
1 **Copula-based frequency analysis of overflow and flooding**
2 **in urban drainage systems**

3
4 **Guangtao Fu*, David Butler**

5
6 Centre for Water Systems, College of Engineering, Mathematics and Physical Sciences,
7 University of Exeter, North Park Road, Harrison Building, Exeter EX4 4QF, UK

8 *Corresponding author. Email: g.fu@exeter.ac.uk; Phone: +44 (0)1392 723692; Fax: +44 (0)1392 217965
9

10 **Abstract:**

11 The performance evaluation of urban drainage systems is essentially based on accurate
12 characterisation of rainfall events, where a particular challenge is development of the joint
13 distributions of dependent rainfall variables such as duration and depth. In this study, the copula
14 method is used to separate the dependence structure of rainfall variables from their marginal
15 distributions. Three one-parameter Archimedean copulas, including Clayton, Gumbel, and Frank
16 families, are fitted and compared for different combinations of marginal distributions that cannot
17 be rejected by statistical tests. The fitted copulas are used, through the Monte Carlo simulation
18 method, to generate synthetic rainfall events for system performance analysis in terms of sewer
19 flooding and Combined Sewer Overflow (CSO) discharges. The copula method is demonstrated
20 using an urban drainage system in the UK, and the cumulative probability distributions of
21 maximum flood depth at critical nodes and CSO discharge volume are calculated. The results
22 obtained in this study highlight the importance of taking into account the dependence structure of
23 rainfall variables in the context of urban drainage system evaluation and also reveal the different

24 impacts of different dependence structures on the probabilities of sewer flooding and CSO
25 volume.

26 **Keywords:** Bivariate distribution, combined sewer overflow, copula, frequency analysis, sewer
27 flooding, urban drainage system

28 **1. Introduction**

29 Urban drainage systems are used in urban areas for flood and pollution control through collection
30 and conveyance of stormwater and dry weather flow (DWF) to receiving waters and wastewater
31 treatment plants. Most systems in the UK and many other countries are combined sewer systems,
32 in which both DWF and stormwater flow in a single pipe network. Such combined sewer
33 systems have two common issues: sewer flooding and combined sewer overflow (CSO)
34 discharges when the flow exceeds the available system capacity (Butler and Davies, 2011). Their
35 economic, social and environmental impacts have been discussed in detail in the literature (e.g.,
36 Schmitt et al., 2004; Fu et al., 2009; Andres-Domenech et al., 2010). Most sewer systems in the
37 developed countries were constructed many decades ago and designed using simple deterministic
38 methods on the basis of design rainfall events, which are usually related to a specified return
39 period and generated from intensity-duration-frequency curves (Hvitved-Jacobsen and Yousef,
40 1988; Butler and Davies, 2011). System performance is affected by many factors that may have
41 changed over time such as system characteristics, land use and climate change. Thus, there is a
42 need to assess the performance of sewer systems regarding sewer flooding and CSO discharging
43 in a changed situation (Korving et al., 2002; Schmitt et al., 2004; Thorndahl and Willems, 2008).
44 This is also driven by strict regulations such as the Water Framework Directive in the member
45 states of EU to improve receiving water quality through better utilization of the sewer system
46 capacity.

47 Many different approaches have been developed for frequency analysis of sewer flooding
48 or CSO discharges in an urban drainage system, for example, analytical probability methods
49 (Benoist and Lijklema, 1989; Adams and Papa, 2000), Bayesian methods (Korving et al., 2002),
50 first-order reliability methods (Thorndahl and Willems, 2008), and imprecise probability
51 methods (Fu et al., 2011). In these methods, the historical rainfall series available are separated
52 into rainfall events, and probability distributions of some rainfall variables are then used to
53 characterise the stochastic nature of rainfall. For example, rainfall depth and duration are often
54 used in the literature (e.g., Vandenberghe et al., 2010; Zegpi and Fernández, 2010; Fu et al.,
55 2011).

56 In many cases, rainfall variables are related, however, due to the difficulty and
57 complexity in generating the joint probability distributions of rainfall variables, the dependence
58 structure between rainfall variables is not explicitly considered in many studies (e.g., Adams and
59 Papa, 2000; Thorndahl and Willems, 2008; Andres-Domenech et al., 2010). Research has shown
60 that the assumption of independence can have a significant effect on the frequency distributions
61 of flood or CSO discharges and may lead to erroneous results (Benoist and Lijklema, 1989).
62 Thus many efforts have been made to consider the correlation relationships between rainfall
63 variables (Córdova and Rodríguez-Iturbe, 1985; Yue, 2000) and to analyse the implications for
64 hydrologic design (Kao and Govindaraju, 2007b).

65 Most recently, there is increasing attention on the use of copulas as a flexible tool to
66 quantify the dependence structure between correlated variables in the fields of hydrology and
67 water engineering (e.g., De Michele and Salvadori, 2003; Kao and Govindaraju, 2007a; Zhang
68 and Singh, 2007; Zegpi and Fernández, 2010; Vandenberghe et al., 2010 and 2011). The use of
69 copulas enables to model the probabilistic dependence structure, independently of marginal

70 distributions, and thus allows for multivariate random events to be described using different
71 types of marginal distributions. This represents a significant advantage compared to conventional
72 multivariate analysis as many variables from hydrological phenomena cannot be described using
73 the same type of probability distributions. An important application of copulas is modelling the
74 stochastic nature of rainfall and flood using historical data (Favre et al., 2004; Vandenberghe et
75 al., 2010). Copulas also provide a convenient way to generate samples of correlated rainfall
76 variables, thus they can be used for flood frequency analysis in conjunction with the Monte Carlo
77 simulation method (e.g., Kao and Govindaraju, 2007a; Fontanazza et al., 2011).

78 The primary aim of this paper is to investigate the use of copulas for assessing the
79 hydraulic performance of a combined sewer system in an urban catchment, which explicitly
80 capture the dependence structure between rainfall depth and duration. The hydraulic performance
81 of the sewer system is represented by the maximum water level over the ground surface (flood
82 depth) at critical manholes and the volume of CSO discharges during a rainfall event. The latter
83 can be used as a performance indicator for receiving water quality as long as its limitations are
84 understood (Lau et al., 2002), although recent research suggests the performance of a sewer
85 system can be better considered in the context of integrated urban wastewater systems (Rauch et
86 al., 2002; Fu et al., 2008; Fu et al., 2009). In this study, the dependence between rainfall depth
87 and duration is represented using the Archimedean copulas, and the Monte Carlo simulation
88 method is then used to generate synthetic rainfall events for system performance analysis. The
89 copula method is demonstrated using an urban drainage system in the UK, and the cumulative
90 probability functions (CDF) of flood depth at one critical node and CSO overflow volume are
91 calculated. The results show the suitability and flexibility of the Archimedean copulas in

92 simulating the dependence of rainfall depth and duration, and the impacts of dependence
93 structure on the performance of urban drainage systems.

94 **2. Methodology**

95 **2.1. Concept of copulas**

96 Copulas can be described as multivariate CDFs with standard uniform marginals and represents
97 the dependence structure of random variables. For two random variables X and Y , their marginal
98 cumulative distribution functions are represented by

$$99 \quad u = F_X(x) \text{ and } v = F_Y(y) \quad (1)$$

100 where u and v are uniformly distributed random variables and $u, v \in [0,1]$. The joint CDF

101 $H_{XY}(x, y) = P(X \leq x, Y \leq y)$ describes the probability of two events: $X \leq x$ and $Y \leq y$. The

102 bivariate CDF $H_{XY}(x, y)$ can be represented as

$$103 \quad H_{XY}(x, y) = C(u, v) \quad (2)$$

104 where $C(u, v)$ is called a copula and can be uniquely determined when u and v are continuous.

105 Through Eq. (2), it is easy to see that the copula is actually a multivariate distribution function

106 with uniform marginals (Nelsen, 2006). This provides two main advantages in determining

107 $H_{XY}(x, y)$: (1) the marginals can be determined using different distributions, and (2) the

108 dependence structure can be described separately from the marginals, which allows for building

109 complex multivariate distributions to model stochastic phenomena such as rainfall without the

110 knowledge of marginal distributions.

111 There are many families of copulas that represent different dependence structures. The

112 one-parameter Archimedean copulas are of special interest for hydrologic analyses, and the

113 general expression of Archimedean copulas can be written as

114
$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) \quad (3)$$

115 where φ , called a generator, is a convex decreasing function defined in $[0, 1]$, satisfying $\varphi(1) = 0$
116 and $\lim_{t \rightarrow 0} \varphi(t) = \infty$. Using different forms of the function φ , different families of Archimedean
117 copulas can be generated, for example, the Gumbel, Frank and Clayton families. These copulas
118 can describe a wide range of dependence level, from negative to positive, and have been used to
119 describe the rainfall characteristics in previous studies (e.g., Kao and Govindaraju, 2007b; Zhang
120 and Singh, 2007) and thus are selected to describe the relationship between rainfall depth and
121 duration for the case study catchment in this study.

122 Only recently, use of copulas in hydrology has gained substantial attention with an intent
123 of describing the probabilistic structures of random variables such as rainfall and flood. The
124 work by De Michele and Salvadori (2003) was perhaps the first application in hydrology, and
125 they used the Frank family of Archimedean copulas to describe the dependence between rainfall
126 intensity and duration. Favre et al. (2004) used the Archimedean copulas to analyse the joint
127 distribution of flood peak flow and volume in two Canadian river catchments. Zhang and Singh
128 (2007) compared several different Archimedean copulas including Gumbel and Frank families to
129 simulate the joint distributions between rainfall intensity, depth and duration. Vandenberghe et
130 al. (2010) applied a number of copulas to investigate the dependence structure of storm variables
131 on the basis of a 105-year rainfall series. There are few applications to urban drainage systems
132 with one exception of the work by Fontanazza et al. (2011), which applied the copula approach
133 to generate synthetic rainfall events for urban flood estimation but focused on analysing the
134 impacts of hyetographs. The work described in this paper will look at the impacts of different
135 copulas on the frequency of sewer flooding and CSO overflow in the urban drainage system.

136 **2.2. Copula fitting**

137 For Archimedean copulas, the simplest method to estimate the parameter θ is through a
138 concordance measurement - Kendall's τ - which is a rank correlation coefficient, defined to
139 measure the orderings of two measured quantities. Kendall's τ is defined in the interval $[-1, 1]$,
140 where 1 represents total concordance, -1 represents total discordance, and 0 represents zero
141 concordance. According to the work by Nelsen (2006), the relationship between parameter θ
142 and Kendall's τ can be determined for the three Archimedean families.

143 Particularly, a closed-form expression can be derived for Clayton and Gumbel families.

144 In addition to the non-parametric method describe above, there are some parametric
145 methods available for parameter estimation, such as the conventional Maximum Likelihood
146 (ML) method, Inference Function for Margins (IFM) method (Joe, 1997) and Canonical
147 Maximum Likelihood (CML) method (Genest et al., 1995), and Minimum Distance Methods.
148 For more information, the reader is referred to the following studies (e.g., Genest and Favre,
149 2007; Chowdhary et al., 2011; Nazemi and Elshorbagy, 2012). The IFM method was used in this
150 study as it has a better performance compared with others according to our preliminary tests.
151 More importantly, it allows to explore the impacts of the choice of parametrically estimated
152 marginal CDFs on copula fitting as prior research has shown that a number of marginal CDFs
153 may not be rejected for rainfall variables under several statistical tests (Fu et al., 2005). The root
154 mean square error is a good indicator of goodness of fit, and can be calculated as:

$$155 \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \{C(u_i, v_i) - C_n(u_i, v_i)\}^2} \quad (4)$$

156 where C_n is the empirical copula. For this measure, the smaller the values, the better the copula
157 fits the data. Formal hypothetical tests are increasingly used to evaluate the goodness-of-fit of
158 different copulas (e.g., Berg, 2009; Genest et al., 2009; Nazemi and Elshorbagy, 2012). The

159 Cramér-von Mises statistic is chosen to compare an estimated copula C with the empirical copula
 160 C_n :

$$161 \quad T_n = n \int_{[0,1]^2} \{C(u,v) - C_n(u,v)\}^2 dC(u,v) = \sum_{i=1}^n \{C(u_i, v_i) - C_n(u_i, v_i)\}^2 \quad (5)$$

162 The p -values for T_n are approximated using the bootstrapping method described by Berg (2009)
 163 and Genest et al. (2009). 10,000 random samples are used in this study according to Genest and
 164 Favre (2007). Higher p -values are desired as they represent higher suitability of the chosen
 165 copulas.

166 In addition to the statistics, graphical methods can be used to verify the appropriateness
 167 of a fitted copula by comparison to the empirical distribution. The empirical distribution can be
 168 estimated using a nonparametric (empirical) approach (e.g., Genest and Rivest, 1993; Kao and
 169 Govindaraju, 2007a; Zhang and Singh, 2007): (1) assume an intermediate random variable z
 170 whose samples can be transformed from the n observations

$$171 \quad z_i = \frac{1}{n} \{\text{number of } (x_j, y_j) \text{ such that } x_j < x_i \text{ and } y_j < y_i\}, i=1, 2, \dots, n \quad (6)$$

172 (2) estimate the empirical distribution K_n using

$$173 \quad K_n(t) = \frac{1}{n} \{\text{number of } z_i \text{ such that } z_i < t\}, i=1, 2, \dots, n \quad (7)$$

174 The theoretical distribution of Archimedean copulas, i.e., $K(t) = P[C(u,v) \leq t]$, can be derived
 175 using the generating function

$$176 \quad K(t) = t - \frac{\varphi(t)}{\varphi'(t)}, \quad 0 < t \leq 1 \quad (8)$$

177 The distribution functions $K(t)$ and $K_n(t)$ transform the two dimension into one dimension,
 178 allowing for visual comparison of the empirical and theoretical copulas. The distribution plots or

179 Q-Q plots can be plotted for examination. This can provide a good indication if the theoretical
180 copula fits the data well.

181 Tail dependence analysis is critical to investigate the magnitude of dependence in the
182 upper and lower tails of a bivariate distribution (e.g., Poulin et al., 2007). It also helps identify
183 the most suitable copula by emphasising on the joint occurrence of extreme values (Nazemi and
184 Elshorbagy, 2012). The tail dependence can be represented by a coefficient. For the Gumbel
185 copula, the upper tail dependence coefficient is

$$186 \quad \lambda_U = 2 - 2^{1/\theta} \quad (9)$$

187 Amongst several estimators of the coefficient, the following estimator was proposed by Frahm
188 et al. (2005):

$$189 \quad \hat{\lambda}_U = 2 - 2 \exp \left[\frac{1}{n} \sum_{i=1}^n \log \left(\sqrt{\log \frac{1}{U_i} \log \frac{1}{V_i}} / \log \frac{1}{\max(U_i, V_i)^2} \right) \right] \quad (10)$$

190 where (U_i, V_i) ($i=1, 2, \dots, n$) are random samples generated from a copula. This estimator has an
191 advantage that no parameters (threshold values) are required for calculation and thus are used in
192 this study.

193 **2.3. Monte Carlo sampling**

194 On the basis of the copula method, Monte Carlo simulations can be used to generate
195 samples for correlated random variables. A general procedure is to generate correlated pairs
196 (u, v) using a copula and then transform them to real values based on the inverse marginal CDFs.
197 This is normally conducted using the following steps (Kao and Govindaraju, 2007a): (1) generate
198 the independently uniformly distributed random pairs (u, t) ; (2) solve v through a simplified
199 expression with copulas $P[V \leq v | U = u] = \frac{\partial C(u, v)}{\partial u} = t$, and thus obtain correlated pairs (u, v) ; (3)

200 assume $F_X(x)=u$ and solve x using the inverse function of X , and similarly assume $F_Y(y)=v$
201 and solve y . In this way, the correlated random samples (x, y) can be generated from their
202 margins (u, v) .

203 **3. Case study**

204 **3.1. The catchment**

205 An urban drainage system in a Scotland town is used to demonstrate the copula methodology
206 developed in this paper. The total catchment area is about 200 hectares, serving a population of
207 4,000. A combined sewer system provides the infrastructure for draining the catchment runoff
208 and routing the wastewater from the town to the treatment facilities. The sewer system consists
209 of 265 nodes, 265 pipes, 2 outfalls and 1 weir, and has a total conduit length of 22,482 metres.
210 The pipe gradients vary from 0.0001 to 0.0439. Flows are diverted downstream via two outfalls:
211 one is connected to a wastewater treatment works and the other to a combined sewer overflow,
212 and both flows are eventually discharged into a river. Fig. 1 shows a satellite image of the study
213 area and provides the layout of the sewer system described.

214 The storm water management model (SWMM), developed by the U.S. Environmental
215 Protection Agency, was used for hydrologic simulation of rainfall-runoff in the urban catchment
216 and for hydrodynamic simulation of in-sewer transport through the urban drainage system. The
217 system model has been well calibrated for the purpose of flood evaluation in the work by
218 Fullerton (2004), and has been used for system design and uncertainty analysis (Fu et al., 2011;
219 Sun et al., in revision).

220

221 **3.2. Rainfall data**

222 A 10-year time series of 5-minute rainfall from one rain gauge station is used in this study. To
223 analyse statistical characteristics of the actual rainfall events in the case study of urban
224 catchment, independent rainfall events are separated from the time series using the concept of
225 inter-event time (IETD) definition, i.e., the time interval between two consecutive events should
226 be no less than a pre-determined IETD. According to the maximum concentration time of the
227 catchment, IETD was set to 20 minutes such that the runoff response from an individual event is
228 not affected by any other. A total of 3405 events were identified from the rainfall series. As the
229 events with a total amount of 3mm cannot generate sewer flooding and CSO discharges, so they
230 are not considered for analysis in this study, and the number of events is reduced to 570, with an
231 average of 57 rainfall events per year.

232 The general characteristics of rainfall events are described by rainfall depth and duration.
233 Fig. 2 shows the scatter plots of the 570 events with marginal histograms of rainfall depth and
234 duration. There is a high frequency of low rainfall depth, about 50% of the rainfall events have a
235 very small depth less than 6 mm. The average rainfall depth is 7.4 mm, but the maximum is up
236 to 42 mm. Similarly, most events last a short period, although about 10 % have a duration over
237 800 minutes. The average rainfall duration over the 570 events is 399 minutes with a maximum
238 of 1725 minutes. The two variables are related to some extent, with a Kendall's tau = 0.27. As
239 can be seen from the marginal histograms, the two variables follow a rather different marginal
240 distribution. This clearly demonstrates the need to separate the marginal distributions and
241 dependence structure in the joint distribution of the two rainfall variables so that the marginal
242 distributions can be simulated by different types of distribution.

243

244 4. Results and discussion

245 4.1. Marginal distributions

246 The following commonly used distributions are used to fit the rainfall depth and duration
247 data: Generalized Pareto (GP), Generalized Extreme Value (GEV), Log-Logistic (Log-log) and
248 Gamma, according to previous studies (e.g., Kao and Govindaraju, 2007b; Vandenberghe et al.,
249 2011). The above functions are fitted using the maximum likelihood estimation method that
250 maximises the log-likelihood function. In the calculation, the maximum number of iterations is
251 specified as 100 and the accuracy of the estimation is set to 1.0×10^5 .

252 Three goodness-of-fit tests, i.e., Kolmogorov Smirnov (K-S), Anderson Darling (A-D)
253 and Chi-Square (χ^2) tests, are used to determine if the data follow one of the specified
254 distributions (the null hypothesis H_0). The hypothesis is evaluated at the 5% significance level.
255 The critical values at this level are 0.056 for K-S, 16.919 for χ^2 , and 2.502 for A-D, respectively.
256 Smaller statistic values indicate better fit to the data. The hypothesis regarding a specific
257 distribution is rejected at the significance level if the test statistic (K_n , A_n^2 , or χ_n^2) is greater than
258 the relevant critical value given above. Appendix A provides the CDFs and goodness-of-fit tests
259 used.

260 The statistics of K-S (K_n), A-D (A_n^2) and Chi-Square (χ_n^2) for rainfall duration and
261 depth are provided in Tables 1 and 2, respectively. The p -values measure the amount of
262 information that is against the null hypothesis H_0 , and are also provided here. The smaller the p -
263 values, the more evidence we have against H_0 . For rainfall duration, the three distributions, i.e.,
264 GEV, Log-log and Gamma, cannot be rejected with all of the three tests and have a decreasing
265 ranking according to the statistics values. Similarly, for rainfall depth, the distribution GP best
266 fits to the data, followed by Log-log and GEV. All of the three distributions cannot be rejected

267 with all of the three tests. According to the A-D test, Gamma and GP are rejected for rainfall
268 duration and depth, respectively. The A-D test is stricter than the K-S test possibly because it
269 gives more weight to the distribution tails. The results confirm that in many cases it is not
270 possible to determine one single best distribution particularly when a relatively short series of
271 data is available (Korving et al., 2002; Kao and Govindaraju, 2007b; Fu et al., 2011). This study
272 considers all the distributions that cannot be rejected to investigate the bivariate distribution
273 using copulas.

274 **4.2. Dependence structure**

275 The selected marginal distributions for rainfall duration and depth are used to fit the
276 Archimedean copulas using the CML method. The parametrically estimated values of parameter
277 θ are provided in Table 3, along with their 95% confidence intervals. Recall that parameter θ
278 can also be derived according to the relationships between θ and τ , and the values for Gumbel,
279 Frank and Clayton copulas are 1.371, 2.589 and 0.741, respectively. It can be seen that the
280 parametric estimates for Gumbel and Frank copulas are in a good agreement with those from the
281 non-parametric method, and are bracketed in the relevant 95% confidence intervals. For the
282 Clayton copula, however, the non-parametric estimate is significantly bigger than the parametric
283 estimates, and is out of the 95% confidence intervals. This is possibly because the rainfall data as
284 shown in Fig. 2 illustrate greater dependence in the upper tail than in the lower tail. On the
285 contrary, Clayton copula exhibits greater dependence in the lower tail than in the upper tail. The
286 Gumbel copula is an asymmetric Archimedean copula with greater dependence in the upper tail
287 than in the lower. The Frank copula is a symmetric Archimedean copula. Thus these two copulas
288 are more appropriate for describing the dependence structure between rainfall duration and
289 depth.

290 Table 4 shows the resulting RMSE and Cramér-von Mises statistic values. The statistic
291 confirms the inappropriateness of the Clayton copula as it has a lower p -value for most the
292 marginal combinations. This statistic shows a more significant difference compared with
293 RMSE. Amongst the three marginal distributions of rainfall depth, GP has the worst
294 performance in terms of copula fitting although this distribution is the best in the marginal
295 distribution fitting according to the statistics. This implies that it is important to consider the
296 goodness-of-fit of both marginal distributions and copulas in order to achieve the best overall
297 performance in constructing a joint distribution of multi-variables. For the three distributions of
298 rainfall duration, there is no significant difference in copula fitting. It can be seen that the
299 Gumbel and Frank copulas are in good agreement.

300 Table 5 shows the upper tail dependence coefficients for Gumbel copulas. It can be seen
301 that the estimated coefficient values are very close to the theoretical ones. The estimator
302 proposed by Frahm et al. (2005) has a high accuracy. More importantly, this implies that the
303 choice of marginal distributions has no impacts on tail dependence, which is mainly controlled
304 by copula selection as expected.

305 Fig. 3 shows the Q-Q plots of Gumbel and empirical copulas for different marginal
306 distribution combinations. The x -axis represents the cumulative probability of empirical copula
307 and y -axis represents that of Gumbel copula. The diagonal straight line represents a perfect
308 match between the parametrically estimated copula and empirical copula. Generally the Q-Q
309 plots confirm the results revealed from the statistic values in Table 4. That is, the GP vs. GEV
310 and GP vs. Gamma pairs provide the worst copula fitting results, while all the other distribution
311 combinations provide a rather good fitting. The pair Log-log vs. GEV is chosen as the base case

312 to analyse the frequency of flooding and CSO discharges and compare the impacts of marginal
313 distributions and copulas.

314 To understand the structure of dependence, Fig. 4 visualizes the CDF and probability
315 distribution function (PDF) of the Gumbel copula on the basis of the Log-log vs. GEV
316 combination. The variables u and v represent the transformed random variables X and Y
317 (rainfall depth and duration) in the unit hypercube, respectively, and have the same ranks as X
318 and Y . Fig. 4a shows the fitted copula (shaded surface) together with the empirical copula
319 (points). The strong dependence in the upper tail is clearly illustrated in Fig. 4b.

320 Selection of the most suitable copula is a complex process and need to consider several
321 different measures including statistics, graphical approaches, and comparison to empirical
322 copulas and data regarding dependence types. A single measure may fail to identify the
323 inappropriate copulas, leading to an overestimate or underestimate of the probability of sewer
324 flood and CSO discharges as demonstrated below.

325 **4.3. Flood and overflow frequency**

326 The theoretically fitted Gumbel copula was used to generate a large set of 10,000 samples for
327 rainfall depth and duration. The number of samples used here is very conservative compared to
328 the previous study by Fu et al. (2011) and can provide very stable simulation results. The
329 synthetic rainfall events were produced by applying a rectangular pulse with duration as the
330 width and average rainfall intensity as the height, and they were then used as inputs to the sewer
331 system model to calculate the flood depth at different nodes. We recognise the impact of
332 different rainfall profiles on the frequency of flood and overflow (Fontanazza et al., 2011; Sun et
333 al., 2012), but this is out of scope of this study.

334 Fig. 5 shows the cumulative probabilities of flood depth at one critical node N126 and
335 overflow volume at the CSO. For the copula results, the dependence structure is represented by
336 the Gumbel copula and marginal distributions for rainfall depth and duration are represented by
337 Log-Log and GEV, respectively. In Fig. 5a, the probability of no flooding occurring (flood
338 depth=0) has a value of 0.43. In other words, the probability of flooding at this node is 0.57 and
339 this is equivalent to the probability of system ‘failure’ in terms of sewer flooding. Note that this
340 probability represents the probability for each rainfall event because the way of rainfall events is
341 simulated in this study. This high number of ‘failures’ at this critical node is caused by the
342 expansion of the network to the (left) upstream due to urban development (Fullerton, 2004).
343 Similarly in Fig. 5b, the probability of no CSO discharges (CSO volume=0) is estimated at 0.92.

344 For comparison, Fig. 5 also shows the results when rainfall depth and duration are
345 assumed as independent. In this case, the probability of sewer flooding, having a value of 0.47, is
346 lower than in the case of correlation. This implies that the flood depth is under-estimated without
347 considering the correlation between rainfall depth and duration. This under-estimation has a
348 more significant impact on flood depth when compared with the uncertainties in the fitted copula
349 parameter, as demonstrated with the 95% confidence intervals as shown in Fig. 5. For CSO
350 volume, the differences between correlation and independence are also significant, but mainly lie
351 in the regions of high cumulative probabilities. This is because the probability of CSO discharge
352 is much lower than sewer flooding, that is, CSO discharges can only results from more ‘extreme’
353 rainfall events.

354 **4.4. Impacts of copulas and marginal distributions**

355 Recall that different marginal distributions for rainfall variables cannot be rejected with the
356 statistical tests in this case study. To investigate the impacts of different marginal distributions,

357 Fig. 6 shows the CDFs of flood depth and CSO volume from three marginal distributions of
358 rainfall duration, i.e., GEV, Log-log and Gamma, combined with the Log-log distribution of
359 rainfall depth. For flood depth, the three CDFs are roughly the same, which implies that the
360 impacts of the different marginals are negligible when compared with other uncertainties such as
361 the copula parameter estimation, as shown in Fig. 5. For CSO volume, similarly the impacts of
362 different marginals are small, although there are some differences in the upper tails, reflecting
363 the importance of distribution tails in rainfall event simulation.

364 Different copulas were compared for the CDFs of flood depth and CSO volume and
365 results are shown in Fig. 7. Note that according to the copula fitting results the Clayton copula is
366 not appropriate to describe the dependence structure of rainfall depth and duration, but it is used
367 here for the purpose of demonstration of its potential impacts. The Clayton copula overestimates
368 the probability of sewer flooding and CSO discharge, i.e., the performance of the sewer system.
369 This can be explained by the dependence structure of the Clayton copula: greater dependence in
370 the lower tail than in the upper tail, which results in more small synthetic rainfall events.
371 Conversely, the cumulative probabilities of sewer flooding and CSO discharges estimated by the
372 Gumbel copula are smaller than those from Clayton and Frank copulas, because more extreme
373 events are generated as a result of the greater dependence in the upper tails.

374 Clearly the copulas have more significant impacts on the CDFs of sewer flooding and
375 CSO volume than the marginal distributions, when comparing the results in Fig. 5 and Fig. 6.
376 This implies the importance of considering the dependence structure of rainfall variables when
377 evaluating the system performance of urban drainage systems through synthetic events based
378 methods.

379 The impact of dependence structure of rainfall variables on system performance can be
380 illustrated by calculating the return periods of the sewer system. Fig. 8 shows the return periods
381 of CSO volumes from the three copulas. The return periods of CSO volumes are derived using
382 the cumulative probabilities in Fig. 7b, considering the average 57 rainfall events per year.

383 **5. Conclusions**

384 This paper highlights the importance of considering the dependence structure of rainfall
385 variables in the context of system performance of urban drainage systems using copulas. The
386 copula method is demonstrated using an urban drainage system in the UK to calculate the
387 cumulative probability distributions of flood depth and CSO volume. The rainfall characteristics
388 in the case study are represented by two variables: rainfall depth and duration. The marginal
389 distributions of these variables are simulated using GP, GEV, Log-log and Gamma, and the
390 dependence structure is represented by the following three one-parameter Archimedean copula
391 families: Gumbel, Frank, and Clayton. On the basis of the copula approach, the Monte Carlo
392 simulation is used to generate synthetic rainfall events to evaluate the probabilistic system
393 performance, represented by the CDFs of flood depth and CSO volume. This new methodology
394 is promising in that it provides a simpler way to construct the joint distribution for rainfall
395 variables by separating the dependence from their marginal distributions, and thus provides a
396 basis for performance evaluation of urban drainage systems. The following conclusions are
397 presented on the basis of this study:

- 398 1. It is necessary to consider all the marginal distributions that cannot be rejected by
399 statistical tests for copula fitting using the IFM method, rather than choose the best
400 ranked distributions. As revealed by the case of bivariate copulas, the pair of the best

401 fitted marginal distributions of rainfall depth and duration cannot produce the best overall
402 performance in constructing the joint distribution of rainfall depth and duration.

403 2. Copula identification should be based on several different measures including statistics,
404 graphical approaches, and comparison to empirical copulas and data regarding
405 dependence types. A single measure may fail to identify the inappropriate copulas,
406 leading to an overestimate or underestimate of the probability of sewer flood and CSO
407 discharges.

408 3. The results obtained show the copulas, i.e., the dependence structures, have more
409 significant impacts on the CDFs of sewer flooding and CSO volume than the marginal
410 distributions. Different copulas affect different parts of the CDFs of sewer flooding and
411 CSO volume, i.e., those with higher or lower return periods.

412 The copula method has the flexibility and advantage in building complex, bivariate
413 probability distributions of rainfall depth and duration for system performance analysis. The
414 results provide a more accurate probabilistic evaluation of sewer flooding and CSO discharges
415 based on the characterization of the dependence structure of rainfall depth and duration. This
416 provides crucial information for more accurate estimation of design storms and the associated
417 risks.

418

419 **Appendix A: Cumulative distribution functions and test statistics**

420 The marginal distribution functions used in this paper are given in equations (A.1)-(A.4).

421 Generalized Extreme Value Distribution (GEV)

$$422 \quad F(x) = \exp\left(-\left[1 + \xi(x - \mu)/\sigma\right]^{-1/\xi}\right) \quad (\text{A.1})$$

423 where $\xi \neq 0$, $\sigma > 0$ and μ are shape, scale and location parameters, respectively.

424 Generalized Pareto Distribution (GP)

$$425 \quad F(x) = 1 - [1 + \xi(x - \mu)/\sigma]^{-1/\xi} \quad (\text{A.2})$$

426 where $\xi \neq 0$, $\sigma > 0$ and μ are shape, scale and location parameters, respectively.

427 Log-Logistic Distribution (Log-log)

$$428 \quad F(x) = [1 + (\beta/x)^\alpha]^{-1} \quad (\text{A.3})$$

429 where $\alpha > 0$ and $\beta > 0$ are parameters.

430 Gamma Distribution

$$431 \quad F(x) = \Gamma_{(x-\gamma)/\beta}(\alpha) / \Gamma(\alpha) \quad (\text{A.4})$$

432 where $\alpha > 0$, $\beta > 0$ and γ are shape, scale and location parameters, respectively. Γ is the Gamma

433 function

$$434 \quad \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt \quad \text{for } (\alpha > 0) \quad (\text{A.5})$$

435 and Γ_z is the incomplete Gamma function

$$436 \quad \Gamma_z(\alpha) = \int_0^z t^{\alpha-1} e^{-t} dt \quad \text{for } (\alpha > 0) \quad (\text{A.6})$$

437 The goodness of fit for the above distributions is considered using three different test statistics:

438 Kolmogorov-Smirnov K_n , Anderson-Darling A_n^2 and Chi-Square χ_n^2 :

$$439 \quad K_n = \max_{1 \leq i \leq n} \left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right) \quad (\text{A.7})$$

$$440 \quad A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln F(x_i) + \ln(1 - F(x_{n-i+1}))] \quad (\text{A.8})$$

$$441 \quad \chi_n^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (\text{A.9})$$

442 where n is the number of data, x_i is the i th sample ($i=1, 2, \dots, n$), and F is the cumulative
443 distribution being tested. As the Chi-Square test is based on binned data, the total number of bins
444 k is determined using the following empirical equation:

$$445 \quad k = 1 + \log_2 n \quad (\text{A.10})$$

446 O_i is the observed frequency for bin i and E_i is the expected frequency for bin i calculated by

$$447 \quad E_i = F(x_2) - F(x_1) \quad (\text{A.11})$$

448 where x_1 and x_2 are the lower and upper limits for bin i .

449 **References**

- 450 Adams, B.J., Papa, F., 2000. *Urban Stormwater Management Planning With Analytical*
451 *Probabilistic Models*. 358 pp., John Wiley & Sons, West Sussex, U.K.
- 452 Andres-Domenech, I., Munera, J.C., Frances, F., Marco, J.B., 2010. Coupling urban event-based
453 and catchment continuous modelling for combined sewer overflow river impact assessment.
454 *Hydrol. Earth Syst. Sci.* 14 (10), 2057-2072.
- 455 Benoist, A.P. and Lijklema, L., 1989. A methodology for the assessment of frequency
456 distributions of combined sewer overflow volumes. *Water Res.* 23(4), 487-493.
- 457 Berg, D. 2009. Copula goodness-of-fit testing: an overview and power comparison. *Eur J*
458 *Finance* 15(7 & 8): 675-701.
- 459 Butler, D., Davies, J.W., 2011. *Urban Drainage*, 3rd ed. Spon Press, London.
- 460 Chowdhary, H., Escobar, L., Singh, V.P., 2011. Identification of suitable copulas for bivariate
461 frequency analysis of flood peaks and flood volumes. *Hydrology Research* 42(2-3), 193-216.
- 462 Córdova, J. R., Rodríguez-Iturbe, I., 1985. On the probabilistic structure of storm surface runoff.
463 *Water Resour. Res.* 21 (5), 755– 763.

464 De Michele, C., and Salvadori, G., 2003. A generalized Pareto intensity-duration model of storm
465 rainfall exploiting 2-copulas. *J. Geophys. Res.* 108(D2), doi:10.1029/2002JD002534.

466 Favre, A., El Adlouni, S., Perreault, L., Thiémonge, N., Bobée, B., 2004. Multivariate
467 hydrological frequency analysis using copulas. *Water Resour. Res.* 40, W01101,
468 doi:10.1029/2003WR002456.

469 Fontanazza, C.M., Freni, G., La Loggia, G., Notaro, V., 2011. Definition of synthetic rainfall
470 events for urban flooding estimation: the integration of multivariate statistics and cluster
471 analysis. 12th International Conference on Urban Drainage, Porto Alegre, Brazil.

472 Frahm, G., Junker, M., and Schmidt, R., 2005. Estimating the tail dependence coefficient:
473 Properties and pitfalls. *Insur. Math. Econ.*, 37(1), 80-100.

474 Fu, G., Butler, D., Khu, S.T., 2008. Multiple objective optimal control of integrated urban
475 wastewater systems. *Environ. Model. Softw.* 23, 225-234.

476 Fu, G., Butler, D., Khu, S.T., 2009. The impact of new developments on river water quality from
477 an integrated system modelling perspective. *Sci. Total Environ.* 407(4), 1257-1267.

478 Fu, G., Butler, D., Khu, S.T., Sun, S., 2011. Imprecise probabilistic evaluation of sewer flooding
479 in urban drainage systems using random set theory. *Water Resour. Res.* W02534, 47,
480 doi:10.1029/2009WR008944.

481 Fullerton, J.N., 2004. *A simulated modelling approach for storm water flow optimisation*, PhD
482 thesis, Centre for Water Systems, University of Exeter, UK.

483 Genest, C., Favre A.C., 2007. Everything you always wanted to know about copula modelling
484 but were afraid to ask. *J Hydrol Eng* 12(4), 347-368.

485 Genest, C., K. Ghouli, and L. P. Rivest, 1995. A semiparametric estimation procedure of
486 dependence parameters in multivariate families of distributions. *Biometrika* 82(3), 543- 552.

487 Genest, C., Rémillard, B., and Beaudoin, D., 2009. Goodness-of-fit tests for copulas: a review
488 and a power study. *Insurance: Mathematics and Economics* 44, 199-213.

489 Genest, C., Rivest, L., 1993. Statistical inference procedures for bivariate Archimedean copulas.
490 *Journal of the American Statistical Association*, 88 (424), 1034-1043.

491 Hvitved-Jacobsen, T., Yousef, Y.A., 1988. Analysis of rainfall series in the design of urban
492 drainage control systems. *Water Res.* 22 (4), 491-496.

493 Joe, H., 1997. *Multivariate models and dependence concepts*. Chapman and Hall, London.

494 Kao, S.C., Govindaraju, R.S., 2007a. Probabilistic structure of storm surface runoff considering
495 the dependence between average intensity and storm duration of rainfall events. *Water*
496 *Resour. Res.* 43, doi:10.1029/2006WR005564.

497 Kao, S.C., Govindaraju, R.S., 2007b. A bivariate frequency analysis of extreme rainfall with
498 implications for design. *J. Geophys. Res.*, 112, D13119, doi:10.1029/2007JD008522.

499 Korving, H., Clemens, F., van Noortwijk, J., van Gelder, P., 2002. Bayesian estimation of return
500 periods of CSO volumes for decision-making in sewer system management. In: *Proc. of 9th*
501 *International Conference on Urban Drainage*, Portland, Oregon, USA.

502 Lau, J., Butler, D., Schütze, M., 2002. Is combined sewer overflow spill frequency/volume a
503 good indicator of receiving water quality impact? *Urban Water* 4 (2), 181-189.

504 Nelsen, R.B., 2006. *An Introduction to Copulas*, Springer, New York.

505 Nazemi, A., Elshorbagy, A., 2012. Application of copula modelling to the performance
506 assessment of reconstructed watersheds. *Stochastic Environmental Research & Risk*
507 *Assessment* 26(2), 189-205.

508 Poulin, A., Huard, D., Favre, A.C., Pugin, S., 2007. Importance of tail dependence in bivariate
509 frequency analysis. *J. Hydrol. Eng.* 12(4), 394-403.

510 Rauch, W., Bertrand-Krajewski, J.L., Krebs, P., Mark, O., Schilling, W., Schütze, M.,
511 Vanrolleghem, P.A., 2002. Mathematical modelling of integrated urban drainage systems.
512 Water Sci. Technol. 45(3), 81-94.

513 Schmitt, T.G., Thomas, M., Etrich, N., 2004. Analysis and modeling of flooding in urban
514 drainage systems. J. Hydrol. 299 (3-4), 300-311.

515 Sun, S., Fu, G., Djordjevic, S., Khu, S.T., 2012. Separating aleatory and epistemic uncertainties:
516 probabilistic sewer flooding evaluation using probability box. J. Hydrol. 420-421, 360-372 .

517 Thorndahl, S., Willems, P., 2008. Probabilistic modelling of overflow, surcharge and flooding in
518 urban drainage using the first-order reliability method and parameterization of local rain
519 series. Water Res. 42(1-2), 455-466.

520 Vandenberghe, S., Verhoest, N.E.C., De Baets, B., 2010. Fitting bivariate copulas to the
521 dependence structure between storm characteristics: A detailed analysis based on 105 year
522 10 min rainfall. Water Resour. Res. 46, W01512, doi:10.1029/2009WR007857.

523 Vandenberghe, S., Verhoest, N.E.C., Onof, C., De Baets, B., 2011. A comparative copula-based
524 bivariate frequency analysis of observed and simulated storm events: A case study on
525 Bartlett-Lewis modeled rainfall. Water Resour. Res., 47, doi:10.1029/2009WR008388.

526 Yue, S., 2000. The Gumbel mixed model applied to storm frequency analysis. Water Resources
527 Management 14(5), 377-389.

528 Zegpi, M., Fernández, B., 2010. Hydrological model for urban catchments – analytical
529 development using copulas and numerical solution. Hydrol. Sci. J. 55(7), 1123-1136.

530 Zhang, L., Singh, V.P., 2007. Bivariate rainfall frequency distributions using Archimedean
531 copulas. J. Hydrol. 332, 93-109.

532 **Figure captions**

533 **Fig. 1 - Layout of the case study network.**

534 **Fig. 2 - Scatter plot of rainfall depth and duration with marginal histograms.**

535 **Fig. 3 - Q-Q plots of Gumbel and empirical copulas for different marginal distribution**
536 **combinations.**

537 **Fig. 4 - Three dimensional plots for the theoretically fitted Gumbel copula.**

538 **Fig. 5 - Cumulative probabilities of flood depth and CSO volume from Gumbel copula.**

539 **Fig. 6 – Impacts of different marginal distributions on the cumulative probabilities of flood**
540 **depth and CSO volume.**

541 **Fig. 7 – Comparison of Gumbel and Frank copulas.**

542 **Fig. 8 – Return periods of CSO discharge volumes.**

543

544

545

546

547

548 **Table 1 - Test Statistics of the fitted CDFs for rainfall duration* .**

Distribution	Distribution Parameter	K-S		χ^2		A-D
		K_n	p -value	χ_n^2	p -value	A_n^2
GEV	$\xi=0.105, \sigma=189.12, \mu=267.9$	0.026	0.832	3.126	0.959	0.433
Gamma	$\alpha=2.0653, \beta=193.05, \gamma=0$	0.034	0.509	8.421	0.492	0.423
Log-Log	$\alpha=2.872, \beta=386.08; \gamma=-54.80$	0.036	0.427	14.177	0.116	0.975
GP**	$\xi=-0.229, \sigma=399.22, \mu=73.82$	0.055	0.063	-	-	107.39

549 *The CDFs are provided in Appendix A.

550 **This distribution is rejected at the significant level of 5% with the A-D test.

551

552

553
554

Table 2 - Test statistics of the fitted CDFs for rainfall depth *

Distribution	Distribution Parameter	K-S		χ^2		A-D
		K_n	p-value	χ_n^2	p-value	A_n^2
GP	$\xi=0.190, \sigma=3.823, \mu=1.941$	0.020	0.910	5.393	0.80	0.496
Log-Log	$\alpha=1.375, \beta=2.413, \gamma=2.976$	0.046	0.176	14.894	0.094	1.804
GEV	$\xi=0.680, \sigma=1.746, \mu=4.594$	0.052	0.091	16.226	0.062	2.369
Gamma**	$\alpha=5.813, \beta=5.263, \gamma=3.0$	0.052	0.091	-	-	5.813

555 *The CDFs are provided in Appendix A.

556 **This distribution is rejected at the significant level of 5% with the A-D test.

557

558

559 **Table 3 - Estimated parameters of copulas and their 95% confidence intervals.**

Depth	Duration	Gumbel		Frank		Clayton	
	GEV	1.375	[1.285 1.466]	2.472	[1.971 2.974]	0.443	[0.317 0.570]
GEV	Log-log	1.399	[1.309 1.489]	2.412	[1.916 2.909]	0.393	[0.273 0.513]
	Gamma	1.372	[1.280 1.465]	2.546	[2.032 3.059]	0.413	[0.288 0.539]
	GEV	1.377	[1.280 1.465]	2.509	[1.997 3.021]	0.375	[0.257 0.494]
Log-log	Log-log	1.403	[1.284 1.470]	2.447	[1.940 2.953]	0.340	[0.227 0.453]
	Gamma	1.375	[1.311 1.496]	2.591	[2.067 3.114]	0.354	[0.237 0.471]
	GEV	1.382	[1.281 1.470]	2.845	[2.290 3.400]	0.507	[0.344 0.671]
GP	Log-log	1.403	[1.288 1.476]	2.447	[1.940 2.953]	0.340	[0.227 0.453]
	Gamma	1.385	[1.311 1.496]	3.063	[2.502 3.623]	0.473	[0.323 0.623]

560

561

562

Table 4 - Goodness-of-fit of copulas for different combinations of marginal distributions.

Depth	Duration	Gumbel			Frank			Clayton		
		RMSE	T_n	p -value	RMSE	T_n	p -value	RMSE	T_n	p -value
	GEV	0.016	0.139	0.372	0.014	0.110	0.534	0.016	0.146	0.388
GEV	Log-log	0.018	0.184	0.239	0.015	0.130	0.431	0.016	0.142	0.394
	Gamma	0.017	0.173	0.268	0.019	0.196	0.229	0.020	0.237	0.162
	GEV	0.012	0.078	0.726	0.013	0.095	0.622	0.020	0.240	0.152
Log-log	Log-log	0.014	0.114	0.491	0.014	0.118	0.488	0.021	0.241	0.150
	Gamma	0.015	0.128	0.427	0.018	0.193	0.234	0.025	0.345	0.070
	GEV	0.065	2.397	0	0.070	2.816	0	0.060	2.030	0
GP	Log-log	0.014	0.114	0.488	0.014	0.118	0.488	0.021	0.241	0.156
	Gamma	0.066	2.521	0	0.075	3.223	0	0.060	2.083	0

564

565

566

567

568

569

570

571

572

573

574

575

576

577 **Table 5 – The upper tail dependence coefficients for Gumbel copulas.**

Depth	GEV			Log-log			GP		
Duration	GEV	Log-log	Gamma	GEV	Log-log	Gamma	GEV	Log-log	Gamma
λ_U	0.344	0.359	0.343	0.346	0.361	0.344	0.349	0.361	0.351
$\hat{\lambda}_U$	0.344	0.363	0.345	0.349	0.361	0.346	0.347	0.360	0.350

578

579

580

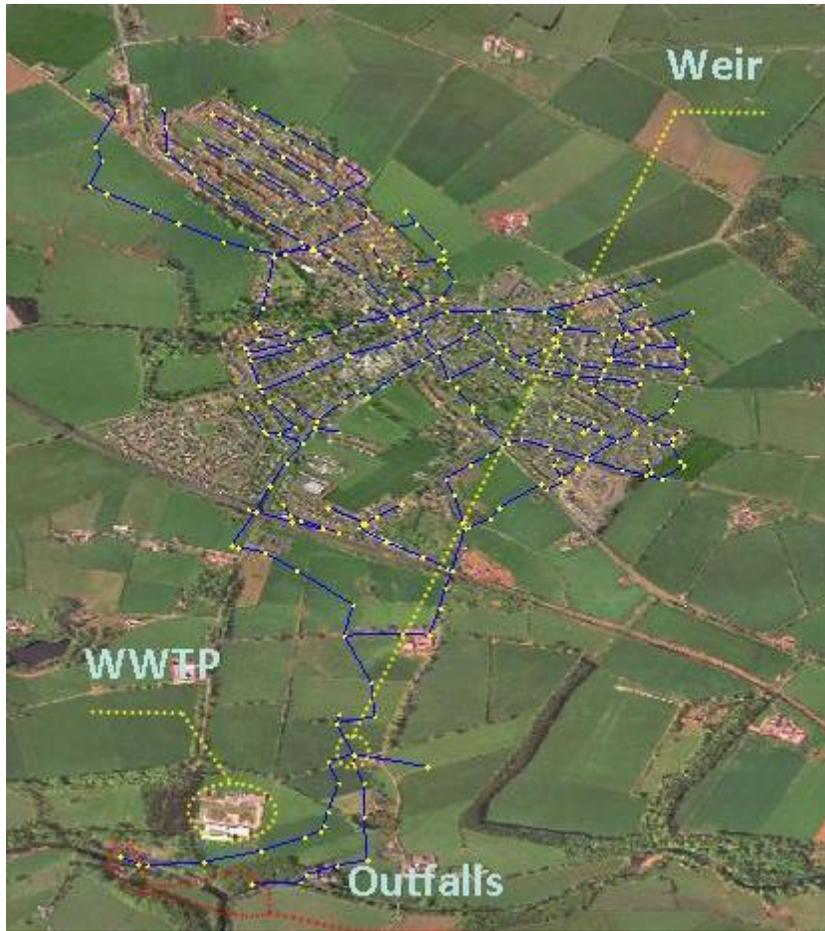
581

582

583

584

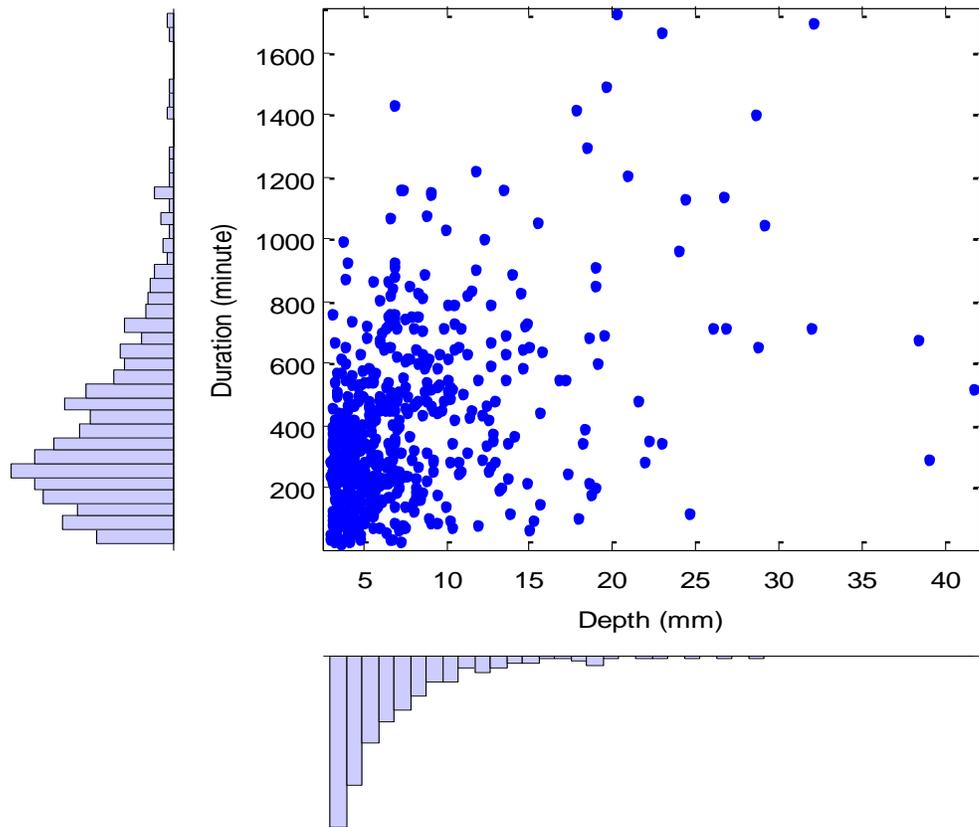
585



586

587

Fig. 1 - Layout of the case study network.

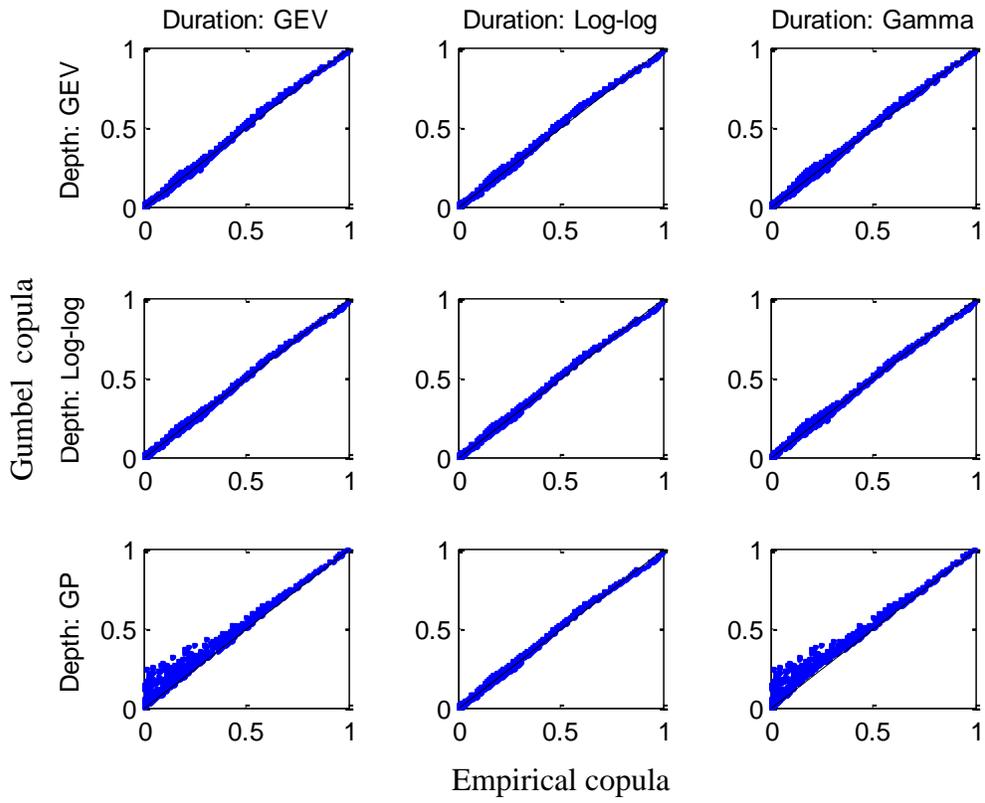


588

589

Fig. 2 - Scatter plot of rainfall depth and duration with marginal histograms.

590



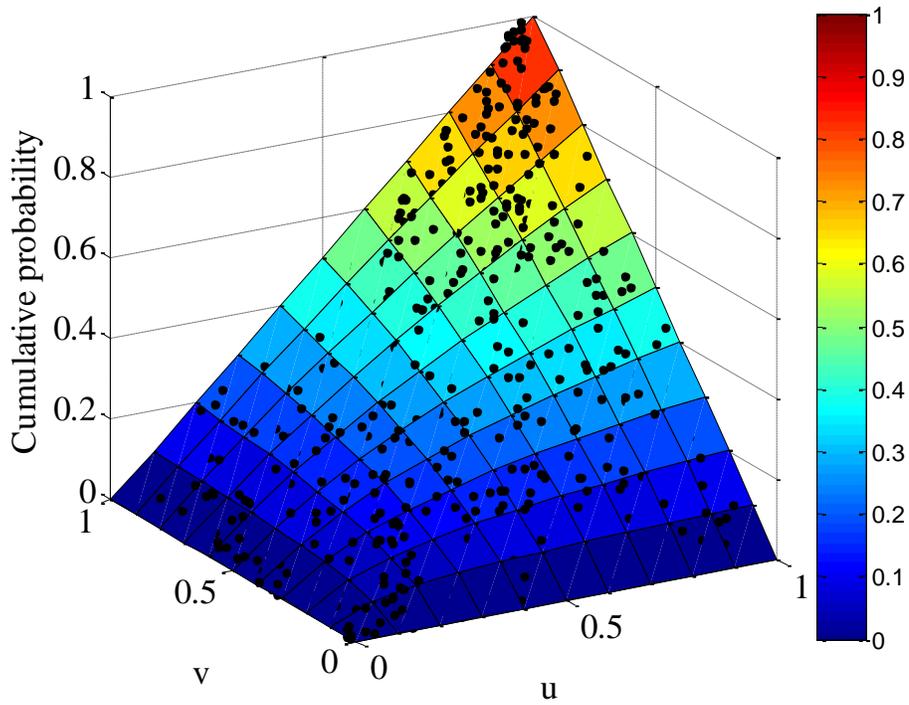
591

592 **Fig. 3 - Q-Q plots of Gumbel and empirical copulas for different marginal distribution**

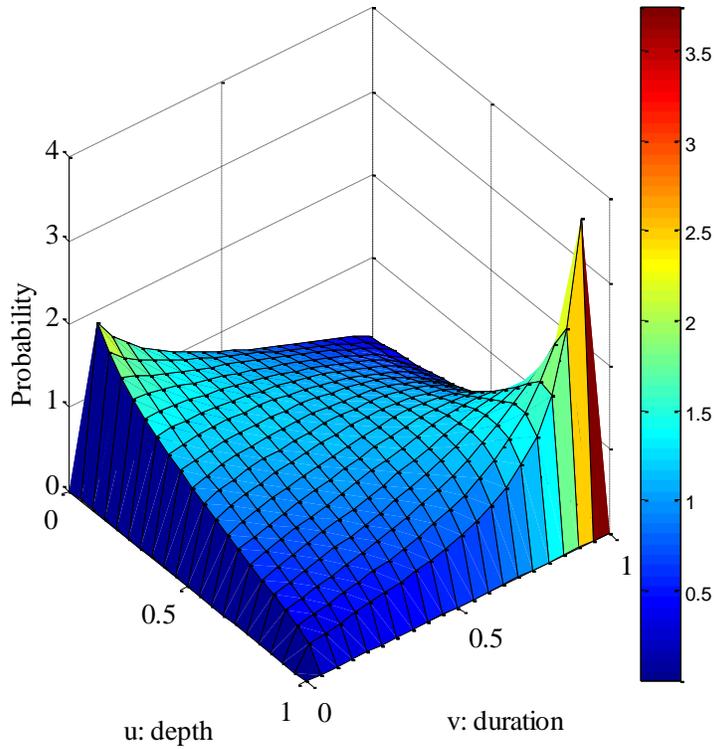
593

combinations.

594



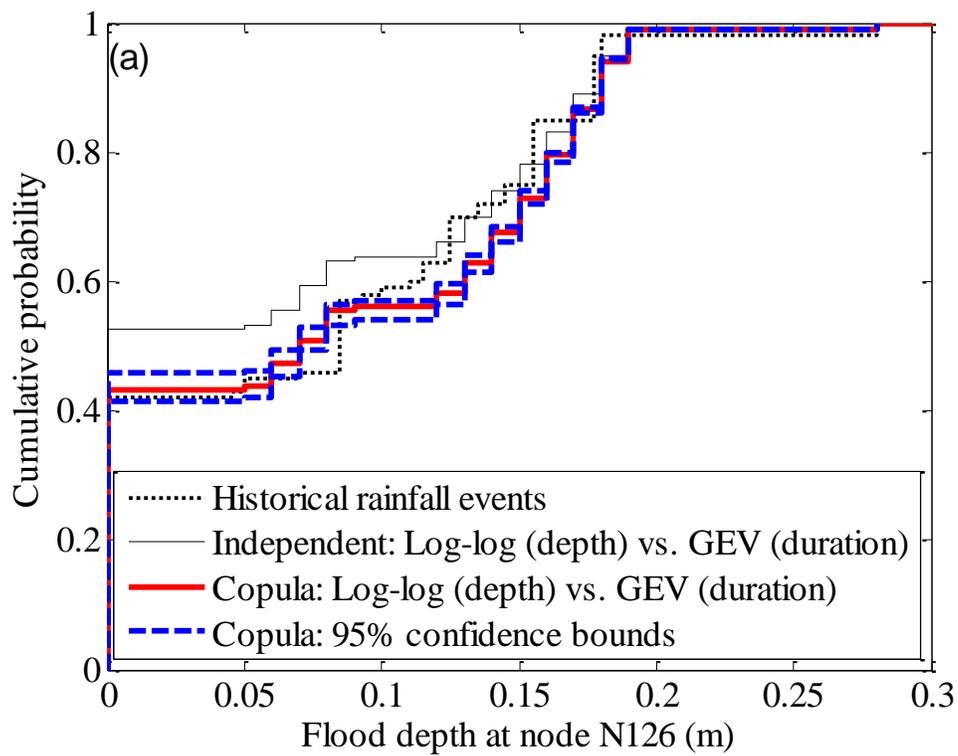
595



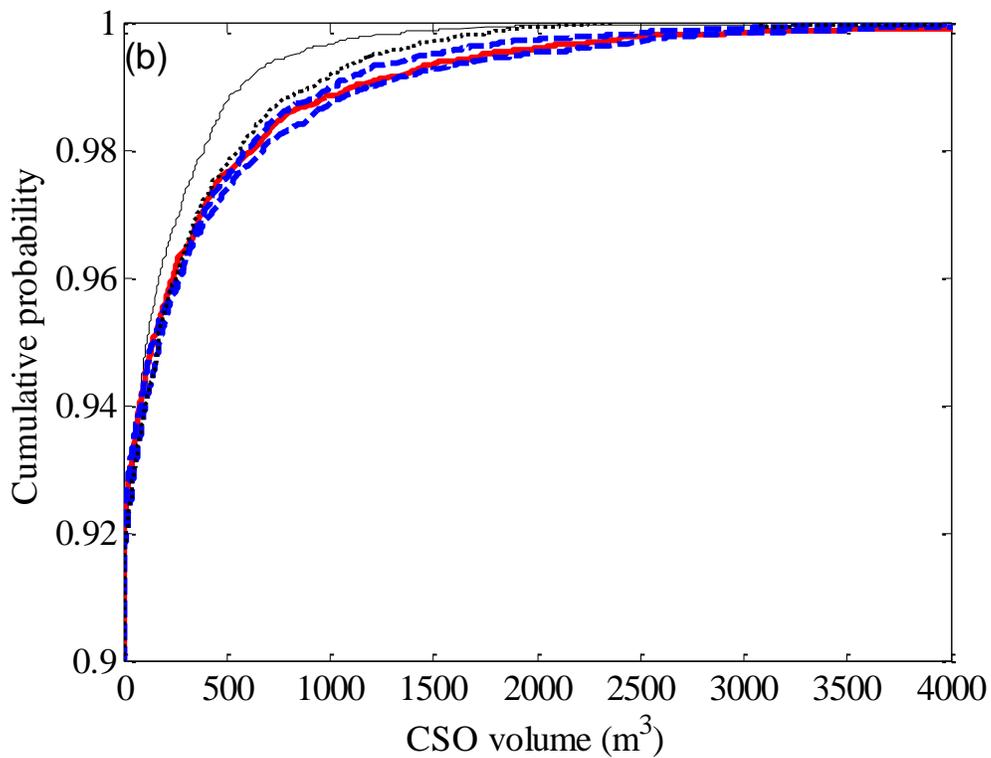
596

597

Fig. 4 - Three dimensional plots for the theoretically fitted Gumbel copula.



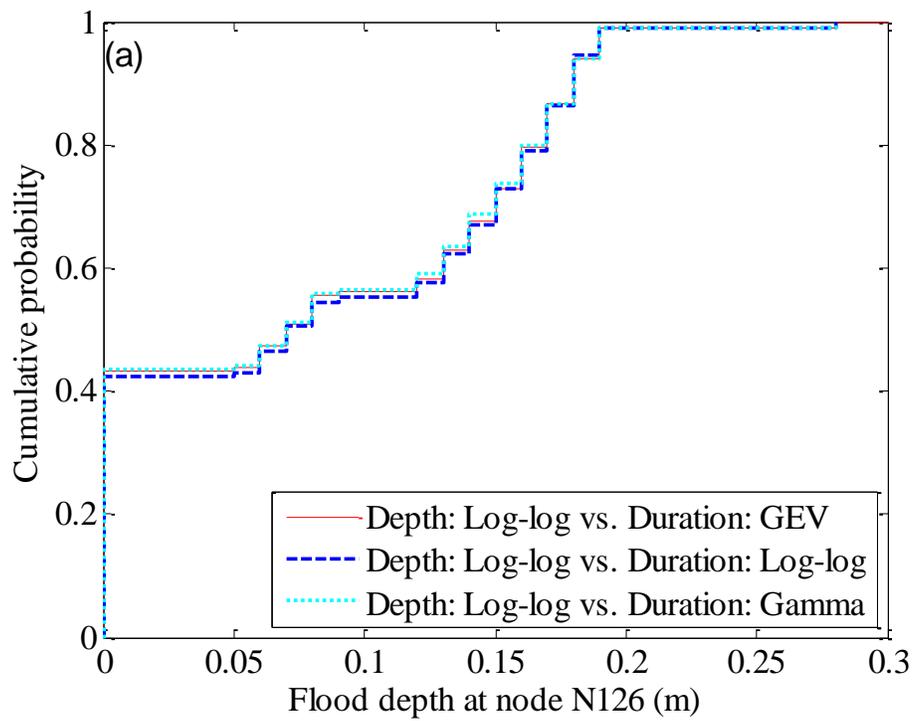
598



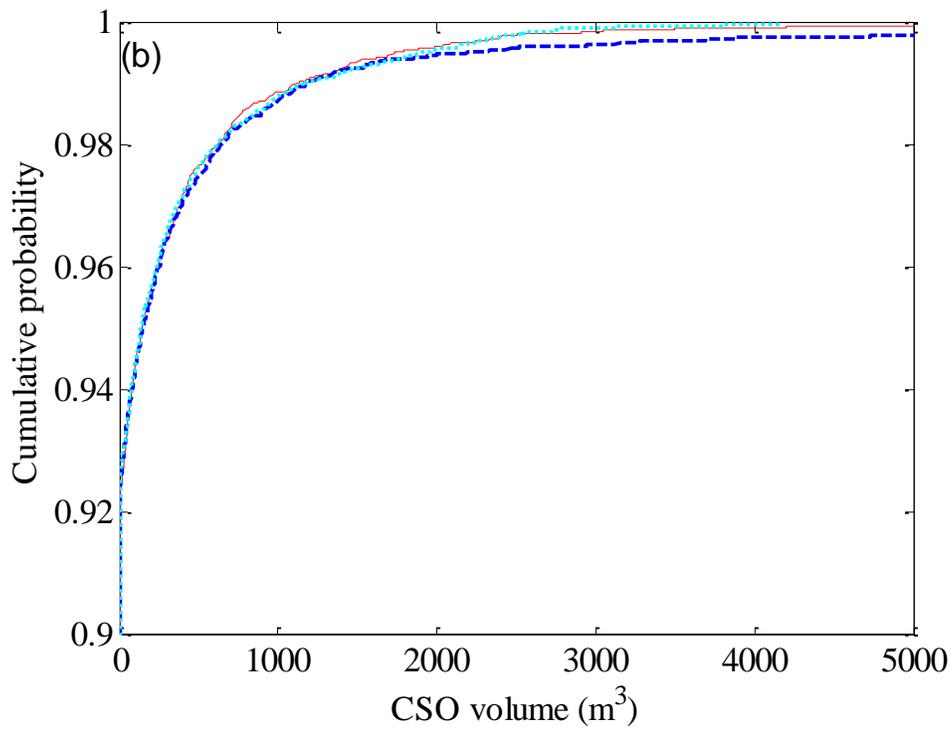
599

600

Fig. 5 - Cumulative probabilities of flood depth and CSO volume from Gumbel copula.



601

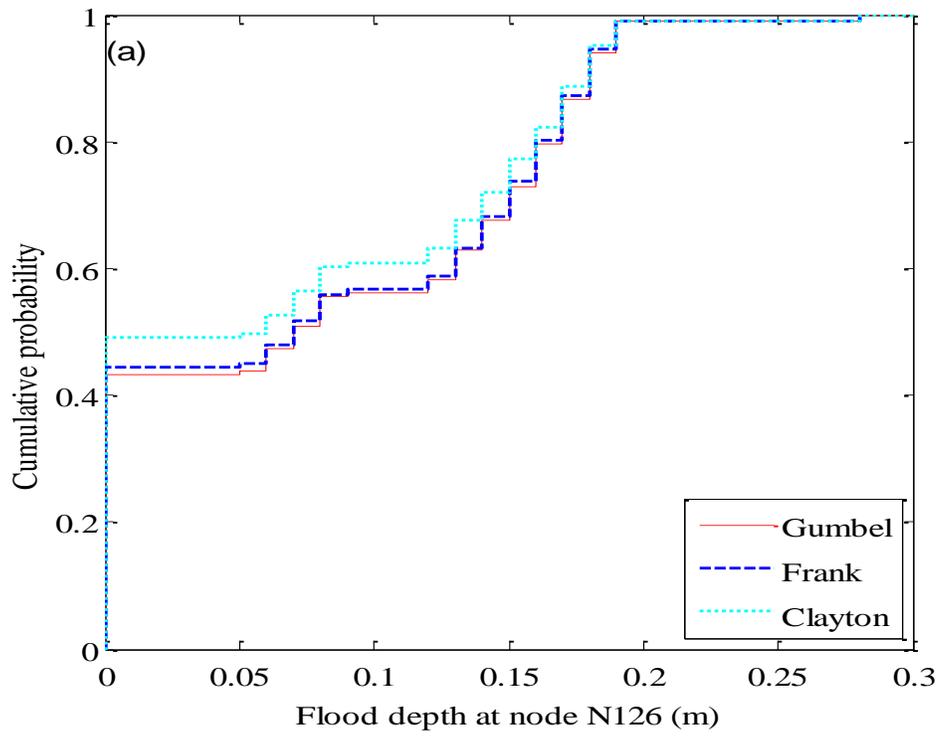


602

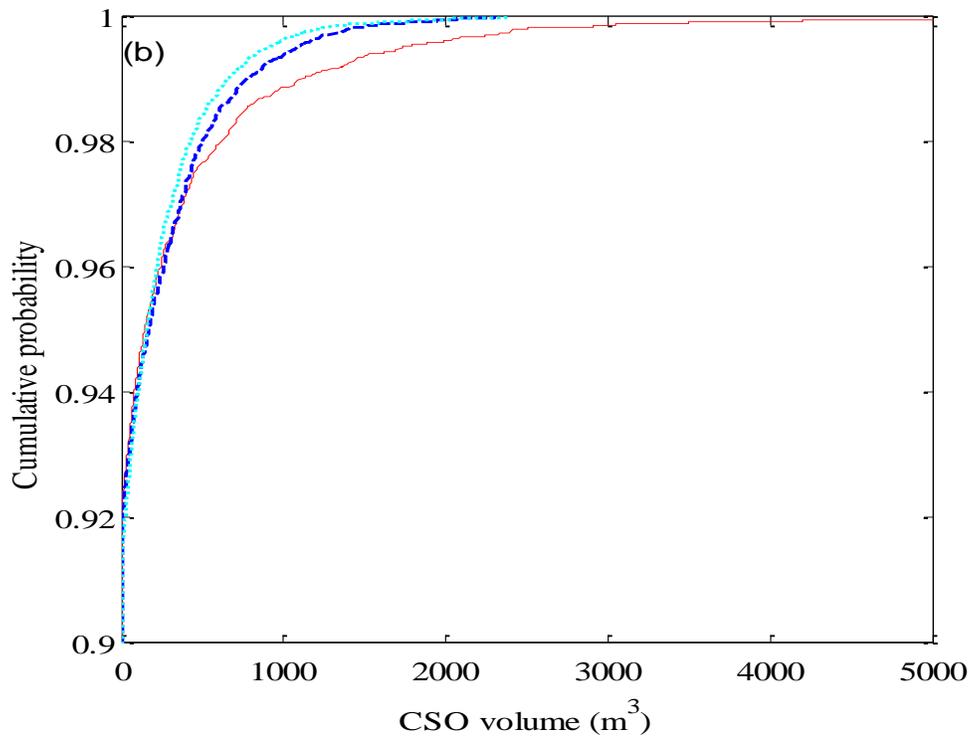
603 **Fig. 6 – Impacts of different marginal distributions on the cumulative probabilities of flood**

604

depth and CSO volume.



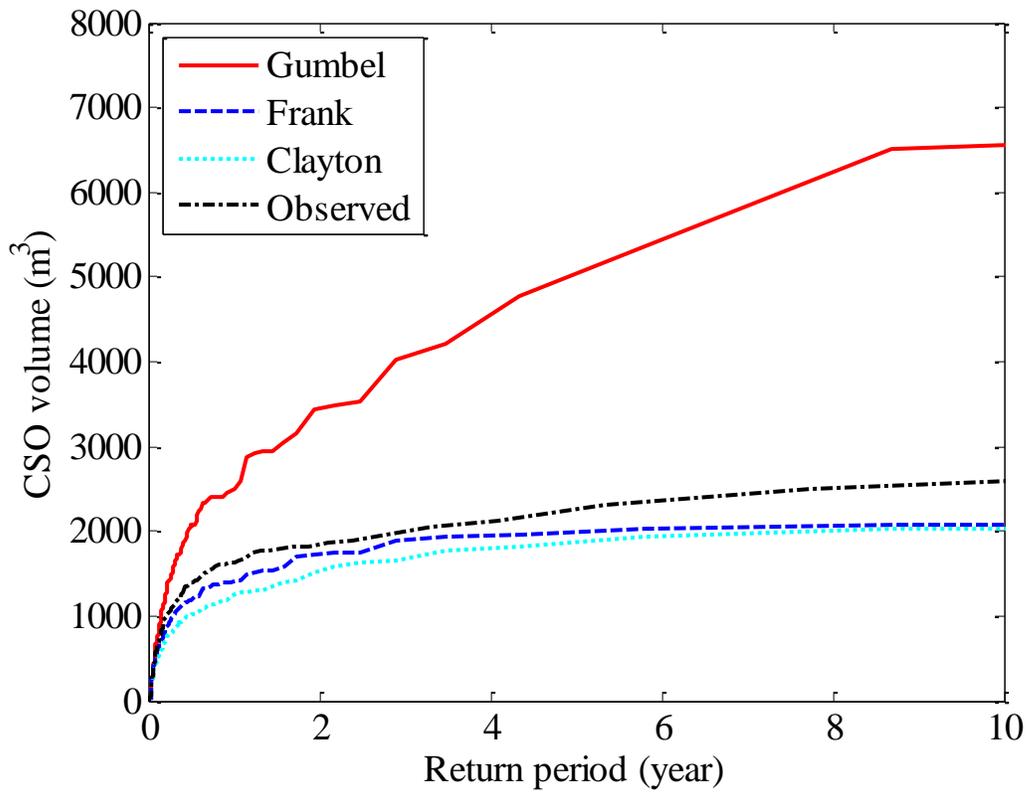
605



606
607

608

Fig. 7 – Comparison of Gumbel and Frank copulas.



609
610

Fig. 8 – Return periods of CSO discharge volumes.