Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140507

Comparative Analysis of the 100-Year Return Level of the Average Monthly Rainfall for South Africa: Parent Distribution versus Extreme Value Distributions

Daniel Mashishi, Daniel Maposa and 'Maseka Lesaoana*

Department of Statistics and Operations Research, University of Limpopo, Private Bag X1106, Sovenga, 0727, South Africa

Received: 12 Apr. 2020, Revised: 10 Jun. 2020, Accepted: 17 Jul. 2020
Published online: 1 Sep. 2020

Abstract: In this paper, we model average monthly rainfall for South Africa using the parent distribution and extreme value theory (EVT). The 100-year return level plays an important role to hydrologists, meteorologists and civil engineers. Hence, the paper focuses on modelling the 100-year return level of average monthly rainfall for South Africa using the parent distribution and EVT. The present paper aims to compare the extreme quantile estimates of the EVT and parent distributions as well as to reveal the risk brought by heavy rainfall in South Africa. The method of maximum likelihood was used to estimate unknown parameters. We first investigate the parent distribution of the average monthly rainfall for South Africa. The results showed that the two-parameter Weibull distribution, which is in the domain of attraction of the Weibull family, is the appropriate parent distribution to model the data. We then perform a comparative analysis of the 100-year return level using the two-parameter Weibull distribution, the generalised extreme value distribution (GEVD), and the Poisson point process. The findings revealed that the 100-year return level of the two-parameter Weibull distribution was lower compared to that of the GEVD and Poisson point process model. The 100-year return level of the GEVD was equal to that of the observed maximum for the series, whereas that of the Poisson point process was slightly higher than the observed maximum average monthly rainfall for South Africa. Moreover, EVT models gave higher quantile estimation of the 100-year average monthly rainfall for South Africa compared to the parent distribution. Furthermore, EVT based estimation gave narrower confidence intervals as compared to the wider confidence interval of the parent distribution. Therefore, EVT models can play an important role in disaster risk reduction and civil engineering constructions, such as bridges and dams.

Keywords: 100-year return level, Extreme value theory, Generalised extreme value, Poisson point process, Two-parameter Weibull distribution.

1 Introduction

Extreme value theory (EVT) is a statistical discipline that develops techniques and models to make inferences about rare events [1]. It can also be used for risk assessment on financial markets, telecommunications for traffic prediction, portfolio adjustment in insurance companies and prediction of the occurrence of environmental phenomena, such as rainfall and storms.

Several studies on the applications of EVT in various disciplines have been conducted. Bali [2] applied generalised extreme value approach to model financial risk measurement. The results revealed that the loss of financial institutions can be accurately estimated using EVT. In a separate study on patients in intensive care unit (ICU) in the United Kingdom, [3] adopted EVT to build a probabilistic detector that identifies patients who are in a deterioration state. The study indicated that about 20,000 unforeseen patients admitted to ICU could be avoided if they had this detector. Diriba et al. [4] used the generalised extreme value distribution (GEVD) and Gumbel distribution to model rainfall data in East London, South Africa. Their findings showed a decreasing trend of extreme rainfall events throughout the past six decades.

Jaruskova and Hanek [5] compared the peaks-over threshold and block maxima approaches in extreme value theory. Their findings revealed that if a series is very long, the shape parameter of the GEVD and generalised Pareto

* Corresponding author e-mail: danmaposa@gmail.com

distribution (GPD) are likely to be very close to each other. It was revealed that GEVD is sensitive to outliers and it often produces high quantile estimates. A separate study by [6] presented an evidence on how heavy rainfall can flash away the seeds during the early stages of ploughing. Dyson and Van Heerden [7] presented an evidence on how heavy rainfall affected the people of Zimbabwe, Mozambique, and South Africa. The report showed that Limpopo province in South Africa suffered a severe loss of R1.3 billion in infrastructure and roads and roughly 200 bridges were also destroyed. In Zimbabwe and Mozambique, approximately 600 people lost their lives and others were forced to leave their own homes. The previous pieces of literature demonstrates that the extreme events are still happening and it is important to model these rare events.

The paper aims to model the average monthly maximum rainfall of South Africa using a parent distribution and EVT with a particular interest in the comparative analysis of the 100-year return level of the parent and EVT distributions. To the best of our knowledge, this type of work has not been done in South Africa and literature is scarce in the other parts of the world. This paper benefits the government officials who undertake the responsibility of disaster management, risk management, food security, and private sectors. The findings also help climatologists, hydrologists, meteorologists, and decision makers manage floods in South Africa. Knowledge of the quantile estimates of the extreme rainfall events will reduce the amount of money spent by the government and insurance companies on disaster relief operations, property recovery and loss of lives since the government will be well prepared for these natural disasters.

The rest of the paper is organised as follows: Section 2 presents the research methodology which comprises, the data source, study area, the family of distributions and extreme value techniques applied in the analysis of the data. Section 3 presents the results of the study in the form of tables and figures, as well as a detailed discussion of the results. Section 4 is devoted to conclusion and recommendations. While acknowledgements and references are presented at the end of the paper.

2 Research Methodology

This paper focuses on modelling average monthly rainfall for South Africa using three approaches: The parent distributions, GEVD and Poisson point process distribution. Four candidate parent distributions are investigated to identify a suitable parent distribution for average monthly rainfall in South Africa. The selected parent distribution is then used to estimate the 100-year return level and compared to the 100-year return level quantile estimates of the EVT models.

2.1 Data source and study area

The data used in this study is secondary data and is average monthly rainfall for South Africa, measured in millimeters (mm), obtained from South African Weather Service (SAWS) for the period 1940 - 2017. The time series data covers the whole of South Africa.

2.2 Augmented Dickey-Fuller Test

The augmented Dickey-Fuller (ADF) test is widely used to test whether a certain data set is stationary or not. According to [8], the ADF test is derived from the expression below:

$$y_t = d_t + \beta_1 y_{t-1} + \sum_{i-1}^{\rho-1} \gamma_i \Delta y_{t-1} + \varepsilon_t,$$

where $d_t = \sum_{i=1}^{\rho} \phi_i t^i$, for $\rho = 0, 1$. The null hypothesis to be tested is given by $H_0 : \beta_1 = 1$ (The data is not stationary) and the alternative hypothesis is: $H_1 :\mid \beta_1 \mid < 1$ (The data is stationary).

2.3 Candidate distributions

The framework of the cumulative distribution functions (CDFs) for the candidate distributions is presented in this subsection. The candidate distributions consist of four parent distributions, presented as follows:

Weibull distribution

We have two types of Weibull distributions: The two-parameter and the three-parameter. The cumulative distribution function (CDF) of the two-parameter Weibull distribution is given by:

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right),$$

while the CDF of the three-parameter Weibull distribution is given by:

$$F(x) = 1 - \exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right),$$

where γ is a location parameter, β is the scale parameter and α is a shape parameter. This distribution is commonly used in hydrology and reliability studies [9], [10], [11].

Gamma distribution

The CDF of the gamma distribution is defined as:

$$F(x) = \frac{\Gamma_{(x-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)},$$

where $\gamma, \beta > 0$ and α are the location, scale and shape parameters, respectively. Γ is called the gamma function [9], [10], [11],.

Pareto distribution

There are two types of Pareto distributions: The two-parameter and three-parameter distribution [12]. Suppose X is a random variable that follows a two-parameter Pareto distribution, then the CDF of X is:

$$F(x) = \left(\frac{x}{\sigma}\right)^{-\xi}, \ x > \sigma,$$

where σ and ξ are the scale and shape parameters, respectively. The CDF of the three-parameter Pareto distribution is given by:

$$F(x) = \left[1 + \left(\frac{x-\mu}{\sigma}\right)\right]^{-\xi}, \ x > \mu,$$

where σ, μ and ξ are the scale, location and shape parameters, respectively.

Log-normal distribution

There are two log-normal distributions, the two-parameter, whose CDF is given by:

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\mu}\right),$$

and the three-parameter log-normal given by:

$$F(x) = \Phi\left(\frac{\ln(x-\gamma) - \mu}{\sigma}\right),$$

where $\gamma, \sigma > 0$ and μ are the continuous location, scale and shape parameters, respectively. Φ is called the Laplace integral [9] [10], [11].

2.4 Extreme value models

This subsection presents two approaches of EVT: The block maxima and peaks-over-threshold (POT). We start by presenting the framework of the GEVD, which represents the block maxima, followed by the model development of Poisson point process for extremes, which represents the POT realisation.

***Generalised extreme value distribution** (GEVD)*

According to [1], the model development of GEVD is based on the statistical behaviour of $M_n = \max(X_1, \cdots, X_n)$, where X_i for $i = 1, \cdots, n$ represents the sequence of independent and identically distributed (iid) random variables that have the same underlying distribution function F. In EVT, the GEVD is said to be the limiting distribution of the normalised maxima of X_i, for $i = 1, \cdots, n$. The unified GEVD for modelling maxima is given by:

$$G(z) = \exp\left(-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right), \qquad (1)$$

which is defined on $\{z : 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0\}$. Equation 1 has three parameters: The location (μ) defined on $-\infty < \mu < \infty$, scale ($\sigma > 0$) and shape (ξ) defined on $-\infty < \xi < \infty$.

The GEVD unites three families of distributions depending on the value of the shape parameter: Frechet type when $\xi > 0$, Weibull when $\xi < 0$ and Gumbel type when $\xi = 0$. In order to estimate the return levels and their corresponding return periods, we first obtain the quantile function using the results in (1). Now, let $G(z_p) = 1 - p$, then (1) becomes:

$$z_p = \mu - \frac{\sigma}{\xi}\left[1 - \left(-\log(1-p)\right)^{-\xi}\right], \ \xi \neq 0. \qquad (2)$$

Note that for very small values of p and also for $\xi < 0$, (2) becomes:

$$z_p = \mu - \frac{\sigma}{\xi}. \qquad (3)$$

The formula in (2) is used to estimate the return levels of historical data.

The Poisson point process for extremes

In EVT, the Poisson point process framework is said to be similar to that of POT approach [1].

Suppose that $X_1, X_2, \cdots$ are independent and identically distributed (iid) random variables with the same distribution function F. Again, suppose that the X_i's are observed values within the data set such that $M_n = \max\{X_1, \cdots, X_n\}$.

Now, if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z),$$

where

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

with z_- and z_+ being the lower and upper endpoints of G, respectively. Then, we have the sequence of point processes

$$N_n = ((i/(n+1)), (X_i - b_n)/a_n : i = 1, \cdots, n)$$

that will converge on this region $(0,1)\times[u,\infty)$, for any $u > z_-$, to a Poisson process with intensity measure on $A = [t_1,t_2]\times[z,z_+]$ which is given according to [1] by:

$$\Lambda(A) = (t_2 - t_1)\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \quad (4)$$

3 Results and Discussion

This section presents the results based on the three quantile estimation approaches of the average monthly rainfall for South Africa. The descriptive statistics of the data are presented, followed by an investigation of the goodness-of-fit of four candidate distributions to select the suitable parent distribution and then estimate the return levels of GEVD, Poisson point process and selected parent distribution.

3.1 Descriptive statistics

Table 1 presents the summary statisitcs of average monthly rainfall for South Africa and stationarity test.

The results in Table 1 reveal that the average monthly rainfall readings for South Africa range from 3.20 mm to 175.00 mm, with a median of 44.80 mm. The data is positively skewed (mean $>$ median) and the kurtosis value which is greater than 3 suggests that the data may follow a heavy-tailed distribution. The level of significance used in this study is 5%. Since p-value is 0.01 in Table 1, the null hypothesis stating that the time series data is not stationary was rejected. According to the ADF test in Table 1, the time series data is stationary.

3.2 Investigating goodness-of-fit of parent distributions

In this subsection, we investigate the goodness-of-fit of four candidate parent distributions: The Weibull, gamma, lognormal and Pareto distributions. Table 2 presents results for the goodness-of-fit based on AIC and BIC.

The results in Table 2, based on the smallest values of both the AIC and BIC, suggest that Weibull distribution provides the best fit to the data and this distribution is the Weibull domain of attraction in the EVT context. We proceed to perform a visual comparison of the Weibull and gamma distributions diagnostic plots since their corresponding AIC and BIC values in Table 2 are very close.

Figure 1 presents the diagnostic plots for the two-parameter Weibull and gamma family of distributions. The diagnostic plots for the Q-Q plot, in Figure 1, reveal that there exists lack of fit at the tails of both the two-parameter Weibull and gamma distributions. However, the P-P plot suggests a reasonably good fit for both the two-parameter Weibull and gamma distributions. Based on the diagnostic plots and the findings in Table 2, the Weibull domain of attraction which includes the two-parameter Weibull distribution is the best parent distribution to model the data.

3.3 Quantile estimation for Weibull, GEVD and Poisson point process models

Table 3 presents the parameter estimates and their corresponding 95% confidence intervals for the two-parameter Weibull, GEVD and Poisson point process models. The shape of the GEVD is negative according to Table 3, which suggests that the underlying distribution of the data belongs to the Weibull domain of attraction. In addition, the non-inclusion of zero in the confidence interval implies that the underlying distribution is strictly in the Weibull domain of attraction. These results are supported by the findings in the previous subsection which indicated that the distribution of the average monthly rainfall for South Africa can be modelled by the Weibull distribution.

According to the findings from Table 2 and Figure 1, the selected parent distribution was Weibull. We therefore proceed to compute the 100-year return level of the Weibull and the two EVT distributions. According to [13], the quantile function of the Weibull distribution is given by:

$$X_p = \left[\beta \ln\left(\frac{1}{1-p}\right)\right]^{\frac{1}{\alpha}}.$$

Table 4 presents the results of the 100-year return levels and their corresponding confidence intervals for the two-parameter Weibull, GEVD and Poisson point process models. Thus, the average monthly rainfall estimates that are expected to be exceeded, at least once every 100 years (99th percentile) are 161.13 mm (two-parameter Weibull),

Table 1: Summary statistics of the rainfall data.

min	mean	median	Q1	Q3	max	kurtosis	skewness	ADF Test
3.20	51.15	44.80	19.57	75.03	175.00	3.05	0.80	-9.41 (p = 0.01)

Table 2: Selection of the most appropriate parent distribution.

Name	Pareto	Weibull	gamma	lognormal
Akaike's Information Criterion (AIC)	9241.879	**9059.184**	9061.132	9122.403
Bayesian Information Criterion (BIC)	9251.563	**9068.867**	9070.815	9132.086

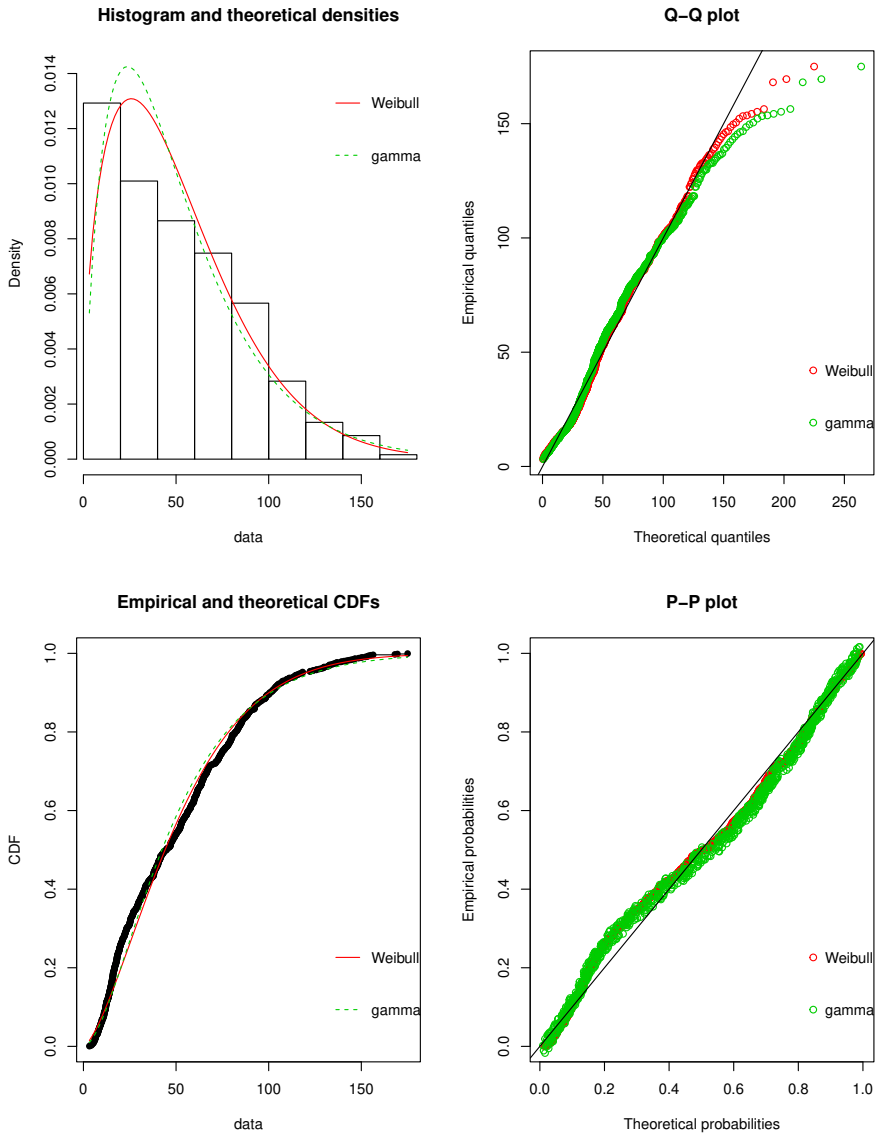


Fig. 1: Diagnostic plots for two-parameter Weibull and gamma distributions (Key: The red line represents the gamma distribution and the green line represents the Weibull distribution).

Table 3: Parameter estimates and the 95% confidence intervals based on the Weibull, GEVD and Poisson point process.

Model	μ	95% CI of μ	σ	95% CI of σ	ξ	95% CI of ξ
Weibull	1.46 (0.04)	(1.38, 1.54)	56.61 (1.33)	(54.00, 59.22)	-	-
GEVD	107 (3.15)	(101.24, 113.58)	24.53 (2.30)	(20.02, 29.04)	-0.25 (0.09)	(-0.43, -0.07)
Poisson	62.48 (3.10)	(56.40, 68.56)	7.68 (0.08)	(6.11, 9.25)	-0.37 (0.03)	(-0.43, -0.31)

Table 4: Quantile estimates and their corrresponding confidence intervals based on the two-parameter Weibull, GEVD and Poisson point process.

Model	100-year return level	95% CI
Weibull	161.13	(133.56, 188.44)
GEVD	174.99	(158.47, 191.32)
Poisson point process	178.44	(170.06, 189.09)

174.99 mm (GEVD) and 178.44 mm (Poisson point process). This implies that the 100-year return level of the point process model is greater than the maximum observed average monthly rainfall of 175 mm, which in turn is equal to the 100-year return level of the GEVD. However the 100-year return level of the two-parameter Weibull distribution is the lowest. The results in Table 4 also reveal that the 95% confidence interval for the two-parameter Weibull is wider than that of the GEVD which in turn is wider than that of the Poisson point process. In other words, the Poisson point process has the narrowest confidence interval.

4 Conclusion and Recommendations

This paper revealed that the average monthly rainfall data for South Africa can be well modelled by a distribution in the Weibull domain of attraction based on the investigation of the candidate distributions and the GEVD. The Poisson point process model also revealed how long South Africans must wait to receive rainfall of greater magnitute than that of February 2000. Moreover, the 100-year return level of the two-parameter Weibull distribution was lower compared to that of the GEVD and Poisson point process models.

The study suggests some future research directions that may help improve the accuracy and reliability of the findings. Since the study revealed some evidences that South Africa might experience higher than expected rainfall in the coming years based on 100-year quantile estimates of the GEVD and Poisson point process, it is recommended that the meteorologists, hydrologists, and geoscientists identify the locations of the likely impact or vulnerable areas. Moreover, the findings might be improved in the future if we utilise multivariate spatio-temporal extreme value modelling and Bayesian approach in extreme value parameter and quantile estimation.

Acknowledgement

Our special thanks go to the South African Weather Service (SAWS), National Research Foundation (NRF) and University of Limpopo for their tremendous support in this study.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article

References

[1] S. Coles. *An introduction to statistical modelling of extreme values*, 1st ed., Springer-Verlag, London, (2004).

[2] T. Bali. A generalised extreme value approach to financial risk measurement, *Journal of Money, Credit and Banking*, **39**, 1613 – 1649 (2007).

[3] S. Hugueny, D. Clifton, and L. Tarassenko. *Probabilistic patient monitoring using extreme value theory: A multivariate, multimodal methodology for detection of patients in deterioration state*, NIHR Biomedical Research Centre, Oxford, 1 – 20 (2010).

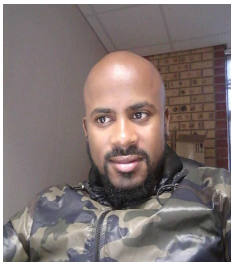
[4] T. Diriba, L.K. Debusho, J. Botai, and A. Hassen. Analysis of extreme rainfall at East London, *South African Statistical Journal*, 25 – 32 (2014).

[5] D. Jaruskova, and M. Hanek. Peaks over threshold method in comparison with block maxima method for estimating high return levels of several Northern Moravia precipitation and discharges series, *Journal of Hydrology and Hydromechanics*, **54**, 309 – 319 (2006).

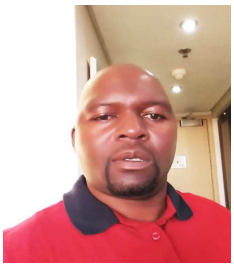
[6] C.J.C. Reason, S. Hachigonta, and R.F. Phaladi. Interannual variability in rainy season characteristics over the Limpopo region of Southern Africa, *International Journal of Climatology*, **25**, 1835 – 1853 (2005).

[7] L. Dyson, and J. Van Heerden. The heavy rainfall and floods over the northeastern interior of South Africa during February 2000, *South African Statistical Journal*, **97**, 80 – 86 (2001).

[8] M. Arltova, and D. Fedorova. Selection of unit root test on the time series and value of AR(1) parameter, *STATISTIKA*, **96**, 47 – 64 (2016).
[9] A.M. Alam, K.E.C. Farnham, and J. Yuan. Best-fit probability distributions and return periods for maximum monthly rainfall in Bangladesh, *Journal of Climate Science*, 1 – 16 (2018).
[10] J. Beirlant, Y. Goegebeur, and J. Teugels. *Statistics of extremes: Theory and applications*, John Wiley and Sons, West Sussex, (2004).
[11] D. Maposa. *Statistics of extremes with applications to extreme flood heights in the lower Limpopo River basin of Mozambique*, Ph.D. Thesis, University of Limpopo, South Africa, (2016).
[12] H. He, N. Zhou, and R. Zhang. On estimation for the Pareto distribution, *Statistical Methodology*, **21**, 49 – 58 (2014).
[13] M.A. Shayib, and A.M. Haghighi. Comparing two Quantiles: the Burr Type X and Weibull cases, *IOSR Journal of Mathematics*, **12**, 28 – 40 (2016).



Daniel Mashishi is a Junior Lecturer in the Department of Statistics and Operations Research, School of Mathematical & Computer Sciences, Faculty of Science & Agriculture, University of Limpopo. He holds a Master of Science degree in Extreme Value Statistics, a structured Master of Science degree in Mathematical Sciences, and an Honours degree in Statistics. His research interests are in the areas of Extreme value Statistics and Data Science. He is also a member of the South African Statistical Association (SASA).



Daniel Maposa is a Senior Lecturer in the Department of Statistics and Operations Research, School of Mathematical & Computer Sciences, Faculty of Science & Agriculture, University of Limpopo. He holds a PhD degree in Extreme Value Statistics, a Master of Science degree in Operations Research and Statistics, and an Honours degree in Applied Mathematics. He has published more than 21 journal research articles in internationally accredited journals, two book chapters and three conference proceedings. In his research activities, he has been all over the world attending international conferences and presenting his research work in statistics of extremes in countries such as New Zealand, Australia, China, Switzerland, Brazil, Morocco, Botswana and Malaysia. Daniel Maposa is a registered professional natural scientist (Pr.Sci.Nat.) in Statistical Sciences & Mathematical Sciences and is a member of International Statistical Institute (ISI) and a member of South African Statistical Association (SASA). He also regularly attends the South African Statistical Association (SASA) conference annually and the ISI World Statistics Congress organised bi-annually.

'Maseka Lesaoana obtained her Doctorate in Operations Research from the University of Southampton, UK (1991) and a Master of Mathematics (MMath) degree in statistics and operations research from the University of Waterloo, Canada (1985). She started her career in academia as a teaching assistant in the Department of Statistics at the National University of Lesotho in 1981. She joined the Human Sciences Research Council in 1994 to establish the South African Data Archive; and the National Department of Labour in 1999 to establish the Labour Market Information and Statistics Unit. While in Pretoria (1994-2002), she was also a part-time lecturer in statistics at the then Technikon South Africa (which has been incorporated into the University of South Africa). She has been with the University of Limpopo since February 2003: as Senior Lecturer in the Department of Statistics and Operations Research (SAOR); Head of Department (SAOR); Director of the School of Mathematical and Computer Sciences; and later as a full-time Research Professor in SAOR from mid-2014 until her retirement at the end of October 2019 and beyond. Her research interests are in applied statistics and operations research. She has successfully supervised and mentored several postgraduate students in both statistics and operations research and has published widely with them and her global collaborators. Since 2017, she has been a member of the National Research Foundation's Strategy and Steering Committee of the newly established National Graduate Academy for the Mathematical and Statistical Sciences in South Africa. She was a member of the Statistics Council of Statistics South Africa (2000-2018). In 2018, she resumed membership of the Umalusi Assessment and Standardisation Committee (that she had left in 2009). She is a member of both the Operations Research Society of South Africa (ORSSA) and the South African Statistical Association (SASA). She is the 2020 President of SASA and a member of the BRICS Planning Committee for the Mathematics and Statistics Conference that will be hosted by South Africa in 2021/2022.