2006). "Bivariate Flood Frequency Analysis Using the Copula Method", *Journal of Hydrologic Engineering*, *11*(2), 150-164.