Menemui Matematik
(Discovering Mathematics)

journal homepage: https://myjms.mohe.gov.my/index.php/dismath/

Modelling of Extreme Streamflow using Copula

Nur Amirah Buliah[1*] and Wendy Ling Shin Yie[2]
[1,2]*Department of Mathematics and Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor*

[1] amira.buliah@gmail.com, [2] sy_ling@upm.edu.my
*Corresponding author

Received: 28 August 2024
Accepted: 1 November 2024

ABSTRACT

This study explores applying various copula models to estimate the dependencies between streamflow and stage data for the Kahang River in Kluang, Johor. Using daily streamflow and stage data, we compared the performance of several copula parameter estimation methods: Maximum Pseudo-Likelihood Estimator (MPLE), Inference Functions for Margins Estimator (IFME), Method-of-Moments Estimator (MoM), Empirical Copula Estimation, and Robust Estimation by Maximum Mean Discrepancy Minimization (MMD). Our findings indicate that different copula performed best for different estimation methods. Specifically, the Student t-copula best fits IFME, the Frank copula for Kendall's tau, Spearman's rho, and the most recent method, MMD. Also, the Joe copula is best for the MPLE and the empirical copula estimation method. The Jackknife interval method produced narrower and more precise confidence intervals across multiple methods, making it the best interval estimator. This comprehensive analysis improves hydrological modelling, facilitating effective water resource management and flood risk assessment.

Keywords: Bivariate Copula, Maximum Pseudo-Likelihood, Inference Functions for Margins, Method-of-Moments, Empirical Copula, Robust Estimation by Maximum Mean Discrepancy Minimization, Streamflow

INTRODUCTION

Hydrology is critical in managing water resources, forecasting floods, and protecting the environment. Accurate streamflow modelling is essential for effective water management and risk assessment in hydrological studies. Traditional statistical methods often fail to capture the complexities and dependencies between hydrological variables, which are crucial for predicting extreme events. These limitations necessitate adopting advanced statistical models that can better describe and predict the behaviour of streamflow data. Shaw and Chithra (2023) implied that multivariate analyses of drought characteristics are limited and that this analysis can provide important information for water management and drought mitigation.

In hydrological modelling, capturing the dependencies between variables, such as streamflow and stage, is crucial for predicting and managing water-related events. Traditional statistical methods often fall short when modelling these dependencies due to their reliance on assumptions of normality and linearity, which are not always valid for hydrological data. These data often exhibit complex, non-linear, and asymmetric relationships, especially during extreme floods.

Copulas provide a robust framework for modelling such dependencies by separating the marginal distributions of each variable from their joint dependency structure. This flexibility allows for a more accurate representation of the intricate interactions between streamflow and stage, which is essential for predicting extreme hydrological events. Copula enables researchers to capture the nuances of tail dependencies, where extreme values in one variable, such as streamflow, may significantly influence the stage level of a river. This capability is particularly important for assessing flood risks and developing effective flood management strategies.

Streamflow and stage data are fundamental to understanding the dynamics of water movement in rivers and streams. These data are essential for flood forecasting, water resource management, and environmental protection applications. However, streamflow and stage data are inherently complex, characterised by non-linear relationships and variability influenced by precipitation, land use, and climate change factors. Modelling extreme streamflow and stage events, such as floods, presents significant challenges due to their impact on infrastructure, ecosystems, and human communities. Accurate modelling of these events is vital for designing effective flood defences and optimising water resource management. Traditional statistical models often struggle to capture the variability and extremes observed in streamflow and stage data, potentially leading to underestimating extreme event probabilities. By focusing on streamflow and stage data, this research aims to develop models that can accurately capture the complexities and risks associated with these hydrological phenomena.

The primary objective of this research is to enhance streamflow and stage data modelling using various copula models and parameter estimation methods. The study aims to identify the most suitable copula models by evaluating the performance of different copula families, including Student t, Frank, Gumbel, Clayton, and Joe copula, in capturing the dependencies between streamflow and stage data. It also seeks to assess the effectiveness of various copula parameter estimation methods, such as Maximum Pseudo-Likelihood Estimator (MPLE), Inference Functions for Margins Estimator (IFME), Method-of-Moments Estimator (MoM), Empirical Copula Estimation, and Robust Estimation by Maximum Mean Discrepancy Minimization (MMD). Additionally, the research aims to compare goodness-of-fit tests using metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) to evaluate and select the best-fitting models. Finally, the study will determine the best confidence interval method by analysing the accuracy and precision of different confidence intervals construction methods, such as Wald, Bootstrap, and Jackknife, to ensure the reliability of the copula parameter estimates. By achieving these objectives, the research seeks to provide a comprehensive framework for accurately modelling the complex dependencies in streamflow data, particularly for extreme events, ultimately contributing to more effective water resource management, improved flood forecasting, and better-informed environmental protection strategies.

The remainder of the article is organised as follows. In the next section, we briefly review the relative literature. The data and methods are introduced in the third section, followed by the research findings. Finally, we will present the conclusion of this research.

LITERATURE REVIEW

Copula, introduced by Sklar in 1959, has become a cornerstone in multivariate statistical analysis, particularly useful for modelling the dependence structure between random variables. Copulas allow the construction of multivariate distribution functions based on one-dimensional margins,

providing a flexible approach to capturing complex dependencies in data (Nelsen, 2006). This flexibility is especially advantageous in fields where relationships between variables are non-linear and intricate, such as hydrology.

However, applying copula in practice comes with significant challenges despite their potential. Tootoonchi et al. (2022) highlighted the lack of practical guidance on various critical aspects, including exploring dependence at different scales, pre-treating data, selecting appropriate copula models, and validating the model fit. These challenges underline the need for a more structured approach in applying copula, particularly in hydrology, where the complexity of data requires careful consideration of model selection and evaluation.

Copula in hydrology have gained traction due to their ability to model dependencies independently of marginal distributions, allowing for more accurate joint distribution constructions. For example, the Archimedean copulas have been effectively used to model the dependence structure between flood characteristics like peak flow, volume, and duration, as demonstrated by Karmakar and Simonovic (2009). Archimedean copula, characterised by their simplicity and analytical tractability, have also been extensively studied. The Clayton, Gumbel, and Frank copula are notable members of this family, each offering different dependency structures that can be tailored to specific data characteristics (Joe, 1997).

Several methods have been explored in parameter estimation, each with advantages and limitations. The Maximum Pseudo-Likelihood Estimator (MPLE) is favoured for its efficiency and simplicity, operating by maximising a pseudo-likelihood function as an approximation of the true likelihood function (Genest et al., 1995). The Inference Functions for Margins Estimator (IFME), a two-step procedure, estimates marginal parameters before copula parameters, simplifying the estimation process for large datasets (Joe & Xu, 1996). Other methods include the Method-of-Moments (MoM), which matches sample moments with theoretical moments. The MOM estimator is consistent and asymptotically normal (Oh and Patton, 2013). Empirical copula estimation is a non-parametric approach useful when the underlying copula form is unknown (Deheuvels, 1979). The Robust Estimation by Maximum Mean Discrepancy Minimization (MMD) stands out for its robustness against outliers, minimising the difference between empirical and theoretical copula distributions in a reproducing kernel Hilbert space (Gretton et al., 2005).

Recent research has delved into the effectiveness of different estimation methods under various conditions. Lokoman and Yusof (2019) found that the IFME method excels for small sample sizes and lower correlation levels, as demonstrated in a rainfall study from Kuala Krai and Ulu Sekor station in Malaysia. For datasets with very strong dependence, the original MPLE is preferred, particularly for larger samples (Joo et al., 2020). Buliah and Yie (2020) applied Archimedean Copula with Maximum Likelihood Estimation (MLE) to model extreme rainfall in Malaysian hydrological stations. Ko and Hjort (2019) introduced a model-robust inference framework for the IFME of copula parameters, while Idiou and Benatia (2021) presented MLE, IFME, and MoM for estimating the Archimedean class. Alquier et al. (2023) proposed an MMD-based method for copula models, highlighting its robustness and consistency even in the presence of outliers or model misspecification.

In terms of constructing confidence intervals, methods such as Wald, Bootstrap, and Jackknife are explored for their comparative performance with hydrological data. The Wald interval is based on the asymptotic normality of parameter estimates, requiring large sample sizes for accuracy (Agresti, 2018). Kummaraka and Srisuradetchai (2023) provided an explicit formula for constructing Wald Confidence intervals for the dependence parameter in a bivariate Clayton

copula. The Percentile Bootstrap, a non-parametric alternative, resamples data to recalibrate estimates, making it particularly useful for smaller samples (Efron & Tibshirani, 1993). The Jackknife method, which iteratively leaves out one observation, offers robustness in the presence of outliers (Tukey, 1958).

The growing use of copula in hydrology, especially for modelling streamflow data, demonstrates their potential to provide a comprehensive understanding of hydrological phenomena. Worland et al. (2019) highlighted the superior performance of copula methods in streamflow estimation, while Sahoo et al. (2020) identified the Clayton and Frank Copula as optimal for the Mahanadi River basin in India. In Malaysia, Latif and Mustafa (2021) utilised a semiparametric copula-based approach to model flood characteristics, and Shiau and Lien (2021) showed that copula-based methods effectively infill missing data in streamflow datasets from eastern Taiwan.

Through advanced parameter estimation methods and robust goodness-of-fit evaluations, copula-based models can significantly enhance the reliability and accuracy of hydrological modelling, ultimately contributing to better water resource management and environmental protection.

METHODOLOGY

Data Used

The data used in this study is the real data of the daily flow and stage of Kahang River in Kluang, Johor, from the Department of Irrigation and Drainage Malaysia. The daily data used are in the form of cubic meter per second (m^3/s) for flow and meter (m) for stage from April 1978 to July 2009.

Copula

A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform on the interval $[0, 1]$. Sklar's Theorem is the cornerstone of copula theory. It states that for any multivariate distribution function F with marginals $F_1, F_2, \dots, F_d$, there exists a copula C such that for all $x \in \mathbb{R}^d$:

$$H(x_1, x_2, \dots, x_d) = C\big(F_1(x_1), F_2(x_2), \dots, F_d(x_d)\big). \quad (1)$$

If the marginals F_i are continuous, then the copula C is unique. Conversely, given a copula C and univariate marginal distribution functions $F_1, F_2, \dots, F_d$, the function H defined above is a valid joint distribution function with the specified marginals.

Parametric Copula

Student-t Copula

The Student t-copula is derived from the multivariate Student t-distribution. It is particularly useful for modelling dependencies with heavy tails, where extreme events in one variable are likely to be associated with extreme events in another. The t-copula with ν degrees of freedom and correlation matrix R is defined by:

$$C(u_1, \dots, u_d; \nu, R) = t_{R,\nu}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)) \quad (2)$$

where $t_{R,\nu}$ is the joint CDF of the multivariate Student t-distribution with ν degrees of freedom and correlation matrix R, and t_ν^{-1} is the inverse CDF of the univariate Student t-distribution.

Clayton Copula

The Clayton copula is an Archimedean copula known for capturing lower tail dependence, meaning it models scenarios where extremely low values in one variable are associated with extremely low values in another. The Clayton copula with parameter $\theta > 0$ is defined by:

$$C(u_1, u_2; \theta) = \left(\max\left[\left(u_1^{-\theta} + u_2^{-\theta} - 1\right), 0\right]\right)^{-\frac{1}{\theta}} \quad (3)$$

Gumbel Copula

The Gumbel copula is another Archimedean copula that captures upper tail dependence, which is useful for modelling situations where extremely high values in one variable are likely to be associated with extremely high values in another. The Gumbel copula with parameter $\theta \geq 1$ is defined by:

$$C(u_1, u_2; \theta) = \exp\left(-\left[(-\log u_1)^\theta + (-\log u_2)^\theta\right]^{\frac{1}{\theta}}\right) \quad (4)$$

Frank Copula

The Frank copula is also an Archimedean copula but is unique in that it can model both positive and negative dependencies. It does not exhibit tail dependence. The Frank copula with parameter $\theta \neq 0$ is defined by:

$$C(u_1, u_2; \theta) = -\frac{1}{\theta} \log\left(\frac{\left(1 + e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}\right) \quad (5)$$

Joe Copula

The Joe copula, another Archimedean copula, captures upper tail dependence like the Gumbel copula but has different dependency structures. The Joe copula with parameter $\theta > 1$ is defined by:

$$C(u_1, u_2; \theta) = 1 - \left[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\right]^{\frac{1}{\theta}} \quad (6)$$

Semi-Parametric Approach

Maximum Pseudo-Likelihood Estimator (MPLE)

The sample of pseudo-observations is defined via

$$U_{i,n} = \left(F_{n,1}(X_{i1}), \dots, F_{n,d}(X_{id})\right), \quad i \in \{1, \dots, n\}.$$

For any $j \in \{1, \dots d\}$, let R_{ij} denotes the rank of X_{ij} among $X_{1j}, \dots, X_{nj}$,

$$F_{n,j}(X_{ij}) = R_{ij}/(n+1), \quad \text{for } i \in \{1, \dots, n\}.$$

Leading to the MPLE

$$\theta_n = \underset{\theta \in \Theta}{\operatorname{argsup}} \sum_{i=1}^{n} \log C_\theta(U_{i,n}). \tag{7}$$

Parametric Approach

The Inference Functions for Margins Estimator (IFME)

IFME is a two-stage estimation procedure designed to reduce the computational burden of MLE in copula models. Estimate the unknown marginal parameter vectors $\gamma_{0,1}, \dots, \gamma_{0,d}$ for each margin *j* by maximising the log-likelihood function:

$$\gamma_{n,j} = \underset{\gamma_j \in \Gamma_j}{\arg\sup} \sum_{i=1}^{n} \log f_{j,\gamma_j}(X_{ij})$$

Here, f_{j,γ_j} is the density function of the *j*-th margin with parameters γ_j. With the estimated marginal parameters $\gamma_{n,1}, \dots, \gamma_{n,d}$, transform the original data X_{ij} to pseudo-observations:

$$U_{i,\gamma_n} = (F_{1,\gamma_{n,1}}(X_{i1}), \dots, F_{d,\gamma_{n,d}}(X_{id})), \quad i \in \{1, \dots, n\}.$$

Estimate the copula parameter vector θ_0 by maximising the log-likelihood of the copula density function:

$$\theta_n = \underset{\theta \in \Theta}{\arg\sup} \sum_{i=1}^{n} \log c_\theta(U_{i,\gamma_n}) \tag{8}$$

Non-Parametric Approach

Method-of-Moments Estimator (MoM)

Method-of-moments estimators in the context of copula are extensions of traditional method-of-moments estimators used in various statistical areas. In the copula setting, moments of random variables are replaced by moments of the copula, such as Kendall's tau or Spearman's rho.
Given a copula family $C = \{C_\theta : \theta \in \Theta\}$, let g_τ and g_{ρ_s} be functions defined by:

$$g_\tau(\theta) = \tau(C_\theta) \quad \text{and} \quad g_{\rho_s}(\theta) = \rho_s(C_\theta), \quad \theta \in \Theta \subseteq \mathbb{R},$$

where $\tau(C_\theta)$ and $\rho_s(C_\theta)$ are the Kendall's tau and Spearman's rho of C_θ, respectively. Method-of-moments estimators based on Kendall's tau (respectively, Spearman's rho) can be used for the family *C* if the function g_τ (respectively, g_{ρ_s}) is one-to-one. In that case, the estimator $\hat{\theta}$ of θ_0 is given by:

$$\hat{\theta}_n = g_\tau^{-1}(\tau_n) \quad \text{(respectively, } \hat{\theta}_n = g_{\rho_s}^{-1}(\rho_{s,n})), \tag{9}$$

where τ_n (respectively, $\rho_{s,n}$) is the sample version of Kendall's tau (respectively, Spearman's rho).

Empirical Copula Estimation

The empirical copula method provides a non-parametric estimate of the copula function. This method is particularly useful when estimating the copula without making strong parametric assumptions about the underlying joint distribution. The empirical copula C_n is defined as follows:

$$C_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1\big(U_{i,n} \le u\big) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} 1(U_{ij,n} \le u_j), \quad u \in [0,1]^d, \tag{10}$$

where $U_{i,n} = (U_{i1,n}, \ldots, U_{id,n})$ are the pseudo-observations and the inequalities $U_{i,n} \le u$ are to be understood component-wise.

The empirical copula is essentially the empirical distribution function of pseudo-observations. It is a consistent estimator of the true copula *C*, and its asymptotic follows from those of the so-called empirical copula process.

Robust Estimation by Maximum Mean Discrepancy Minimization (MMD)

The MMD is a method proposed to address robustness issues in the presence of outliers and model misspecifications. This method utilises the concept of MMD, which measures the distance between two probability distributions using their embeddings in a reproducing kernel Hilbert space (RKHS). The MMD between two probability distributions, *P* and *Q,* is defined as:

$$D(P,Q) = \sup_{f\in\mathcal{F}} \left| \int f dP - \int f dQ \right|,$$

where $\mathcal{F}$ is the unit ball in an RKHS with a kernel *K*. *K* is the popular Gaussian kernel $K_G(u,v) = \exp(\|u - v\|^2/\gamma^2)$. This can be rewritten using the kernel mean embeddings μ_P and μ_Q:

$$D(P,Q) = \| \mu_P - \mu_Q \|_{\mathcal{H}},$$

where $\mu_P = E_P[K(\cdot, X)]$ and $\mu_Q = E_Q[K(\cdot, Y)]$ for $X \sim P$ and $Y \sim Q$. When estimating copula parameters using MMD, find the parameter θ that minimises the distance between the empirical mean embedding and the model mean embedding:

$$\hat{\theta} = \arg\min_\theta \| \mu_{C_n} - \mu_{C_\theta} \|_{\mathcal{H}}, \tag{11}$$

where C_n is the empirical copula, and C_θ is the parametric copula with parameter θ.

Goodness-of-Fit Tests

Root Mean Squared Error (RMSE)

RMSE is the square root of the MSE. It provides a measure of the average magnitude of the error in the same units as the observed copula values, *C*. RMSE is particularly useful for understanding the scale of the errors.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(C(u_i,v_i) - C_\theta(u_i,v_i)\big)^2}$$

Mean Absolute Error (MAE)

MAE measures the average absolute difference between the observed copula values, *C,* and the values predicted by the copula model, C_θ. MAE is a robust metric that is less sensitive to outliers than MSE and RMSE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |C(u_i, v_i) - C_\theta(u_i, v_i)|$$

Mean Relative Error (MRE)

MRE measures the average relative difference between the observed copula values, *C,* and the values predicted by the copula model, C_θ. MRE compares errors across scales and explains the model's relative performance.

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n} \frac{C(u_i, v_i) - C_\theta(u_i, v_i)}{C(u_i, v_i)}$$

Confidence Intervals

Wald Interval

The Wald interval is a common method for constructing confidence intervals based on the asymptotic normality of estimators. Assuming $\hat{\theta}$ is approximately normally distributed, the $(1-\alpha)\%$ confidence interval for θ is given by:

$$\hat{\theta} \pm z_{\alpha/2} \cdot SE(\hat{\theta})$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

Percentile Bootstrap Interval

The percentile bootstrap method involves resampling the data to estimate the distribution of the estimator directly. Determine the $\alpha/2$ and $1-\alpha/2$ percentiles of the empirical distribution. The $(1-\alpha)\%$ confidence interval is given by:

$$(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2})$$

where $\hat{\theta}^*_{\alpha/2}$ and $\hat{\theta}^*_{1-\alpha/2}$ are the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ percentiles of the bootstrap distribution, respectively.

Jackknife Interval

The jackknife method is a resampling technique that systematically leaves out one observation at a time from the sample set to estimate the bias and variance of an estimator. Assuming the jackknife estimator follows a normal distribution, the $(1-\alpha)\%$ confidence interval is:

$$\hat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})_{jack}}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

Simulation Studies

Simulation studies are conducted in this section to investigate the performance of estimation of copula parameter ($\hat{\theta}$) using MPLE. The simulations are conducted as follows:

1. Generate random samples for each combination of three different sample sizes (n = 100, 500, 1000) and levels of dependence parameter (τ = 0.2, 0.5, 0.8).
2. Estimate the Clayton copula parameter using MPLE, IFME, Kendall's Tau, Spearman's Rho, and MMD parameter estimation methods.
3. Repeat the steps above 100 times.
4. Calculate RMSE, MAE, and MRE to evaluate and compare the performance of the methods.

RESULTS AND DISCUSSION

The study compared and evaluated the estimation performance of semi-parametric, parametric, and non-parametric approaches based on the goodness-of-fit statistics. A simulation study using MPLE was conducted to compare the parametric copulas' performance before the real streamflow data was fitted.

Data Simulation

Data simulation is a critical process in evaluating and understanding the behaviour of statistical models under controlled conditions. This study simulated data using the Clayton copula function, which models the dependency structure between two random variables. This approach is particularly relevant in hydrological studies where understanding the joint behaviour of related variables is essential. The simulation involved various methods of copula parameter estimation, including MPLE, IFME, Kendall's tau, Spearman's Rho, and MMD. The simulation was conducted across different sample sizes (n = 100, 500, 1000) and correlation values (τ = 0.2, 0.5, 0.8), comprehensively evaluating each estimation method's performance.

The performance of each estimation method was evaluated using three key metrics: RMSE, MAE, and MRE, as shown in Table 1. These metrics assess the accuracy and reliability of the copula parameter estimates under varying conditions. For small sample sizes (n = 100), the RMSE and MAE values were relatively close across all correlation levels (τ = 0.2, 0.5, 0.8), indicating that the methods performed similarly regarding absolute error. However, the MMD method consistently exhibited slightly lower RMSE and MAE values, particularly in scenarios with low and high correlations, suggesting better performance in small samples. The IFME method demonstrated better performance with medium correlations. The MRE values revealed greater variability among the methods, with Kendall's tau showing notably higher MRE, particularly at τ = 0.2 and τ = 0.8, making it less reliable for parameter estimation under these conditions. In contrast, MMD and IFME maintained relatively stable and lower MRE values, indicating better robustness for small sample sizes.

As the sample size increased to n = 500, the MMD method showed competitive performance, with the lowest RMSE and slightly lower MAE at τ = 0.2, while MPLE exhibited better performance with low to medium correlations. Kendall's tau showed better performance with medium to high correlations, while IFME demonstrated superior performance with high correlations. However, Kendall's tau and Spearman's Rho exhibited increased MRE at τ = 0.5,

suggesting that these rank-based methods may introduce bias or variability in moderate sample sizes.

For large sample sizes (n = 1000), Kendall's tau exhibited the lowest RMSE and MAE values, suggesting better performance with low correlations. Spearman's Rho and MMD demonstrated better performance with medium correlations, while IFME performed well in large samples with high correlations. The MRE values stabilized across all methods, although IFME showed some variability at lower correlation levels (τ = 0.2). The consistent performance of all methods across all sample sizes suggests that they offer a balanced trade-off between bias and variance, making them reliable choices for large datasets.

Table 1: RMSE, MAE, and MRE of Clayton Copula Parameter Estimation

Sample size	Method	$\tau = 0.2$			$\tau = 0.5$			$\tau = 0.8$		
n		RMSE	MAE	MRE	RMSE	MAE	MRE	RMSE	MAE	MRE
100	MPLE	0.4059	0.3321	6.3066	0.4084	0.3342	4.2488	0.4053	0.3320	4.7344
	IFME	0.4086	0.3336	4.0332	0.4040	0.3299	4.3302	0.4077	0.3344	4.4162
	Kendall's Tau	0.4073	0.3327	16.196	0.4108	0.3357	4.9864	0.4091	0.3348	96.518
	Spearman's Rho	0.4082	0.3342	6.1060	0.4088	0.3348	4.5137	0.4084	0.3347	5.6167
	MMD	0.4038	0.3305	4.6484	0.4091	0.3350	5.3321	0.4033	0.3296	6.0673
500	MPLE	0.4073	0.3323	5.6694	0.4084	0.3338	5.3980	0.4102	0.3355	6.6567
	IFME	0.4079	0.3329	4.9585	0.4085	0.3339	6.5974	0.4066	0.3323	4.9293
	Kendall's Tau	0.4082	0.3333	5.1594	0.4084	0.3333	7.5461	0.4071	0.3318	4.9466
	Spearman's Rho	0.4082	0.3333	5.4183	0.4094	0.3345	7.9069	0.4085	0.3339	5.2393
	MMD	0.4072	0.3324	5.9619	0.4091	0.3345	5.2426	0.4072	0.3326	5.3119
1000	MPLE	0.4088	0.3337	4.5178	0.4083	0.3336	4.5274	0.4083	0.3334	5.5505
	IFME	0.4080	0.3334	43.466	0.4088	0.3336	6.3033	0.4069	0.3323	6.3528
	Kendall's Tau	0.4079	0.3330	7.1222	0.4085	0.3337	5.7323	0.4074	0.3326	6.0825
	Spearman's Rho	0.4090	0.3341	6.9499	0.4080	0.3331	5.2929	0.4077	0.3325	7.6419
	MMD	0.4094	0.3342	5.7488	0.4080	0.3331	5.1138	0.4085	0.3337	4.8699

The data simulation analysis shows that the performance of copula parameter estimation methods varies depending on sample size and correlation levels. The MMD method consistently provides the most accurate estimates for small sample sizes, particularly in low and high-correlation scenarios, while IFME performs well with medium-correlation data. As sample sizes increase to a medium range, MPLE is the most reliable method for low to medium correlations. Kendall's tau performs well for medium to high correlations, while MMD is also effective for low correlations. Additionally, IFME shows strong performance in high-correlation data within medium-sized samples. For large sample sizes, the accuracy of all methods becomes more comparable, with Kendall's tau slightly better for low correlations and Spearman's Rho and MMD preferred for medium correlations. IFME remains a robust choice for high-correlation data in large samples. Overall, IFME and MMD are recommended for small datasets, MPLE, IFME, Kendall's tau, and MMD for medium datasets, and IFME, Kendall's tau, Spearman's Rho or MMD for large datasets, depending on the correlation level.

Modelling of Streamflow using Copula

The application of copula modelling on the streamflow data of Kahang River (flow and stage) involves different estimation methods: semi-parametric, parametric, and non-parametric. This section evaluates the performance of these methods using various statistical metrics and confidence intervals.

Data and Descriptive Analysis

This research uses the daily flow and stage of the Kahang River. The data is sourced from the Department of Irrigation and Drainage Malaysia. The 95^{th} percentile method is used to extract heavy or extreme streamflow, which results in 571 observations. Descriptive statistics provide the measures and the summaries of streamflow data; hence, they play an important role in describing the fundamental features of the data used in research. The descriptive statistics of the daily flow and stage of Kahang River are shown in Table 2.

Table 2: Descriptive Statistics of Daily Flow and Stage for Kahang River

Streamflow Variable	Mean	Median	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum
Flow ($\boldsymbol{m^3/s}$)	193.699	145.690	138.422	3.171	15.362	97.860	1,152.430
Stage ($\boldsymbol{m}$)	6.408	6.320	1.153	0.409	2.982	3.700	10.610

From Table 2, the minimum daily flow and stage during heavy streamflow recorded for Kahang River is 97.860 m^3/s and 3.700 m respectively. The maximum daily flow and stage recorded is 1,152.430 m^3/s and 10.610 m respectively. This flow level is quite high, typically seen in large rivers or during extreme weather events like heavy rainfall. The flow data is averaged at 193.699 m^3/s with a standard deviation of 138.422 m^3/s. Meanwhile, the stage data is averaged at 6.408 m with a standard deviation of 1.153 m.

The Dependence Level Between the Streamflow Data

The dependence between the streamflow data was measured first using Kendall's tau and Spearman's rho method. The dependence of flow and stage is shown in Table 3.

Table 3: The Dependence Level between Flow and Stage

	Correlation	p-value
Kendall's tau	0.4624	0.0000
Spearman's rho	0.6129	0.0000

From Table 3, the flow is positively associated with the stage. The dependence level of the streamflow data using Kendall's tau and Spearman's rho is 0.4624 and 0.6129, respectively, which is moderate in the association. The p-values for both methods are 0.0000 at the significance level of $\alpha = 0.05$. Since the p-value is less than 0.05, the correlation for the streamflow data is significant.

Semi-Parametric Approach

The semi-parametric approach presents and discusses the results of applying the MPLE method to estimate the copula parameters.

Maximum Pseudo-Likelihood Estimator (MPLE)

MPLE method was employed to estimate the parameters of various copula families $(\hat{\theta})$. These estimates, their corresponding standard errors (SE), and GOF statistics are presented in Table 4.

The performance of these copulas was further evaluated using error metrics such as RMSE, MAE, and MRE for the copula, as shown in Table 4. The Clayton copula exhibited the lowest RMSE at 0.4033, suggesting it provides the best fit by minimising the overall prediction error. Regarding MAE and MRE, the Joe copula outperformed the others with a value of 0.3288 and -

2.0411, indicating that it offers the most accurate predictions on average and effectively minimises relative errors, although it may not be the best performer in RMSE.

Table 4: Copula Parameter Estimation using MPLE and GOF statistics

Copula		$\hat{\theta}$	SE	MLE	RMSE	MAE	MRE
Student t	ρ	0.6390	0.0230	146.8000	0.4081	0.3345	-2.1805
	df	18.1070	-				
Clayton		1.7210	0.0910	98.4800	0.4033	0.3293	-2.3128
Gumbel		1.6660	0.0710	135.1000	0.4147	0.3384	-2.4140
Frank		4.4460	0.4260	129.3000	0.4142	0.3396	-2.6825
Joe		1.8920	0.1010	114.1000	0.4052	0.3288	-2.0411

Table 5: MPLE Interval

Copula		Wald	Bootstrap	Jackknife
Student t	ρ	(0.5939, 0.6841)	(0.5964, 0.6839)	(0.6220, 0.6560)
	df	-	(12.7903, 23.3023)	(15.5677, 20.6464)
Clayton		(1.5415, 1.8996)	(1.4818, 2.0441)	(1.7087, 1.7333)
Gumbel		(1.5278, 1.8046)	(1.5615, 1.8041)	(1.6628, 1.6692)
Frank		(3.6114, 5.2798)	(3.8761, 5.1787)	(4.4280, 4.4640)
Joe		(1.6929, 2.0904)	(1.6897, 2.0850)	(1.8876, 1.8964)

Confidence intervals for the copula parameter estimates were calculated using Wald, Bootstrap, and Jackknife methods, as presented in Table 5. These intervals are consistent across the different methods, reinforcing the reliability of the estimates. Notably, the Student t-copula showed wider intervals for the degrees of freedom, indicating greater variability and uncertainty in this parameter. The Clayton copula's intervals were narrow and consistent, emphasizing the stability of its parameter estimates. The Gumbel, Frank, and Joe copula also demonstrated consistent intervals, supporting the robustness of their estimates.

In conclusion, the Joe copula is the best-fitting model under the MPLE method because of its lowest MAE and MRE values. While the Clayton copula has the lowest RMSE, the Joe copula offers the most balanced and accurate fit across the key error metrics. Therefore, the Joe copula is recommended as the optimal model for the dataset when using the MPLE method.

Parametric Approach

The parametric approach presents and discusses the results of applying the IFME method to estimate the copula parameters.

The Inference Functions for Margins Estimator (IFME)

Fitting Data to Marginal Distributions

The parametric approach in the analysis utilizes the IFME method to estimate the copula parameters, requiring the determination of the marginal distributions first. The goodness-of-fit (GOF) statistics for different distributions—Lognormal, Gamma, and Weibull—fitted to the flow and stage data of the Kahang River are presented in Table 6.

Table 6: Fitting Flow and Stage to Marginal Distributions

	Flow			Stage		
Distribution	χ^2	AD	R^2	χ^2	AD	R^2

Lognormal	0.0000	21.2966	0.8663	0.0408	1.1952	0.9926
Gamma	0.0000	33.2456	0.8383	0.0149	1.2489	0.9909
Weibull	0.0000	43.3446	0.6790	0.0000	5.0552	0.9391

Based on the analysis of the GOF statistics, the Lognormal distribution is identified as the best-fitting model for both the flow and stage data of the Kahang River. The Lognormal distribution consistently shows superior performance across the Chi-square, Anderson-Darling, and R-squared tests, particularly excelling in the AD and R^2 metrics, which are crucial for assessing the distribution's fit to the data. Therefore, the lognormal distribution is recommended as the optimal choice for modelling flow and stage variables using the IFME method.

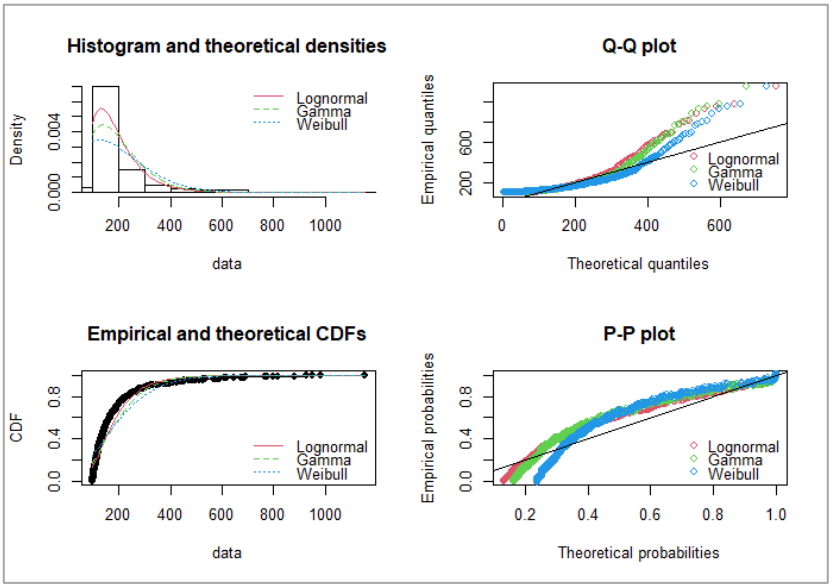

Figure 1: Comparison Between the Distributions Fitted to the Flow of the Kahang River

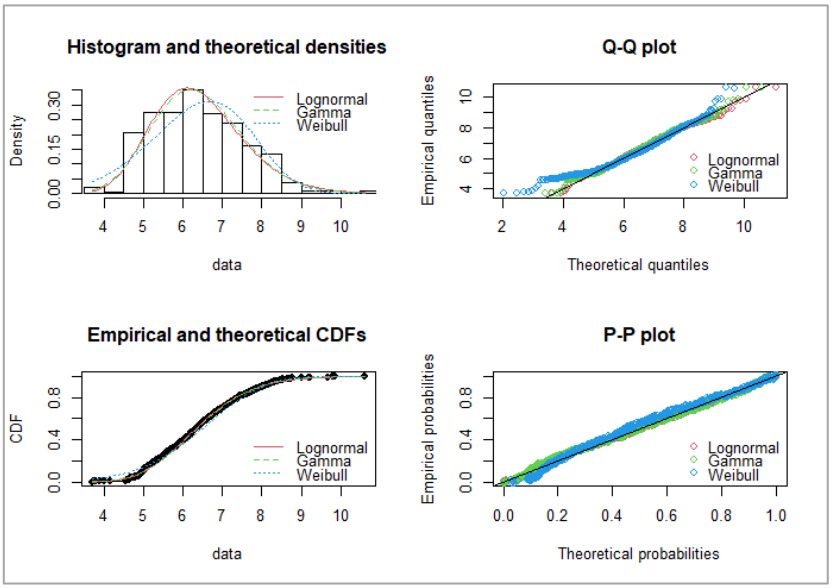

Figure 2: Comparison Between the Distributions Fitted to the Stage of the Kahang River

Figures 1 and 2 support the statistical findings, with the lognormal distribution closely matching the empirical data for both flow and stage. The Weibull distribution shows the least fit, particularly for the stage data.

Copula Parameter Estimation using IFME

The IFME method was employed to estimate the parameters of various copula families. These copula parameter estimates $(\hat{\theta})$, their corresponding standard errors (SE), and GOF statistics are presented in Table 7.

Table 7: Copula Parameter Estimation using IFME and GOF statistics

Copula		$\hat{\theta}$	SE	MLE	RMSE	MAE	MRE
Student t	ρ	0.5878	0.0260	119.8000	0.4000	0.3280	-2.5402
	df	12.5961	2.4780				
Clayton		1.7210	0.0970	75.1200	0.4076	0.3340	-2.2449
Gumbel		1.4260	0.0440	80.7600	0.4105	0.3394	-2.1041
Frank		4.3030	0.4310	121.3000	0.4120	0.3397	-2.3587
Joe		1.4620	0.0560	58.8000	0.4072	0.3367	-2.5541

Table 8: IFME Interval

Copula		Wald	Bootstrap	Jackknife
Student t	ρ	(0.5376, 0.6379)	(0.5438, 0.6466)	(0.5859, 0.5897)
	df	(7.7391, 17.4532)	(8.9444, 17.7623)	(11.4717, 13.7205)
Clayton		(1.5309, 1.9103)	(1.4396, 1.9541)	(1.7087, 1.7333)
Gumbel		(1.3384, 1.5126)	(1.3235, 1.5381)	(1.4228, 1.4292)
Frank		(3.4587, 5.1466)	(3.8500, 4.8894)	(4.2850, 4.3210)
Joe		(1.3526, 1.5724)	(1.3203, 1.6081)	(1.4576, 1.4664)

Confidence intervals for the copula parameter estimates were calculated using Wald, Bootstrap, and Jackknife methods, as presented in Table 8. These intervals are consistent across the different methods, reinforcing the reliability of the estimates.

The Student t-copula emerges as the best-fitting model under the IFME method. Its lowest RMSE value indicates that it effectively minimizes squared prediction errors, making it a reliable choice for data modelling. The Student t-copula minimises absolute errors, and the Gumbel copula in relative errors. The Student t-copula offers the most balanced and accurate fit across the key error metrics. Therefore, the Student t-copula is recommended as the optimal model for the dataset when using the IFME method, particularly when overall prediction accuracy is a priority.

Non-Parametric Approach

The non-parametric approach presents and discusses the results obtained from applying the MoM, empirical copula estimation, and MMD method to estimate the copula parameters.

Method-of-Moments Estimator (MoM)

Kendall's Tau Estimation

The copula parameter estimation using Kendall's Tau is provided in Table 9. The estimated copula parameters $(\hat{\theta})$, their standard errors (SE) and GOF statistics for different copula families are:

Table 9: Copula Parameter Estimation using Kendall's Tau

Copula		$\hat{\theta}$	SE	RMSE	MAE	MRE
Student t	ρ	0.6642	0.0250	0.4096	0.3323	-2.0538
	df	4.0000	-			
Clayton		1.7210	0.1490	0.4117	0.3367	-2.2175
Gumbel		1.8600	0.0750	0.4102	0.3357	-2.9027
Frank		5.0920	0.1440	0.3972	0.3203	-2.6896
Joe		2.5880	-	0.3971	0.3245	-2.2507

Table 10: Kendall's Tau Interval

Copula		Wald	Bootstrap	Jackknife
Student t	ρ	(0.6145, 0.7139)	(0.6177, 0.7209)	(0.6621, 0.6663)
Clayton		(1.4278, 2.0133)	(1.4717, 2.0792)	(1.7087, 1.7333)
Gumbel		(1.7139, 2.0067)	(1.7275, 2.0121)	(1.8538, 1.8662)
Frank		(4.8095, 5.3752)	(4.5612, 5.5983)	(5.0631, 5.1209)
Joe		-	(2.3213, 2.8919)	(2.5762, 2.5998)

The intervals from the Wald and bootstrap method shown in Table 10 are generally wider, suggesting more variability in the estimates compared to the jackknife methods. While the Joe copula has the lowest RMSE, indicating a better fit in terms of prediction error, the Frank copula has the lowest MAE, and the Student t-copula has the lowest MRE. Considering all metrics, the Frank copula appears to be the best copula estimated for Kendall's Tau Estimation method due to its lowest MAE, and the Frank copula offers the most balanced and accurate fit across the key error metrics.

Spearman's Rho Estimation

The copula parameter estimation using Spearman's Rho is provided in Table 11. The estimated copula parameters $(\hat{\theta})$, their standard errors (SE), and GOF statistics for different copula families are:

Table 11: Copula Parameter Estimation using Spearman's Rho

Copula	$\hat{\theta}$	SE	RMSE	MAE	MRE
Clayton	1.5740	0.1490	0.4147	0.3362	-2.3417
Gumbel	1.7900	0.0730	0.4160	0.3374	-2.6109
Frank	4.6170	0.3370	0.4117	0.3364	-2.2312

Table 12: Spearman's Rho Interval

Copula	Wald	Bootstrap	Jackknife
Clayton	(1.2816, 1.8655)	(1.3271, 1.9446)	(1.5616, 1.5864)
Gumbel	(1.6461, 1.9342)	(1.6736, 1.9368)	(1.7839, 1.7961)
Frank	(3.9572, 5.2767)	(4.0629, 5.2941)	(4.5891, 4.6449)

The intervals from the Wald and bootstrap method presented in Table 12 are generally wider, suggesting more variability in the estimates compared to the jackknife methods. While the Frank copula has the lowest RMSE and MRE, indicating a better fit for prediction error and relative error, the Clayton copula has the lowest MAE.

Considering all metrics, the Frank copula appears to be the best copula estimated for Spearman's Rho Estimation method due to its lowest RMSE and MRE. The Frank copula offers the most balanced and accurate fit across the key error metrics.

Empirical Copula Estimation

The GOF statistics for empirical copula estimation are provided in Table 13. The GOF statistics for different copula families are:

Table 13: GOF statistics for Empirical copula Estimation

Copula	RMSE	MAE	MRE
Clayton	0.0069	0.0054	0.0812
Gumbel	0.0058	0.0043	0.0484
Frank	0.0065	0.0054	0.0670
Joe	0.0047	0.0039	0.0954

While the Joe copula has the lowest RMSE and MAE, indicating it best fits the copula, the Gumbel copula has the lowest MRE. The MRE for the Joe copula is the highest among the copula, suggesting it might have more relative error than others.

Considering all metrics, the Joe copula appears to be the best copula estimated for the empirical copula estimation due to its lowest RMSE and MAE despite having a higher MRE. The fit for prediction and absolute errors are more favourable with the Joe copula.

Robust Estimation by Maximum Mean Discrepancy Minimization (MMD)

The copula parameter estimation using the MMD method is provided in Table 14. The estimated copula parameters $(\hat{\theta})$ and GOF statistics for different copula families are:

Table 14: Copula Parameter Estimation using MMD

Copula		$\hat{\theta}$	RMSE	MAE	MRE
Student t	ρ	0.6100	0.4138	0.3402	-2.5992
	df	15.8300			
Clayton		1.0500	0.4163	0.3408	-2.8218
Gumbel		1.5600	0.4381	0.3623	-2.8294
Frank		3.6400	0.4047	0.3299	-1.9657
Joe		1.8200	0.4049	0.3271	-2.9226

Table 15: MMD Interval

Copula		Bootstrap	Jackknife
Student t	ρ	(0.5319, 0.6606)	(0.6084, 0.6116)
	df	-	(15.7937, 15.8664)
Clayton		(0.7536, 1.1850)	(0.7457, 1.3543)
Gumbel		(1.4423, 1.6785)	(1.4851, 1.6349)
Frank		(3.0681, 4.1531)	(2.9693, 4.3107)
Joe		(1.6607, 1.9166)	(1.8188, 1.8212)

The intervals from the bootstrap method shown in Table 15 are generally wider, suggesting more variability in the estimates compared to the jackknife method. The error measures show that

the Frank copula has the lowest RMSE and MRE, suggesting a better fit than the others. The Joe copula has the lowest MAE. Given these criteria, the Frank copula appears to be the best copula estimated for the MMD method, as it has the lowest RMSE, MRE, and one of the lowest MAE values.

CONCLUSION

The research contributes significantly to the field of hydrology by advancing the application of copula in modelling streamflow data. It introduces a comparative analysis of different copula models—such as the Student t-copula, Clayton copula, Gumbel copula, Frank copula, and Joe copula—across multiple estimation methods, including Maximum Pseudo-Likelihood Estimator (MPLE), Inference Functions for Margins Estimator (IFME), the Method-of-Moments Estimator (MoM), empirical copula estimation and Maximum Mean Discrepancy Minimization (MMD). The study comprehensively evaluates these copula models by considering three performance metrics—RMSE, MAE, and MRE.

One of the key contributions is identifying different copula best suited to specific estimation methods. For instance, the Student t-copula performed exceptionally well with the IFME. The Frank copula was the best for Kendall's tau, Spearman's Rho, and MMD estimation. The Joe copula was identified as the most suitable for the MPLE and the empirical copula estimation method.

Furthermore, the study makes a notable contribution by evaluating the precision of confidence intervals using various methods, concluding that the Jackknife method provides the most precise intervals, while the Wald and Bootstrap methods effectively capture variability and robustness. Overall, the research supports using copula, particularly the Frank and Joe copula, for effective hydrological modelling and accurate dependency estimation.

REFERENCES

Agresti, A. (2018). *Statistical Methods for the Social Sciences.* Pearson.

Alquier, P., Chérief-Abdellatif, B. E., Derumigny, A., & Fermanian, J. D. (2023). Estimation of copula via maximum mean discrepancy. *Journal of the American Statistical Association*, **118(543)**:1997-2012.

Buliah, N. A., & Yie, W. L. S. (2020, October). Modelling of extreme rainfall using copula. In *AIP Conference Proceedings* (Vol. 2266, No. 1). AIP Publishing.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. *Acad. Roy. Belg. Bull. Cl. Sci.*, 65, 274-292.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

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

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005, October). Measuring statistical dependence with Hilbert-Schmidt norms. In International conference on algorithmic learning theory (pp. 63-77). Berlin, Heidelberg: Springer Berlin Heidelberg.

Idiou, N., & Benatia, F. (2021). Simulation techniques of Archimedean Copula Estimators: Parametric and Semi-Parametric Approaches. *European Journal of Mathematics and Statistics*, **2(3)**:52-60.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall.

Joe, H., & Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. *Technical Report.*

Joo, K., Shin, J. Y., & Heo, J. H. (2020). Modified maximum pseudo likelihood method of copula parameter estimation for skewed hydrometeorological data. *Water*, **12(4)**:1182.

Karmakar, S., & Simonovic, S. P. (2009). Bivariate flood frequency analysis. Part 2: A copula‐based approach with mixed marginal distributions. *Journal of Flood Risk Management*, **2(1)**:32-44.

Ko, V., & Hjort, N. L. (2019). Model robust inference with two-stage maximum likelihood estimation for copula. *Journal of Multivariate Analysis*, **171**:362-381.

Kummaraka, U., & Srisuradetchai, P. (2023). Interval estimation of the dependence parameter in bivariate Clayton copula. *Emerging Science Journal*, **7(5)**:1478-1490.

Latif, S., & Mustafa, F. (2021). Bivariate joint distribution analysis of the flood characteristics under semiparametric copula distribution framework for the Kelantan River basin in Malaysia. *Journal of Ocean Engineering and Science*, **6(2)**:128-145.

Lokoman, R. M., & Yusof, F. (2019). Parametric estimation methods for bivariate copula in rainfall application. *Jurnal Teknologi*, **81(1)**.

Nelsen, R. B. (2006). *An Introduction to Copula.* Springer Science & Business Media.

Oh, D. H., & Patton, A. J. (2013). Simulated method of moments estimation for copula-based multivariate models. *Journal of the American Statistical Association*, **108(502)**:689-700.

Sahoo, B. B., Jha, R., Singh, A., & Kumar, D. (2020). Bivariate low flow return period analysis in the Mahanadi River basin, India using copula. *International Journal of River Basin Management*, **18(1)**:107-116.

Shaw, B., & Chithra, N. R. (2023). Copula-based multivariate analysis of hydro-meteorological drought. *Theoretical and Applied Climatology*, **153(1)**:475-493.

Shiau, J. T., & Lien, Y. C. (2021). Copula-based infilling methods for daily suspended sediment loads. *Water*, **13(12)**:1701.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**:229-231.

Tootoonchi, F., Sadegh, M., Haerter, J. O., Räty, O., Grabs, T., & Teutschbein, C. (2022). Copula for hydroclimatic analysis: A practice‐oriented overview. Wiley Interdisciplinary Reviews: *Water*, **9(2)**:e1579.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, **29(2)**:614-623.

Worland, S. C., Steinschneider, S., Farmer, W., Asquith, W., & Knight, R. (2019). Copula theory as a generalised framework for flow‐duration curve based streamflow estimates in ungaged and partially gaged catchments. *Water Resources Research*, **55(11)**:9378-9397.