4open 2020, **3**, 12
© L.M. Canno Ferreira Fais, et al., Published by EDP Sciences, 2020
https://doi.org/10.1051/fopen/2020011

Available online at:
www.4open-sciences.org

RESEARCH ARTICLE OPEN ACCESS

A copula based representation for tailings dam failures

Laura Maria Canno Ferreira Fais[1], Verónica Andrea González-López[2], Diego Samuel Rodrigues[1], and Rafael Rodrigues de Moraes[2,*]

[1] School of Technology, University of Campinas, Paschoal Marmo 1888, 13484-332 Limeira, S.P., Brazil
[2] Department of Statistics, University of Campinas, Sergio Buarque de Holanda 651, 13083-859 Campinas, S.P., Brazil

Received 16 March 2020, Accepted 13 August 2020

Abstract – In this article, we model the dependence between *dam factor* and $D_{\max}$, where *dam factor* is an indicator of risk of a tailings dam failure, which involves the height H of the tailings dam, the volume of material housed by the tailings dam V_T and the volume dispensed by the tailings dam, V_F, when the dam breaks. And, $D_{\max}$ is the maximum distance traveled by the material released by the tailings dam, after the collapse. With the dependence found via copula models and Bayesian estimation, given a range of *dam factor*, we estimate the probability of the released material to exceed a certain threshold. Since the *dam factor* involves the released volume V_F (unknown before the dam break), we present a naive way to estimate it using V_T and H. In this way, it is possible to estimate the *dam factor* of a tailings dam and with such a value to identify the probability of the tailings dam to show a $D_{\max}$ that exceeds a certain threshold.

Keywords: Copula models, Bayesian estimation, Conditional probability

Introduction

Tailings dams are scattered throughout various regions of the world, many of them already inactive, others still operating. Some countries have turned their attention to the conditions of such reservoirs, as entire regions have suffered from the impact of tailings dams breaks. The impacts of such ruptures have been incalculable, taking lives, leaving useless areas, and polluting rivers, projecting damages miles away from the reservoir site. Despite the demand imposed by several entities, the problem of regularizing these dams is complex due to several reasons. Many reservoirs are not used and their maintenance requires specialized equipment/technicians. In other cases, it is not easy to identify those responsible for the reservoirs because they work (or worked) for companies which are no longer in operation.

Several aspects must be taken into account in establishing the reliability of a tailings dam, the construction method, its size, the region where it is located, etc. For example, when talking about the region where it is located, factors such as the proximity of watercourses or cities raises the risk of the tailings dam in case of failure. Moreover, it is very common to find entire cities standing in the vicinity of reservoirs of this type, since the mining activity that surrounds the reservoir generates jobs and as a consequence facilitates the rising of communities that are sustained mainly by this activity. As expected, the databases that can help understand the dynamics of these tailings dams break are small (few cases), or at least those from which comparable information has been collected.

A valuable database of 35 tailings dam breaks reported around the world is presented by [1]. Information such as the volume of the material in a tailings dam (V_T) could be used to determine the volume of the material to be released in a dam breaking (V_F). The height H, attained by the material contained in the dam, and the volume of the released material could be used to estimate the distance reached by the released material ($D_{\max}$). Among the amounts to consider, the following two quantities are usually incorporated into the studies: (i) $H \times V_F$ (dam factor), see [2], (ii) $H_f = H \times \frac{V_F}{V_T} \times V_F$ (dam factor related to the fractional volume $\frac{V_F}{V_T}$), see [1]. In the literature there are several proposals to determine V_F and $D_{\max}$. The most recent according to our knowledge is also given in [1], where the authors propose a stochastic linear relationship between $\log(V_F)$ and $\log(V_T)$, and also a stochastic linear relationship between $D_{\max}$ (or $\log(D_{\max})$) and the *dam factor* version (i) or (ii).

Several aspects arise in the statistical perspective used in [1]. The first one is about the estimation of the parameters of the linear relations. A frequentist view as it is used could be compromised by the limited number of cases available in the

*Corresponding author: rafael.moraes@gmx.de database, since the law of large numbers (large sample size) is not feasible, which results in a lack of good properties in parameter estimators. To overcome this limitation, a Bayesian approach is then recommended. The second aspect presented in [1] is the imposition of linearity in the relationships to be estimated. This assumption is quite ambitious and, although the authors take the precaution of analyzing the goodness of the fit, the validity of the diagnoses also ends up being compromised by the size of the data set. Using the same data as [1], here we address both issues, making more flexible the relationship between V_T and V_F and also between *dam factor* and $D_{\max}$ using copula models and from a Bayesian perspective.

This article is organized as follows. Section Theoretical Background exposes the theoretical foundations of the work, with a focus on dependence models (see [3]). Inspecting the dependence relationships of the data provided in [1], the model selection procedure and the referred Bayesian parameter estimation are presented in Section Model Representation. Section Chances for Large Values of $D_{\max}$ shows a *dam factor*-based prediction study in order to determine the distance that the released material can travel after a breakdown. In this same section, we offer a strategy to identify the reservoir's *dam factor*. Finally, Section Conclusion closes the paper with some general remarks and conclusions.

Theoretical background

What really interests us is the relationship between X_1 and X_2, where X_1 is the variable V_T (or *dam factor*) and X_2 is the variable V_F (or $D_{\max}$) corresponding to tailings dams. Figures 1a, 2a and 3a show the scatter plots between the pairs (see [1]), with the most extreme cases highlighted.

We approach this problem using the concept of copula, where if H is the cumulative distribution function of (X_1, X_2), being those continuous random variables, there is a function C such that for all $(x, y) \in$ Image (X_1, X_2),

$$H(x,y) = C(F_1(x), F_2(y)), \quad \text{with} \quad F_1(x) = H(x,\infty) \quad \text{and} \quad F_2(y) = H(\infty, y). \tag{1}$$

If the function C is the 2-copula of (X_1, X_2), $C(u, v) =$ Prob $(F_1(X_1) \leq u, F_2(X_2) \leq v)$, for $u, v \in [0,1]$, then, C is the joint distribution of the variables $U := F_1(X_1)$ and $V := F_2(X_2)$. The function C is the one that we want to identify. The result given by (1) – see [4] – allows the decomposition of H into three functions: the marginal distribution F_1, the marginal distribution F_2, and the copula function C. F_1 and F_2 are marginally determined, that is, it is enough to inspect the stochastic behavior of X_1 to determine F_1 and similarly, X_2 is enough to determine F_2. Already to determine C, we need both variables duly transformed by their marginal distributions. Copula functions cover all dependence types, since $\forall (u, v) \in [0, 1]^2$ and for all copula $C(u, v)$,

$$W(u,v) = \max\{0, u + v - 1\} \leq C(u,v) \leq \min\{u,v\} = M(u,v). \tag{2}$$

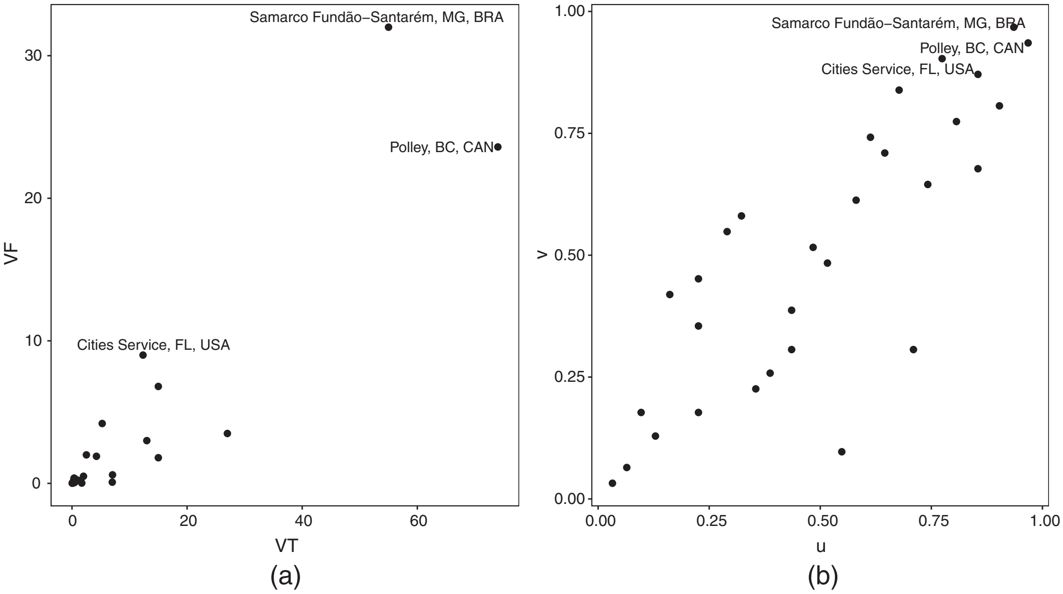
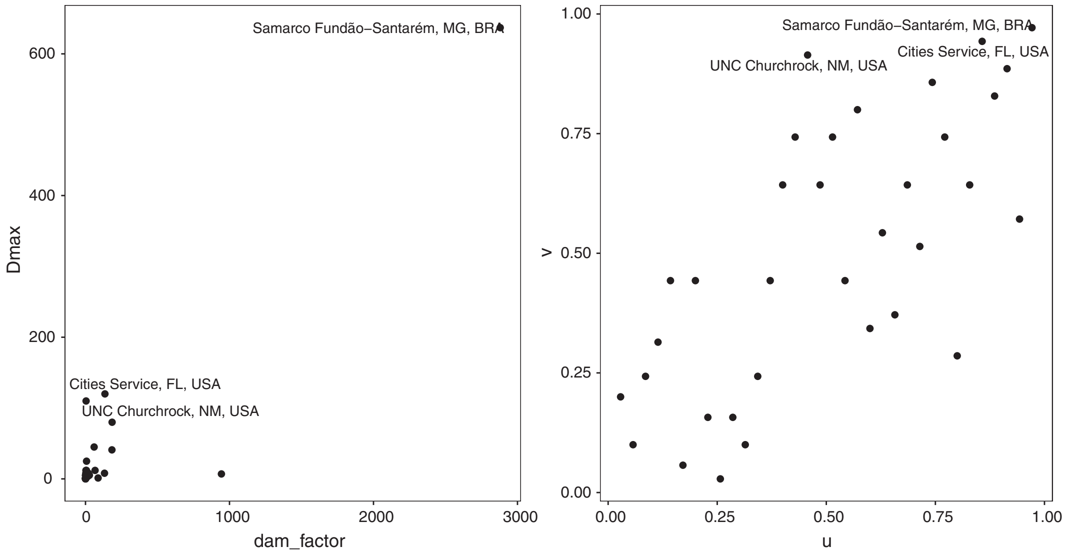
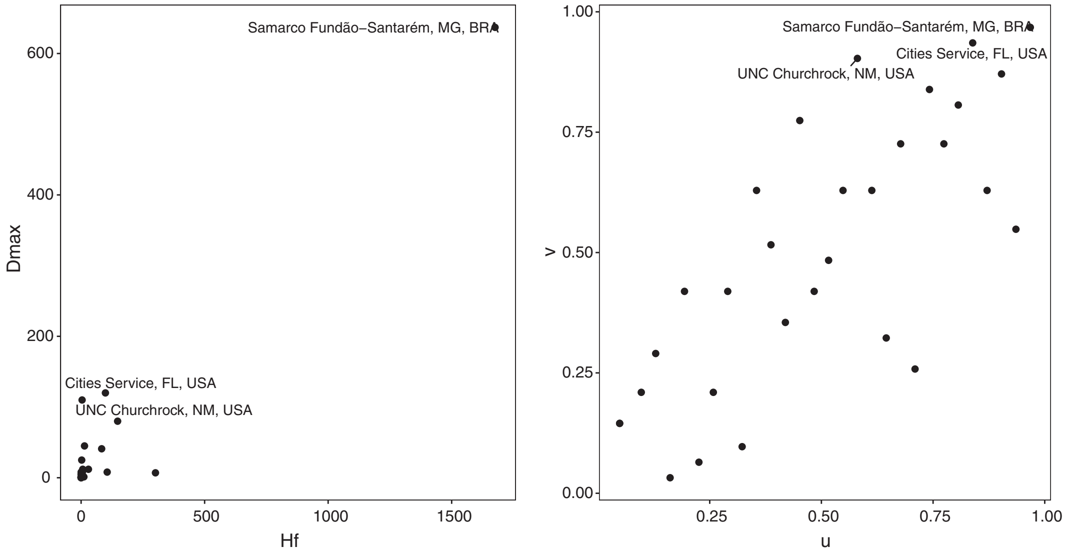


Figure 1. Scatterplot for V_T ($\times 10^6$ m^3) vs. V_F ($\times 10^6$ m^3) with (a) original observations and (b) observations transformed to [0, 1] by scaling ranks, U vs. V.



(a) (b)

Figure 2. Scatterplot for $H \times V_F$ vs. $D_{\max}$ with (a) original observations and (b) observations transformed to [0, 1] by scaling ranks, U vs. V. H in meters (m), $D_{\max}$ in kilometres (km).



(a) (b)

Figure 3. Scatterplot for H_f vs. $D_{\max}$ with (a) original observations and (b) observations transformed to [0, 1] by scaling ranks, U vs. V. H in meters (m), $D_{\max}$ in kilometres (km).

(X_1, X_2) has 2-copula M (W) if and only if X_2 is a monotone nondecreasing (nonincreasing) funcion of X_1 almost surely (see [5]). This property demonstrates that for X_1 and X_2 verify a linear relationship the 2-copula of (X_1, X_2) can only be M (Spearman correlation coefficient = 1) or W (Spearman correlation coefficient = −1). In the next example, we show a family of copulas indexed by a parameter $\theta \in [1, \infty)$. This allows covering the cases in the spectrum considered by the result (2).

Example 2.1

Consider $(u, v) \in [0, 1]^2$, $\theta \in [1, \infty)$, *the Gumbel–Hougaard copula is given by,*

$$C(u, v|\theta) := \exp(-((-\ln(u))^{\theta} + (-\ln(v))^{\theta})^{\frac{1}{\theta}}).$$

The Gumbel–Hougaard family is a family inside the Archimedean class (see [3]). That being, then there is a function ϕ: $[0, 1] \rightarrow [0, \infty]$ associated with this family which generates it. In this case, $\phi_\theta(t) := (-\ln(t))^\theta$ and the generation of C occurs since the following analytical expression is allowed:

$$C(u, v|\theta) = \phi_\theta^{-1}(\phi_\theta(u) + \phi_\theta(v)), \tag{3}$$

where $\phi_\theta^{-1}(z) = \exp(-z^{\frac{1}{\theta}})$. Note that $C(u,\ v|\theta = 1) = uv$ (independence between the random variables) and $C(u,\ v|\theta) \rightarrow M(u,\ v)$ when $\theta \rightarrow \infty$ (positive and strict dependence between the variables). This means that the Gumbel–Hougaard family allows to model positive dependence types with a spectrum that goes from independence to perfect positive dependence, not being possible for this family the cases from independence to the copula W. Also, the Kendall's tau coefficient is given by $\tau_\theta = \frac{\theta - 1}{\theta} \in [0,\ 1]$ because $\theta \in [1,\ \infty)$. So, the spectrum of τ_θ summarizes the positive dependence attained by the family. $\{C(\cdot, \cdot|\theta)\}_{\theta \in [1, \infty)}$ is a positively ordered family since, given θ, θ' such that $1 \leq \theta \leq \theta' < \infty$, $C(u,\ v|\theta) \leq C(u,\ v|\theta')$, $\forall (u,\ v) \in [0,\ 1]^2$. Features such as those mentioned for the Gumbel–Hougaard copula serve as a framework to interpret the dependence of random variables. This subject will be central in the next sections when the dependence between the variables of interest is determined.

In the application, we compare several copula models, some of them to be formulated considering the equation (3) (Archimedean copulas) with appropriate generators ϕ_θ, for which the pseudo-inverse of ϕ_θ is defined as $\phi_\theta^{[-1]}(s) := \phi_\theta^{-1}(s)$ when $0 \leq s \leq \phi_\theta(0)$ and $\phi_\theta^{[-1]}(s) := 0$ if $\phi_\theta(0) \leq s \leq \infty$. For instance, if $\phi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1)$, $\theta \in [-1,\ \infty)\backslash\{0\}$, the model generated by equation (3) is called Clayton family, if $\phi_\theta(t) := -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right)$, $\theta \in (-\infty,\ \infty)\backslash\{0\}$, the model generated by equation (3) is called Frank family. And, if $\phi_\theta(t) = -\ln(1 - (1 - t)^\theta)$, $\theta \in [1,\ \infty)$, the model generated by equation (3) is called Joe family. Clayton and Frank reach the limits W and M, see equation (2). Joe's family follows the same pattern as the Gumbel–Hougaard, only allowing cases from the independence to M. We also consider two well-known copulas belonging to the family of elliptical copulas with the shape,

$$C(u, v|\rho) := \psi(\psi^{-1}(u), \quad \psi^{-1}(v)|\rho), \quad (u, v) \in [0, 1]^2, \tag{4}$$

for an appropriate function ψ and parameter $\rho \in [-1,\ 1]$. (i) The Gaussian copula given by $\psi(t) := \Phi(t)$ which is the usual cumulative standard Gaussian distribution, $N(0,\ 1)$ and $\psi(s,\ t|\rho) := \Phi(s,\ t|\rho)$ which is the bivariate standard Gaussian distribution centered in zero, $N_2(\mathbf{0},\ \mathbf{P})$ with $\mathbf{P} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$; (ii) the t-Student copula given by $\psi(t) = T_\eta(t)$ which is the cumulative of the univariate t-Student distribution with η degrees and $\psi(s,\ t|\rho) := T_\eta(s,\ t|\rho)$ that is the bivariate cumulative t-Student distribution with η degrees of freedom and ρ correlation.

As can be verified by the examples of copulas registered here, there is a huge diversity concerning the possibilities of definition for the relationships between X_1 and X_2. Many of these are available in statistical libraries (see for example copula package in *R project* [R-project],[1] that in our case is used for the modeling).

With the variety of models introduced previously, we wish to cover a considerable range of dependence types that allow us to determine the best representation of the dependence between X_1 and X_2. For comparison between the models, we will adopt a model selection criterion (see [6]).

Model representation

In order to proceed with the estimation of the model, the original observations $\{(x_{1i}, x_{2i})\}_{i=1}^n$ are replaced by their re-scaled marginal ranks to $[0,\ 1]$, $u_i := \frac{|\{j:1 \leq j \leq n,\ x_{1j} \leq x_{1i}\}|}{n+1}$ and $v_i := \frac{|\{j:1 \leq j \leq n,\ x_{2j} \leq x_{2i}\}|}{n+1}$, $i = 1, \ldots, n$, where $|A|$ denotes the cardinal of the set A. This step has two purposes, namely to neutralize the effect arising from the (possibly) difference in measurement scale of the original observations and to mitigate the problems of evaluating the copula likelihood function in the boundaries of the region $[0,\ 1] \times [0,\ 1]$. This latter task is done by dividing the ranks by $n + 1$ so that the pseudo-observations, $\{u_{1i}\}_{i=1}^n$ (or $\{v_{2i}\}_{i=1}^n$), never attain 0 or 1. Pseudo observations are nonparametric estimates of $U = F_1(X_1)$ (or $V = F_2(X_2)$), furthermore, the function C is the distribution of $(U,\ V)$, which leads us to infer that the dependence exposed by equation (1) is revealed when exploring the dispersion between the pseudo-observations. See the scatterplot for the three cases: (1) V_T vs. V_F, (2) $H \times V_F$ vs. $D_{\max}$, (3) H_f vs. $D_{\max}$, of the values U vs. V represented by the pseudo-observations $\{(u_i, v_i)\}_{i=1}^n$, Figures 1b, 2b and 3b.

In order to guarantee some conditions required by the models, we test for stochastic independence and exchangeability. Originally proposed by [7], the independence test applied here, is based on the distance between the empirical copula and

[1] https://www.r-project.org.

the product copula $\Pi(u, v) = uv$ and is implemented using simulations executed by the indepTest() function from the copula R-package. We test H_0: U and V are independent and, H_0 is rejected with p-value < 0.001 in all the cases. Such evidence supports the search for a model that represents the dependence. A rather desirable property of the dependence is the exchangeability, a condition requested by many families of copulas including the Archimedean and the elliptical ones. A copula is exchangeable if the property $C(u, v|\theta) = C(v, u|\theta)$ is verified for all $(u, v) \in [0, 1]^2$ and the test proposed by [8], based on the empirical copula, is implemented using simulations by the function exchTest() of the copula R-package. Then, we test if H_0: U and V are exchangeable, and H_0 is not rejected in all the cases, with p-values much greater than the generally used significance level of 5%. Such evidence supports to use the models given by Archimedean and elliptical copulas. In order to define the appropriate copula in the three situations we use the function fitCopula(), with arguments (a) copula and (b) method being (a) "claytonCopula(dim = 2)", "frankCopula(dim = 2)", "gumbelCopula(dim = 2)", "joeCopula(dim = 2)", "normalCopula(dim = 2)", "tCopula(dim = 2)" and being (b) method = "mpl" (maximum pseudo likelihood) which is the maximum log-likelihood (MLL) method evaluated on the pseudo observations.

Given a copula C its density c is computed and the log-likelihood is given by $\ln(\prod_{i=1}^{n} c(u_i, v_i))$ which is maximized in the underlying parameters to obtain MLL $(C, \{(u_i, v_i)\}_{i=1}^{n})$, related to the model C and the set $\{(u_i, v_i)\}_{i=1}^{n}$. Note that five of these models have one parameter while the t-Student copula model has two parameters, so that a penalty is applied to this model in order to promote a fair selection. We consider the *Bayesian Information Criterion* (BIC) for this purpose, see [6],

$$\text{BIC}\,(C, \{(u_i, v_i)\}_{i=1}^{n}) := \text{MLL}\,(C, \{(u_i, v_i)\}_{i=1}^{n}) - \frac{1}{2} N\ \ln(n), \tag{5}$$

where N is the total number of parameters of C and $n = 30$ is the number of observations in the dataset. The BIC takes into account the trade-off between the maximum log-likelihood, MLL $(C, \{(u_i, v_i)\}_{i=1}^{n})$, and the penalty imposed by the second term, $\frac{1}{2} N\ \ln(n)$, which is done to enforce parsimony, i.e., simpler models having fewer parameters. According to the BIC, the higher the value taken by the equation (5), the better the model.

In the next subsection, we show the results produced by the selection procedure (selecting the best copulas) and also the estimation of the parameters. For the latter, we show the classical estimator as well as two Bayesian versions.

Estimation

We note that in the three cases, the two best models (copulas) are the Gaussian and the Gumbel–Hougaard (see Tabs. 1 and 2). It is clear that due to the size of the database ($n = 30$) the reliability of the classical parameter estimates may not be the best possible. This aspect can be addressed in a timely manner, but for now we can get some conclusions about the type of relationship in each case.

The Gaussian copula confirms that assuming linearity between X_1 and X_2 could not be indicated. If the relationship between these variables were in fact linear, the Gumbel–Hougaard copula could have detected this fact by attributing a very large estimate for θ (in the direction of the $M(u, v)$ copula), see Section Theoretical Background. For all the cases, (1) V_T vs. V_F, (2) $H \times V_F$ vs. $D_{\max}$, and (3) H_f vs. $D_{\max}$, the estimation of θ is moderate $\hat{\theta} \in (2, 3]$ showing evidence of the non-applicability of the M copula.

In order to introduce a level of confiability to the parametric estimation process, we conduct a Bayesian estimation. We apply Hamiltonian Monte Carlo (HMC) simulations through the rstan R-package in two Bayesian strategies, described below and in general terms, Non-Informative (NI) and Informative (I). In Table 3, we show the results for each of the three cases and for each of the two best models pointed out by the BIC, see Tables 1 and 2. For the Non-Informative cases of the Gaussian copula, we use an uniform prior distribution for ρ, properly accommodated in the interval $[-1, 1]$, which is the range of possibilities of the ρ parameter. The Informative cases of Gaussian copula are implemented assuming a Beta priori distribution for ρ, accommodated in $[-1, 1]$, where its mode is the one derived from the relation between Kendall's tau coefficient and the copula association parameter. For instance, by means of the empirical estimation of Kendall's tau coefficient we can obtain an estimation of the parameter (see iTau() function from copula R package). The NI cases of

Table 1. Parameters estimated by *Maximum Log-Likelihood (MLL)* and BIC values for the copula between the pseudo-observations *ranks*(V_T) and *ranks*(V_F); degrees of t-Student $\hat{\eta} = 3.529$ also estimated by MLL using the tCopula() function of the copula R-package.

Copula	Parameter estimate	BIC
Gaussian	0.875	17.619
Gumbel H	2.950	17.484
t	0.868	16.620
Frank	9.385	15.267
Joe	3.641	15.125
Clayton	2.929	14.814

Table 2. Parameters estimated by *Maximum Log-Likelihood (MLL)* and BIC values for the copula between the pseudo-observations *ranks(dam factor)* and *ranks*($D_{\max}$). Left: *dam factor* = $H \times V_F$, $\hat{\eta} = 12378.75$; Right: *dam factor* = H_f, $\hat{\eta} = 1479.85$. In both cases the degrees η are estimated by MLL using the tCopula() function of the copula R-package.

$H \times V_F$ vs. $D_{\max}$			H_f vs. $D_{\max}$		
Copula	Parameter estimate	BIC	Copula	Parameter estimate	BIC
Gaussian	0.744	9.902	Gaussian	0.799	11.306
Gumbel H	2.055	9.869	Gumbel H	2.329	11.015
Frank	5.826	9.076	Frank	7.237	10.708
Joe	2.548	8.750	*t*	0.799	9.602
t	0.744	8.201	Joe	2.888	9.435
Clayton	1.254	5.993	Clayton	1.558	7.387

Table 3. Summaries of the Bayesian estimation.

	prior	mean	s–m	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
					Case V_T vs. V_F – see Table 1						
Gaussian	NI	**0.856**	0.001	0.043	0.752	0.834	**0.863**	0.886	0.916	1161	1.000
Gaussian	I	**0.859**	0.001	0.040	0.765	0.838	**0.865**	0.887	0.918	1198	1.003
Gumbel H	NI	**3.023**	0.012	0.445	2.243	2.707	**3.001**	3.325	3.938	1315	1.003
Gumbel H	I	**3.021**	0.012	0.444	2.222	2.712	**3.005**	3.298	3.931	1320	1.001
					$H \times V_F$ vs. $D_{\max}$ – see Table 2 (left)						
Gaussian	NI	**0.696**	0.003	0.088	0.490	0.648	**0.712**	0.760	0.822	1167	1.002
Gaussian	I	**0.707**	0.002	0.082	0.506	0.663	**0.722**	0.765	0.824	1305	1.000
Gumbel H	NI	**2.100**	0.009	0.317	1.512	1.881	**2.089**	2.302	2.743	1158	1.003
Gumbel H	I	**2.111**	0.009	0.319	1.540	1.883	**2.089**	2.317	2.805	1222	1.001
					H_f vs. $D_{\max}$ – see Table 2 (right)						
Gaussian	NI	**0.768**	0.002	0.067	0.612	0.732	**0.779**	0.817	0.864	1286	1.002
Gaussian	I	**0.771**	0.002	0.065	0.618	0.736	**0.782**	0.819	0.864	1192	1.003
Gumbel H	NI	**2.392**	0.010	0.358	1.746	2.135	**2.385**	2.629	3.120	1273	1.001
Gumbel H	I	**2.399**	0.012	0.362	1.742	2.144	**2.372**	2.642	3.135	959	1.001

n_eff: final number of simulations used for the estimation; sd: standard deviation; s–m = sd/n_eff$^{1/2}$; Rhat: potential scale reduction factor on split chains (at convergence, Rhat = 1). In bold, the Bayesian estimates of ρ for Gaussian copula and θ for Gumbel–Hougaard copula, by quadratic loss function on left, by multilinear loss function on right. On the top, Non-Informative (NI) prior; on the bottom, Informative (I).

Gumbel–Hougaard copula are implemented assuming as a prior distribution on θ an improper prior distribution (proportional to 1), while the Informative cases are implemented by assuming a Gamma distribution on θ accommodated in $[1, \infty)$ with mode indicated by the parameter estimation derived from the iTau() function.

While the informative scenario (I) aims to follow the trend of the data, the non-informative (NI) seeks objectivity, without risking in declaring trends. Table 3 shows convergence indicators of the simulations and a descriptive summary of the values taken by the posterior distributions of ρ (or θ) obtained by the simulations. We note that Bayesian estimates (by quadratic and multilinear loss) return values close to the classic estimates reported in Tables 1 and 2, which gives coherence to the adjustments.

In this section, we have determined the best representation of the dependence between X_1 and X_2 in each of the three cases, which happens to be given by the Gaussian copula in the first instance and by the Gumbel–Hougaard copula in the second instance. We also implemented the Bayesian estimation of the parameters involved, determining among other aspects an estimation of the dependence structure, given by $C(u, v|\hat{\rho}_B)$ in the case of Gaussian copula (Eq. (4)) and $C(u, v|\hat{\theta}_B)$ for the Gumbel–Hougaard copula (Example 2.1), where $\hat{\rho}_B$ and $\hat{\theta}_B$ are the Bayesian estimators given by a certain loss function.

Chances for large values of $D_{\max}$

In this section, we approach the specific problem of calculating the probability that the material released by the dam will exceed a certain distance, given a specific risk represented by dam factor values. To address this issue, we use the copula determined by the BIC and the Bayesian estimate presented in Table 3. Given that the perspective via copulas translates the problem to the ranks of the observations, we first address the questions for the ranks of the observations, in this paper denoted by U and V. Now we operate with the vector $(U,\ V)$ represented by the pseudo-observations $\{(u_i, v_i)\}_{i=1}^{n}$ coming from the paired values of dam factor and $D_{\max}$ transformed to [0, 1] by marginal scaling ranks. Then, for each threshold $v \in [0,\ 1]$ and interval $(a,\ b] \subset [0,\ 1]$,

$$\begin{aligned}\text{Prob}\,(V > v|U \in (a,b]) &= 1 - \text{Prob}\,(V \le v|U \in (a,b]) = 1 - \frac{\text{Prob}\,(V \le v, U \in (a,b])}{\text{Prob}\,(U \in (a,b])}\\ &= 1 - \frac{\text{Prob}\,(V \le v, U \le b) - \text{Prob}\,(V \le v, U \le a)}{\text{Prob}\,(U \le b) - \text{Prob}\,(U \le a)} = 1 - \frac{C(b,v)}{b-a} + \frac{C(a,v)}{b-a}.\end{aligned} \tag{6}$$

Under the Gaussian copula, we estimate the probability (6) using the Bayesian estimator by quadratic loss function of ρ, say $\widehat{\rho}_B$. In the same way, under the Gumbel–Hougaard copula, we estimate the probability (6) using the Bayesian estimator by quadratic loss function of θ, say $\hat{\theta}_B$. Then, the estimators of (6) are,

$$\widehat{\text{Prob}}\,(V > v|U \in (a,b]) = 1 - \frac{C(b,v|\widehat{\rho}_B)}{b-a} + \frac{C(a,v|\widehat{\rho}_B)}{b-a}, \tag{7}$$

and,

$$\widehat{\text{Prob}}\,(V > v|U \in (a,b]) = 1 - \frac{C(b,v|\widehat{\theta}_B)}{b-a} + \frac{C(a,v|\widehat{\theta}_B)}{b-a}, \tag{8}$$

respectively. Each one of the estimates (Eqs. (7) and (8)) presents two versions, one under the *Non-Informative* setting and another under the *Informative* setting, see Table 3. Tables 4 and 5 record the estimated values according to equations (7) and (8). Table 4 shows the estimates when the *dam factor* is $H \times V_F$ and Table 5 shows the estimates when the *dam factor* is H_f. We note that the highest probabilities occur with the estimate implemented by the Gumbel–Hougaard copula (Eq. (8)) for the intervals of the extremes (0, 0.25] and (0.75, 1], while for the central intervals, the Gaussian copula (Eq. (7)) offers the highest values. This behavior is displayed for both types of *dam factor* (Tabs. 4 and 5). The Gumbel–Hougaard copula is well known as having more realistic properties in extreme values (see [9]); in the copula's framework, this means near to the values 0 or 1.

Table 4. Case $H \times V_F$ *vs.* $D_{\max}$ (see Tab. 3). From left to right: interval $(a,\ b]$, v lower threshold of V, $P_i = \text{Prob}_i\,(V > v|U \in (a,\ b])$ following equation (8), $i = 1, 2$ and equation (7), $i = 3, 4$, *Non Informative* settings $i = 1, 3$ and *Informative* settings $i = 2, 4$. In bold, the highest probability per line.

Interval $(a,\ b]$	Threshold v	Gumbel H P_1 (NI)	Gumbel H P_2 (I)	Gaussian P_3 (NI)	Gaussian P_4 (I)
(0, 0.25]	0.7000	**0.0369**	0.0362	0.0353	0.0325
	0.7500	**0.0238**	0.0233	0.0224	0.0204
	0.8000	**0.0141**	0.0137	0.0130	0.0116
	0.8500	**0.0073**	0.0071	0.0065	0.0057
	0.9000	**0.0029**	0.0028	0.0025	0.0021
	0.9500	**0.0006**	**0.0006**	0.0005	0.0004
(0.25, 0.5]	0.7000	0.1114	0.1104	**0.1520**	0.1477
	0.7500	0.0739	0.0731	**0.1087**	0.1048
	0.8000	0.0447	0.0441	**0.0718**	0.0686
	0.8500	0.0235	0.0231	**0.0419**	0.0394
	0.9000	0.0096	0.0094	**0.0195**	0.0180
	0.9500	0.0021	0.0021	**0.0052**	0.0046
(0.5, 0.75]	0.7000	0.3037	0.3035	**0.3414**	0.3411
	0.7500	0.2215	0.2211	**0.2696**	0.2684
	0.8000	0.1460	0.1454	**0.1996**	0.1978
	0.8500	0.0824	0.0818	**0.1336**	0.1314
	0.9000	0.0354	0.0350	**0.0742**	0.0721
	0.9500	0.0081	0.0080	**0.0260**	0.0247
(0.75, 1]	0.7000	0.7480	**0.7499**	0.6712	0.6786
	0.7500	0.6808	**0.6826**	0.5993	0.6064
	0.8000	0.5952	**0.5968**	0.5156	0.5220
	0.8500	0.4869	**0.4881**	0.4181	0.4235
	0.9000	0.3520	**0.3528**	0.3039	0.3078
	0.9500	0.1891	**0.1894**	0.1683	0.1702

Table 5. Case H_f vs. $D_{\max}$ (see Tab. 3). From left to right: interval $(a,\ b]$, v lower threshold of V, $P_i = \text{Prob}_i\ (V > v | U \in (a,\ b])$ following equation (8), $i = 1$, 2 and equation (7), $i = 3$, 4, *Non Informative* settings $i = 1$, 3 and *Informative* settings $i = 2$, 4. In bold, the highest probability per line.

Interval $(a,\ b]$	Threshold v	Gumbel P_1 (NI)	Gumbel H P_2 (I)	Gaussian P_3 (NI)	Gaussian P_4 (I)
(0, 0.25]	0.7000	**0.0220**	0.0218	0.0179	0.0172
	0.7500	**0.0133**	0.0131	0.0102	0.0097
	0.8000	**0.0073**	0.0072	0.0051	0.0049
	0.8500	**0.0034**	**0.0034**	0.0022	0.0020
	0.9000	**0.0012**	**0.0012**	0.0006	0.0006
	0.9500	**0.0002**	**0.0002**	0.0001	0.0001
(0.25, 0.5]	0.7000	0.0868	0.0863	**0.1195**	0.1179
	0.7500	0.0539	0.0535	**0.0799**	0.0785
	0.8000	0.0301	0.0299	**0.0484**	0.0473
	0.8500	0.0144	0.0142	**0.0251**	0.0244
	0.9000	0.0051	0.0051	**0.0099**	0.0095
	0.9500	0.0009	0.0009	**0.0020**	0.0018
(0.5, 0.75]	0.7000	0.2992	0.2991	**0.3378**	0.3375
	0.7500	0.2093	0.2090	**0.2589**	0.2582
	0.8000	0.1297	0.1294	**0.1838**	0.1829
	0.8500	0.0670	0.0666	**0.1156**	0.1146
	0.9000	0.0253	0.0251	**0.0582**	0.0573
	0.9500	0.0047	0.0046	**0.0169**	0.0164
(0.75, 1]	0.7000	0.7920	**0.7929**	0.7248	0.7274
	0.7500	0.7235	**0.7244**	0.6511	0.6536
	0.8000	0.6329	**0.6336**	0.5627	0.5649
	0.8500	0.5152	**0.5158**	0.4571	0.4590
	0.9000	0.3684	**0.3687**	0.3313	0.3326
	0.9500	0.1942	**0.1943**	0.1811	0.1816

In Tables 6 and 7, we associate the $[a,\ b)$ intervals $(a,\ b \in [0,\ 1])$ with *dam factor* intervals in the originally observed scale. We also associate the v values with the $D_{\max}$ thresholds in the original scale, according to the original data. Since it is a small database of 30 observations, each interval (in Tab. 6) on the original scale corresponds to a small number of cases, on average approximately $30/4$.

To compute probabilities in real cases, it is enough to identify on the original scale the *dam factor* interval associated with $U \in (a,\ b]$ and also the threshold in V, v associated with $D_{\max}$. For instance, consider the case 11 of [1], first line in Table 8. Both versions of *dam factor* are in the interval (66.5, 2880], which is associated with $(a,\ b] = (0.75,\ 1]$, see Table 6. This means that, according to the estimations given by Tables 4 and 5, the probability $P(V > 0.75 | U \in (0.75,\ 1])$ is ≥ 0.6. While the threshold $v = 0.75$ is related to $D_{\max} = 25$ (see Tab. 7). This already indicates that there is a high probability of the released material traveling 25 km or more. Besides, the computation by the Gumbel–Hougaard copula offers estimates that exceed 0.7. In fact, this case has an observed $D_{\max} = 120$. This simple exercise of using Tables 4 and 5 shows the ability of the copulas to identify risks in real situations, for example, if one wants to calculate the probability of the material released reaching a certain point as a city, a river, etc. The practical impact of these estimates also exposes the need of defining an estimator of the *dam factor*. That, according to the notions used here $H \times\ V_F$ and $H_f = H \times \frac{V_F}{V_T} \times\ V_F$, implies estimating V_F.

The next subsection is intended to estimate V_F, since an adequate estimate of V_F guarantees a reliable estimate of the *dam factor*. For such, we will take the V_T value as an available information.

Expected Value for V_F given V_T

As proved in [10], the expected value $\mathbb{E}(V|U \in (c, d])$ can be computed by integrating the copula function,

$$\mathbb{E}(V|U \in (c, d]) = 1 - \frac{1}{d - c}\left(\int_0^1 C(d, v)\mathrm{d}v - \int_0^1 C(c, v)\mathrm{d}v\right). \tag{9}$$

Under the Gaussian copula we estimate the equation (9) by means of the Bayesian estimator by quadratic loss function of ρ, say $\widehat{\rho}_B$ (see Tab. 3). The estimator of equation (9) is defined as,

$$\hat{\mathbb{E}}(V|U \in (c, d]) = 1 - \frac{1}{d - c}\left(\int_0^1 C(d, v|\widehat{\rho}_B)\mathrm{d}v - \int_0^1 C(c, v|\widehat{\rho}_B)\mathrm{d}v\right). \tag{10}$$

Table 6. Relation between the intervals $(a,\ b]$ and the *dam factor* intervals.

$(a,\ b]$	$(0,\ 0.25]$	$(0.25,\ 0.5]$	$(0.5,\ 0.75]$	$(0.75,\ 1]$
$H \times V_F$ interval	$(0.066,\ 1.08]$	$(1.08,\ 7.31]$	$(7.31,\ 66.5]$	$(66.5,\ 2880]$
H_F interval	$(0.01,\ 0.22]$	$(0.22,\ 3.24]$	$(3.24,\ 13.85]$	$(13.85,\ 1675.64]$

Table 7. Relation between v and the $D_{\max}$ thresholds.

v	0.7	0.75	0.8	0.85	0.9	0.95
$D_{\max}$ threshold	12	25	41	45	110	120

Table 8. Special cases.

Place	Year	H	V_T	$D_{\max}$	V_F	$H \times V_F$	H_f
Cities Service, Fort Meade, Florida	1971	15	12.34	120	9	135	98.46
Minas Gerais, Brazil (Samarco)	2015	90	55	637	32	2880	1675.64

Table 9. $\mathbb{E}_i$, $i = 1, 2$ computed using Gumbel–Hougaard copula (Eq. (11)), $\mathbb{E}_i$, $i = 3, 4$ computed using Gaussian copula (Eq. (10)), $i = 1, 3$ from *Non Informative* settings, $i = 2, 4$ from *Informative* settings. In brackets, the standard deviation of the expected values. In bold, the highest mean expected values per line.

Interval $(c,\ d]$	Gumbel H (NI) $\hat{\mathbb{E}}_1(V\vert U \in (c,\ d])$	Gumbel H (I) $\hat{\mathbb{E}}_2(V\vert U \in (c,\ d])$	Gaussian (NI) $\hat{\mathbb{E}}_3(V\vert U \in (c,\ d])$	Gaussian (I) $\hat{\mathbb{E}}_4(V\vert U \in (c,\ d])$
$(0,\ 0.25]$	0.1976 (18.9 × 10^{-3})	**0.1977** (19.0 × 10^{-3})	0.1826 (16.8 × 10^{-3})	0.1815 (15.7 × 10^{-3})
$(0.25,\ 0.50]$	0.3888 (5.09 × 10^{-3})	0.3889 (5.13 × 10^{-3})	**0.4033** (7.03 × 10^{-3})	0.4029 (6.65 × 10^{-3})
$(0.50,\ 0.75]$	0.5826 (9.49 × 10^{-3})	0.5826 (9.51 × 10^{-3})	0.5967 (7.03 × 10^{-3})	**0.5971** (6.65 × 10^{-3})
$(0.75,\ 1]$	**0.8309** (14.5 × 10^{-3})	0.8308 (14.6 × 10^{-3})	0.8174 (16.8 × 10^{-3})	0.8185 (15.7 × 10^{-3})

In the same way, and under the Gumbel–Hougaard copula we estimate the equation (9) using the Bayesian estimator by quadratic loss function of θ, say $\hat{\theta}_B$ (see Tab. 3). Here is the estimator:

$$\hat{\mathbb{E}}(V|U \in (c,d]) = 1 - \frac{1}{d-c}\left(\int_0^1 C(d,v|\widehat{\theta}_B)\mathrm{d}v - \int_0^1 C(c,v|\widehat{\theta}_B)\mathrm{d}v\right). \qquad (11)$$

In Table 9, we clearly visualize the effect of the copula on the magnitude of the mean values, the Gumbel–Hougaard copula offers the highest values in the extreme intervals and the Gaussian copula does it in the central intervals. As expected, with increasing the U interval (V_T scaled to $[0,\ 1]$) the mean of V increases (V_F scaled to $[0,\ 1]$).

In Table 10, we record in the first column the ranges of V_F values corresponding to the intervals to the left of Table 9, these last ranges made in the percentages of V_F. Table 10 offers naive estimates for V_F, based on V_T intervals. For example, $\hat{\mathbb{E}}_2(V_F|V_T \in (c,d])$ is the estimator of V_F when $V_T \in (c,\ d]$ and computed from the Gumbel–Hougaard copula with an Informative setting.

Next, we compare the intervals in which the naive estimates fall, with those reported in Table 8 and, according to Table 6. Regarding line 1 of Table 8: since $V_T = 12.34$, this value is associated with the interval (7.040, 74.00] of Table 10, which brings us to an average of V_F estimated as $\hat{\mathbb{E}}_i(V_F|V_T \in (7.04,\ 74]) = 32$. When calculating the dam factor-(i), we obtain $H \times \hat{\mathbb{E}}_i(V_F|V_T \in (7.04,\ 74]) = 15 \times 32 = 480 \in (66.5,\ 2880]$, note that the observed value 135 (Tab. 8) is also in the interval (66.5, 2880]. Computing the dam factor-(ii), we obtain $H \times \frac{\hat{\mathbb{E}}_i(V_F|V_T \in (7.04,74])}{V_T} \times \hat{\mathbb{E}}_i(V_F|V_T \in (7.04,\ 74]) = 15 \times 32/12.34 \times 32 = 1244.733 \in (13.85,\ 1675.64]$, which is in fact the interval in which was found the observed data (98.46). In other words, the estimates and observed values are accommodated in the same risk interval, reported in Table 6. If we consider the Samarco case (second case in Tab. 8), we have that the estimation is $\hat{\mathbb{E}}_i(V_F|V_T \in (7.04,\ 74]) = 32$, since the observed $V_T = 55 \in (7.04,\ 74]$, and in fact the observed value is $V_F = 32$. Then, the dam factor estimates are the same as those recorded in Table 8.

Table 10. Original scale. $\mathbb{E}_i$, $i = 1, 2$ computed using Gumbel–Hougaard copula, $\mathbb{E}_i$, $i = 3, 4$ computed using Gaussian copula, $i = 1, 3$ from *Non Informative* settings, $i = 2, 4$ from *Informative* settings.

Interval $(c, d]$	Gumbel H (NI) $\hat{\mathbb{E}}_1(V_F \| V_T \in (c, d])$	Gumbel H (I) $\hat{\mathbb{E}}_2(V_F \| V_T \in (c, d])$	Gaussian (NI) $\hat{\mathbb{E}}_3(V_F \| V_T \in (c, d])$	Gaussian (I) $\hat{\mathbb{E}}_4(V_F \| V_T \in (c, d])$
(0.038, 0.300]	0.021	0.021	0.021	0.021
(0.300, 1.520]	0.085	0.085	0.085	0.085
(1.520, 7.040]	0.600	0.600	0.600	0.600
(7.040, 74.00]	32.00	32.000	32.000	32.000

Conclusion

In this article, we model the dependence between each version of the *dam factor* and $D_{\max}$ with *dam factor* given by the risk factor of a reservoir that depends on its dam's height (total volume) H (V_T) and the volume of material released V_F after a collapse. $D_{\max}$ is given by the maximum distance reached by the dispensed material. In this way, it is possible to accurately describe the performance of the dependence relationship between each *dam factor* and $D_{\max}$. Furthermore, using the BIC criterion we can point out the best representation, given in our case by the Gaussian copula (see Tab. 2). We were also able to delineate the dependence between the volume available from a reservoir V_T and the material's volume V_F capable of being expelled after a collapse (see Tab. 1).

Through a Bayesian study, we offer scenarios that aim to give flexibility to the representation of the dependence. Under an assumption of lack of prior information, the NI scenario would be the one indicated (see Tab. 3). With the alliance between Bayesian tools and copula models we can determine tail probabilities of $D_{\max}$ values scaled to [0, 1] and conditioned at *dam factor* intervals (see Tabs. 4 and 5). With them, we determine risks for reservoirs, depending on the *dam factor*.

We also present in this article a Bayesian and non-parametric method to infer the *dam factor* of a reservoir (see Sect. Expected Value for V_F Given V_T). We show its efficiency for some available cases in the literature. Note that such a method could be improved by univariate modeling techniques in future research studies. We see that the Gaussian copula has a competitor, which is the Gumbel–Hougaard copula (Tabs. 1 and 2), which offers slightly different (and higher) probabilities for extreme events (see Tabs. 4 and 5). On the other hand, when using the Gumbel–Hougaard copula in calculating the *dam factor* (see Tabs. 9 and 10), it produces results comparable to those of the Gaussian copula, which shows that the notion of *mean* hides certain stochastic characteristics. For a complement of other aspects addressed in this problem, see [11]. We consider this work as the starting point of other approaches enriched with marginal models, since here we treat the marginal effect only from a non-parametric perspective (ranks), which may show restrictions, in the presence of few data.

Acknowledgments

R. Rodrigues de Moraes gratefully acknowledge the partial financial support provided by CAPES and CNPq with fellowships from the Master Program in Statistics of University of Campinas. Also, the authors wish to thank the two referees for their many useful comments and suggestions on an earlier draft of this paper.

References

1. Concha Larrauri P, Lall U (2018), Tailings dams failures: Updated statistical model for discharge volume and runout. Environments 5, 2, 28.
2. Rico M, Benito G, Diez-Herrero A (2008), Floods from tailings dam failures. J Hazard Mater 154, 1–3, 79–87.
3. Nelsen RB (2007), An introduction to copulas. Springer Science & Business Media.
4. Sklar M (1959), Fonctions de repartition à n dimensions et leurs marges. Publ Inst Statist Univ Paris 8, 229–231.
5. Mikusinski P, Sherwood H, Taylor MD (1991), The Fréchet bounds revisited. Real Anal Exch 17, 2, 759–764.
6. Schwarz G (1978), Estimating the dimension of a model. Ann Stat 6, 2, 461–464.
7. Deheuvels P (1981), A non parametric test for independence. Publ Inst Stat Univ Paris 26, 29–50.
8. Genest C, Nešlehová J (2012), Tests of symmetry for bivariate copulas. Ann Inst Stat Math 64, 4, 811–834.
9. Genest C, Rivest LP (1989), A characterization of Gumbel's family of extreme value distributions. Stat Prob Lett 8, 3, 207–211.
10. González-López VA, Rodrigues de Moraes R (2020), A copula-based quantifying of the relationship between race inequality among neighbourhoods in São Paulo and age at death, 4open 3, 11.
11. Rodrigues de Moraes R (2020), Eventos Caudais na Prática. Modelagem Bayesiana via Cópulas (Unpublished Master's Thesis).

Cite this article as: Canno Ferreira Fais LM, González-López VA, Rodrigues DS & Rodrigues de Moraes R 2020. A copula based representation for tailings dam failures. 4open, **3**, 12.