14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Probabilistic Modelling of Soil Shear Strength by Maximum Entropy Quantile Functions

Jian Deng
Associate Professor, Dept. of Civil Engineering, Lakehead University, Thunder Bay, Canada

Jasleen Kaur Bheora
Research Student, Dept. of Civil Engineering, Lakehead University, Thunder Bay, Canada

Shumsun Siddique
Graduate Student, Dept. of Civil Engineering, Lakehead University, Thunder Bay, Canada

Eltayeb Eltahir Mohamedelhassan
Professor, Dept. of Civil Engineering, Lakehead University, Thunder Bay, Canada

Marcus Hardy, Alyssa Ray
Red Rock Indian Band, Nipigon, Ontario, Canada

ABSTRACT: This paper proposes a novel probabilistic method to model soil properties using quantile functions, based on fractional probability-weighted moments, principle of maximum entropy, and Akaike information criterion. The quantile function is a counterpart to distribution functions of a random variable since the quantile function is mathematically the inverse cumulative distribution function. The maximum entropy method is presented to generate unbiased quantile functions for measured soil properties. The use of the fractional probability-weighted moments facilitates more accurate quantification of soil uncertainties by the entropy-based quantile functions than the probability density or cumulative distribution functions. Akaike information criterion is then used to locate the optimal order of maximum entropy quantile functions. Maximum entropy quantile distributions are compared to the traditional quantile distributions to evaluate their performance. The analytical entropy quantile distribution obtained can be used in probabilistic reliability analysis.

1. INTRODUCTION

In application of probability and statistics in the planning, analysis, and design of civil engineering systems, quantification of uncertainty is an essential step before the probability of failure can be estimated. This includes collection of relevant data, determination of underlying distribution of the random variable, estimation of distribution parameters, and performance of statistical tests. Probabilistic modelling of soil parameters is especially necessary because randomness is ubiquitous in soil properties of geotechnical engineering projects.

Soils are among the most variable of all engineering materials, and thus, soil sample data are extremely amenable to comprehensive probability, statistics, and reliability treatment. Using this approach, the uncertainties are fully modeled by either distribution functions or quantile functions (QFs). Distribution functions comprise probability density functions (PDFs) and cumulative distribution functions (CDFs). The CDF is the integral of the PDF, and the QF is the inverse of the CDF. Uncertainties are also partially characterized by descriptive statistics, such as moments. The moments of a random variable can be obtained from its distribution functions. Several methods have been proposed to recover a PDF from the first four moments, such as the saddlepoint approximation, the Gram-Charlier and Edgeworth expansions, and the maximum entropy principle (MEP) [Kennedy and Lennox 2000; Singh 2015; Zhao and Lu 2021]. The MEP is attractive and widely used because it can generate a “minimally prejudiced probability distribution which maximizes the entropy subject to constraints supplied by the given information” [Rao and Hsieh 1987]. The following advantages of fractional moments over integral moments have been highlighted in recent research: (1) they can extract more information from samples with fat-tailed distributions; (2) they contain lower sampling variability; and (3) fewer fractional moments are needed to accurately characterize a distribution [Deng and Pandey 2023].

This paper proposes a novel probabilistic method to model soil properties using quantile functions, based on the principle of maximum entropy constrained by fractional probability-weighted moments. Maximum entropy quantile distributions are compared to the traditional quantile distributions to evaluate their performance.

2. SOIL SAMPLING AND TESTING

A large landslide on April 23, 1990, on the east bank of the Nipigon River about 8 km north of Nipigon Township and 8 km south of Alexander Generating Station in Northwestern Ontario, Canada, moved approximately 300,000 m^3 of soil and extended almost 350 m inshore, with a maximum width of approximately 290 m. Soil from the landslide was pushed into the Nipigon River 300 m upstream and downstream and formed several islands in the river. These islands redirected the river current and caused subsequent erosion on the west bank of the river opposite of the slide area. This likely caused several landslides to occur farther south one month later [Radhakrishna et al, 1992]. This landslide incurred significant negative environmental and economic impacts. Although the riverbank soil erosion and landslides of various scales have taken place along all stretch of the Nipigon River, but the 1990 landslide was the first one that received public and academic attention, and the investigations were conducted to locate the main factors which caused it and to understand the landslide’s failure mechanism.

Landslide investigation and soil sampling in the Nipigon River area were conducted in August 2022 with the help of Red Rock Indian Band, which is shown in Figure 1. Thirty samples were obtained and transported to the soil laboratory at Lakehead University, then direct shear tests were performed (Figure 2) to estimate the soil effective shear strength parameters (the effective cohesion, c and the effective angle of internal friction, ϕ) by following the procedure of the direct shear testing of soils suggested by ASTM D3080-90[Bowles 1992]. The soil was sandy to clayey silt on the surface layer of the Nipigon River valley. The results of direct shear tests provide fundamental data to calculate slope stability, soil bearing capacity, pavement designs, and lateral earth pressures on retaining structures.

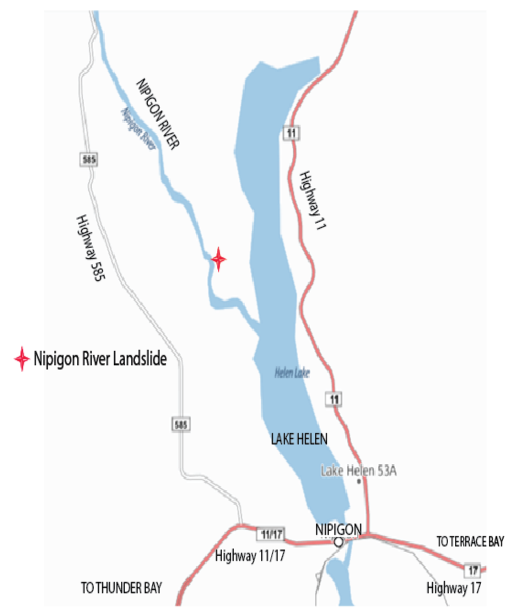

Figure 1: Nipigon River landslide and Soil sampling

Figure 2: Direct shear test equipment.

The direct shear box tests were conducted using the ELE Direct Shear Apparatus EL28-007 series and associated data acquisition system. The size of the chosen shear box is 60 mm × 60 mm × 40 mm. The temperature and humidity of the soil laboratory were 23°C and 40%, respectively throughout the tests. Before testing, we installed the shear box with the soil sample and assembled the dial gauges to obtain the readings of vertical loads and displacements. Consolidated-drained conditions were chosen during the tests: A vertical normal force is applied on the soil sample. The horizontal shear force is delayed until all settlement stops. The shear force is then applied so slowly that the small pore pressure that develops in the sample can be ignored. The strain rate to be applied in the test was fixed as 0.25 mm/min using the gears provided on the device, such that pore pressure buildup in the soil samples can be prevented. Readings of the shear force and displacement dials are recorded in the computer. Perform the test for three vertical loadings for each sample: 10, 25, and 50 kg which are 27.2, 68.1, and 136.2 kPa, respectively. All 30 samples were tested using the same procedure with the same strain rate and vertical load values. Only the effective cohesion of the soil is listed in Table 1. It is evident that the data are highly variable even if all possible measures were taken to ensure uniform conditions during the soil collecting and testing. In next section, maximum entropy quantile function will be developed to describe the randomness in the effective cohesion of soil.

Table 1: Effective cohesion of soil (kPa)

35.5	11.7	16.6	18.4	19.1	15.5	14.9	21.7
18.9	24.4	15.0	20.3	16.5	17.9	26.4	9.9
15.0	15.7	12.8	16.4	14.0	15.7	8.5	15.9
10.8	18.7	3.1	14.7	12.0	12.2		

3. METHODOLOGY

The maximum entropy principle constrained by the fractional probability-weighted moments from a sample of data is used to derive a series of unbiased quantile functions. Akaike information criterion is then used to locate the optimal order of the maximum entropy quantile function.

3.1 Fractional probability weighted moments

This section defines and estimates the fractional probability weighted moment (FPWM). The FPWM is used to numerically model uncertainty in a random variable, which is defined as

$$\beta_s = \int_0^1 x(F) F^s \mathrm{d}F, \tag{1}$$

where $F \equiv F(x) = P(X \leq x)$ is the probability of non-exceedance, $\mathrm{E}[\cdot]$ is the mathematical expectation, and s is a real number (fractional number). β_s is called the fractional probability weighted moment (FPWM) of the quantile function $x(F)$.

If a sample of data $x_i (i = 1,2,\cdots,n)$ is available, then the FPWM can be estimated by

$$b_s \approx \beta_s = \int_0^1 x(F) F^s \mathrm{d}F \approx \frac{1}{n}\sum_{i=1}^{n}[(F_i)^s x_i], \tag{2}$$

where n is the sample size, F_i is a proper formula for plotting position of the element x_i,

$$F_i = \frac{i-a}{n}, 0 < a < 1; \tag{3}$$

or

$$F_i = \frac{i-a}{n+1-2a}, -\frac{1}{2} < a < \frac{1}{2}. \tag{4}$$

For example, $a = 0.44$ is called the Gringorten plotting position.

3.2 Maximum entropy principle using FPWMs

This section derives the maximum entropy quantile function using the maximum entropy principle and sample FPWMs. The maximum entropy principle is often used to infer a probability distribution based on only partial information of a random variable, usually in terms of moments. The entropy of a random variable X is defined by

$$H[x(F)] = -\int_0^1 [x(F)\ln[x(F)]\,\mathrm{d}F, \qquad (5)$$

where $x(F)$ is the QF, $H[x(F)]$ is the entropy, and F is the probability of non-exceedance. The available infromation are given by FPWMs

$$\int_0^1 x(F)F^{\delta_s}\mathrm{d}F = \beta_s, \text{ for } s = 0,1,2,\cdots,K, \qquad (6)$$

where β_s is the sth order FPWM with the total of K FPWMs, δ_s is a real number and $\delta_0 = 0$. β_s can be determined from a sample of data in terms of b_s in Eq. (2), i.e., $\beta_s \approx b_s$.

The maximum entropy principle requires that the entropy $H[x(F)]$ be maximized under the constraints in Eq. (6). To do so, the Lagrangian function $\bar{H}$ is coined as

$$\bar{H} = -\int_0^1 [x(F)\ln[x(F)]\,\mathrm{d}F - (\lambda_0 - 1)\left[\int_0^1 x(F)\mathrm{d}F - b_0\right] - \sum_{s=1}^{K} \lambda_s \left[\int_0^1 x(F)F^{\delta_s}\mathrm{d}F - b_s\right], \qquad (7)$$

where $\bar{H}$ is the Lagrangian and λ_s is the Lagrangian multiplier. Substitution of Eq. (7) into the condition of maximization of $\bar{H}$

$$\frac{\partial \bar{H}}{\partial x(F)} = 0 \qquad (8)$$

leads to

$$x(F) \approx x_K(F|\lambda,\delta) = \exp\left[-\sum_{s=0}^{K} \lambda_s F^{\delta_s}\right], \qquad (9)$$

where $x_K(F)$ is referred to as the maximum entropy quantile function (MEQF) based on the K order sample FPWMs.

Substitution of Eq. (9) into Eq. (6) gives $K+1$ nonlinear equations, $s = 0,1,\cdots,K$.

$$\int_0^1 F^{\delta_s} \exp\left[-\sum_{i=0}^{K} \lambda_i F^{\delta_i}\right]\mathrm{d}F = b_s. \qquad (10)$$

If K and δ_i have already known, solving this set of equation gives the Lagrangian multipliers λ_s, which can be done by the command "*fsolve*" in Matlab (the Gauss-Newton method with numerical gradient and Jacobian).

3.3 Akaike information criterion

This section is to estimate K and δ_i of the maximum entropy quantile function in Eq. (9).

If $x(F)$ is the true QF, and $x_K(F|\lambda,\delta)$ or $x_K(F)$ is the estimated QF given in Eq. (9). The proximity between $x_K(F|\lambda,\delta)$ and $x(F)$ can be estimated by Kullback-Leibler (KL) entropy,

$$\mathrm{KL}[x(F), x_K(F|\lambda,\delta)] = \int_0^1 x(F)\ln\frac{x(F)}{x_K(F|\lambda,\delta)}\mathrm{d}F = C - L(\lambda,\delta,K), \qquad (11)$$

$$C = \int_0^1 x(F)\ln x(F)\,\mathrm{d}F, \qquad (12)$$

$$L(\lambda,\delta,K) = \int_0^1 x(F)\ln x_K(F|\lambda,\delta)\,\mathrm{d}F. \qquad (13)$$

The smaller the $\mathrm{KL}[x(F), x_K(F|\lambda,\delta)]$, the closer $x_K(F|\lambda,\delta)$ to $x(F)$. Therefore, $x_K(F|\lambda,\delta)$ should be chosen such that the KL entropy is minimized,

$$\min_K\left\{\min_{\lambda_0,\cdots,\lambda_K}\left\{\min_{\delta_0,\cdots,\delta_K}\{\mathrm{KL}[x(F), x_K(F|\lambda,\delta)]\}\right\}\right\}. \qquad (14)$$

The term C in Eq. (11) is independent of $x_K(F|\lambda,\delta)$, and the term $L(\lambda,\delta,K)$ can be taken as the expectation of the function $\ln x_K(F|\lambda,\delta)$. So $L(\lambda,\delta,K)$ can be estimated from a sample of data by $\hat{L}(\lambda,\delta,K)$

$$\hat{L}(\lambda,\delta,K) = \frac{1}{n}\sum_{i=1}^{n}[x_i \ln x_K(F_i|\lambda,K)], \qquad (15)$$

$$\widehat{\mathrm{KL}}(\lambda,\delta,K) = C - \hat{L}(\lambda,\delta,K), \qquad (16)$$

where $\widehat{\mathrm{KL}}(\lambda,\delta,K)$ is an estimate of the KL entropy. The minimization in Eq. (14) is given by

$$\min_{\lambda,\delta,K}\{\widehat{\mathrm{KL}}(\lambda,\delta,K)\} = C + \min_{\lambda,\delta,K}\{-\hat{L}(\lambda,\delta,K)\} = C + \min_{\lambda,\delta,K}\left\{-\frac{1}{n}\sum_{i=1}^{n}[x_i \ln x_K(F_i|\lambda,\delta,K)]\right\}. \qquad (17)$$

The minimization is equivalent to the maximum of the term $\frac{1}{n}\sum_{i=1}^{n}[x_i \ln x_K(F_i|\lambda,\delta,K)]$, which is a likelihood estimate. It was noted that the maximum likelihood estimate is often biased. One of the unbiased estimates of $-\hat{L}(\lambda,\delta,K)$ is given by Akaike information criterion

$$\hat{\Gamma}(\lambda, \delta, K) = -\hat{L}(\lambda, \delta, K) + \frac{K}{n}, \quad (18)$$

or

$$\hat{\Gamma}(\lambda, \delta, K) = \sum_{s=0}^{K} \lambda_s \left\{\frac{1}{n}\sum_{i=1}^{n}\left[x_i(F_i)^{\delta_i}\right]\right\} + \frac{K}{N} = \sum_{s=0}^{K}(\lambda_s b_s) + \frac{K}{N} = \widehat{H}[x(F)] + \frac{K}{N}. \quad (19)$$

This is done by substituting Eq. (9) into the term $\ln x_K(F_i|\lambda, \delta, K)$ of Eq. (17) and considering Eq. (2).

3.4 Procedure

The procedure to infer an optimal and unbiased MEQF from sample FPWMs is given as follows.

(1) Obtain a sample of data $x_i(i = 1,2,\cdots,n)$ for random variable X, and specify the highest order of FPWM κ. For example, $\kappa = 10$.

(2) For a specific K $(K = 1, 2,\cdots,\kappa)$, generate K distinct real random numbers as the fractional exponents $\delta_s(s = 1, 2,\cdots,K)$ and $\delta_0 = 0$ by performing Monte Carlo simulation.

(3) Calculate the FPWM b_s using Eq. (2).

(4) Determine the $K + 1$ Lagrangian multiplier, λ_s, in Eq. (9) using Eq. (10).

(5) Repeat steps (3)–(4) until the cycle of Monte Carlo simulation for δ_s terminates. Choose the smallest $\hat{\Gamma}(\lambda, \delta, K)$ of Eq. (19) in all Monte Carlo simulations for the given K. The corresponding λ and δ are the entropy coefficients and fractional exponents for this K.

(6) Repeat steps (2)–(5) until the highest order of fractional moments κ arrives. We obtain the relation of $\hat{\Gamma}(\lambda, \delta, K)$ in Eq. (35) as a function of K only.

(7) Locate the optimal order K that minimizes the $\hat{\Gamma}(\lambda, \delta, K)$. This order is the optimal order of the MEQF based on FPWMs.

4. MAXIMUM ENTROPY QUANTILE FUNCTIONS OF SOIL SHEAR STRENGTH

4.1 Quantile functions of soil shear strength

This section is to derive the maximum entropy quantile function of the soil from the Nipigon River area by following the procedure in Section 3.4. A sample of the soil data was obtained in Table 1. A MATLAB code has been developed by specifying $\kappa = 10$ and using the Gringorten plotting position formula. The cycle of Monte Carlo simulation is 5000. The function "rand" in MATLAB is used to generate the fractional exponents $\delta_s(s = 1, 2,\cdots,K)$. We obtain 10 values of $\hat{\Gamma}(\lambda, \delta, K)$: -46.1867; -46.1594; -46.2462; -46.2147; -46.2114; -46.1794; -46.1681; -46.1366; -46.1054; -46.0733. Therefore, the optimal order is $K = 3$ since -46.2462 is the smallest value. The parameters of the optimal maximum entropy quantile function are listed in the upper half of Table 2. The maximum entropy quantile function is

$$x(F) = \exp[-3.45387055 + 5.38446475F^{1.21150197} - 7.70107589F^{2.02596742} + 4.18244208F^{3.99411995}]. \quad (20)$$

*Table **2.** Parameters of quantile functions (FPWM order K=3)*

	δ_s	λ_s
s=0	0	-3.45387055
s=1	1.21150197	5.38446475
s=2	2.02596742	-7.70107589
s=3	3.99411995	4.18244208
Normal distribution	μ =16.2745	σ =5.8704
Lognormal distribution	μ=2.7184	ζ=0.4174

To further verify the accuracy of the proposed method in Section 3, comparison is made to normal distribution and lognormal distribution, two commonly used probability distributions in geotechnical engineering. Their parameters are determined by the method of moments and are given in the lower half of Table 2. The quantile functions can be obtained from their PDFs as follows.

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \frac{1}{5.8704\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{x-16.2745}{5.8704}\right)^2\right], \quad (21)$$

$$f_2(x) = \frac{1}{\zeta\, x\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{ln\, x-\mu}{\zeta}\right)^2\right] = \frac{1}{0.4174\, x\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{ln\, x-2.7184}{0.4174}\right)^2\right], \quad (22)$$

where $f_1(x)$ is the normal PDF and $f_2(x)$ is the lognormal PDF for the cohesive shear strength X.

4.2 Comparison and accuracy

Comparison between the maximum entropy quantile functions with various orders ($K = 1 - 5$) and the normal and lognormal distributions is given in Figure 3. The sample data were drawn by using the Gringorten plotting position formula. A semi-log plot of probability of exceedance in Figure 4 illustrates that the lognormal quantile function overestimates the sample tail region, but the normal quantile function underestimates the sample tail region, especially in the domain between 10^{-1} and 10^{-2}. The optimal maximum entropy quantile function with order 3 behaves between the normal QF and the lognormal QF, so a better fit is found to the soil cohesion strength.

Furthermore, two more indexes, relative root mean square error (RRMSE) and relative absolute error (RAE), are introduced to quantitatively compare accuracy of these quantile functions:

$$\text{RRMSE}=\left[\frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{Q_{0i}-Q_{Ci}}{Q_{0i}}\right)^2\right]^{1/2}, \tag{23}$$

$$\text{RAE}=\frac{1}{n}\sum_{k=1}^{n}\left|\frac{Q_{0i}-Q_{Ci}}{Q_{0i}}\right|, \tag{24}$$

where n is the sample size; Q_{0i} and Q_{Ci} are the ith sample element and the ith computed value of the i th given probability, respectively. The smaller RAE and RRMSE, the better fit of the distribution. Table 3 shows that the maximum entropy quantile function with order 3 in Eq. (20) has smaller of both RAE and RRMSE than the lognormal distribution, and has smaller RAE than the normal distribution. However, the negative domain of the normal distribution makes it an adverse quantile function for the soil cohesion, which is always positive values.

Another interesting point is the higher order of MEQF, the smaller of both RAE and RRMSE. However, the higher order MEQF has a potential to overfit the sample of data. The Akaike information criterion can effectively prevent overfitting and underfitting, thus lead to an optimal maximum entropy quantile function.

5. CONCLUSIONS

This paper proposes a novel probabilistic method to model the soil shear strength properties using the quantile functions, which are derived by the maximum entropy principle, Akaike information criterion, and fractional probability-weighted moments. The maximum entropy principle is constrained by the fractional probability-weighted moments from a sample of data and is used to derive a series of unbiased quantile functions. Akaike information criterion is then used to locate the optimal order of the maximum entropy quantile function. The use of fractional probability-weighted moments facilitates more accurate quantification of soil uncertainties by the entropy-based quantile functions than the probability density or cumulative distribution functions. The maximum entropy quantile distributions are compared to the traditional quantile distributions to evaluate their performance. The analytical entropy quantile distribution obtained can be used in the probabilistic reliability analysis. The maximum entropy quantile distributions are only applicable to random samples with positive values.

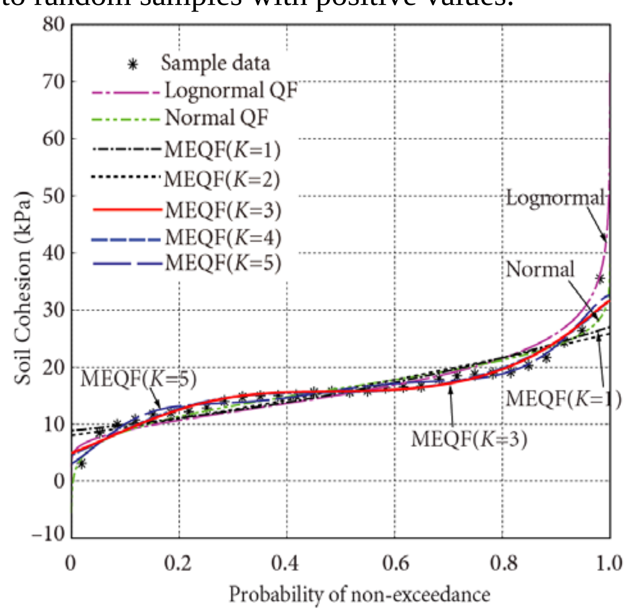

Figure 3: Quantile functions.

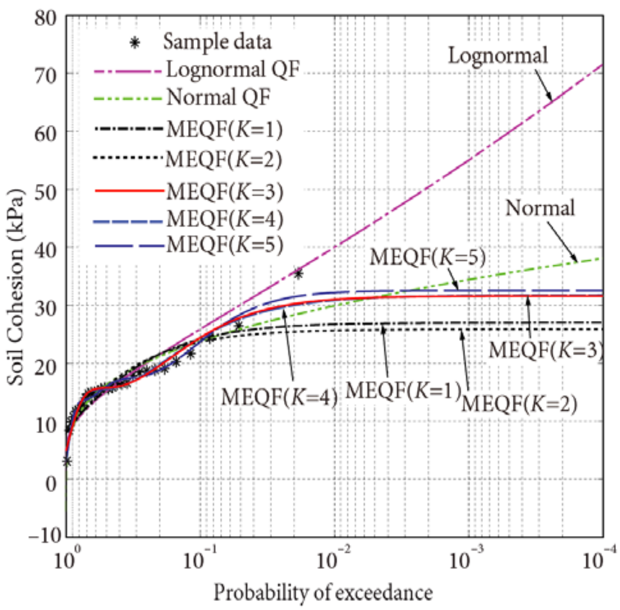

Figure 4: Semi-log quantile fonctions.

Table 3. RRMSE and RAE of quantile functions.

QF	RRMSE	RAE
Normal	0.109	0.087
Lognormal	0.219	0.124
MEQF($K = 1$)	0.365	0.144
MEQF($K = 2$)	0.322	0.139
MEQF($K = 3$)	0.162	0.076
MEQF($K = 4$)	0.149	0.073
MEQF($K = 5$)	0.091	0.058

6. ACKNOWLEDGEMENTS

We acknowledge the support in part by funding from the Government of Canada's New Frontiers in Research Fund (NFRF) [NFRFR-2021-00262] and in part by the Natural Sciences and Engineering Research Council of Canada through discovery grant. Assistance from the Red Rock Indian Band in Northwestern Ontario, Canada (http://rrib.ca/) was greatly acknowledged by Dr Jian Deng's research team.

7. REFERENCES

Bowles, J.E. (1992). *Engineering Properties of Soils and Their Measurement* (4th ed). McGraw Hill.

Deng, J., Pandey, M.D. (2023) Optimal maximum entropy quantile function for fractional probability weighted moments and its applications in reliability analysis. *Applied Mathematical Modelling*, 114, 230–251.

Kennedy, C.A. and Lennox, W.C. (2000). Solution to the practical problem of moments using nonclassical orthogonal polynomials, with applications for probabilistic analysis. *Probabilistic Engineering Mechanics*, 15(4), 371–379.

Radhakrishna, H.S., Bechai, M., Lau, K.C., Law, K.T., and Hale, I. (1992). *Nipigon River Landslide*. GeoHazards 92. First Canadian Symposium on Geotechnique and Natural Hazards, Vancouver, British Columbia. Vol 1 - Geotechnique and Natural Hazards, 127-137.

Rao, R. and Hsieh, C.H. (1987) Maximum entropy probability distributions for flood frequency analysis. *Civil Engineering Systems*, 4, 67–76.

Singh, V.P. (2015). *Entropy Theory in Hydrologic Science and Engineering*. McGraw-Hill Education.

Zhang, X.F. and Pandey M.D. (2013). Structural reliability analysis based on the concepts of entropy, fractional moment and dimensional reduction method. *Structural Safety*, 43, 28–40

Zhao, Y.G. and Lu, Z.H. (2021). *Structural Reliability: Approaches from Perspectives of Statistical Moments*. Wiley.