Comparison of sum of two correlated gamma variables for Alouini's model and McKay distribution

R. Zakaria [a], **J.W. Boland** [b], **N.H. Moslim** [a]

[a]*Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak 26300 Gambang, Kuantan Pahang, Malaysia*
[b]*School of Information Technology and Mathematical Sciences, University of South Australia, Australia*
Email: roslinazairimah@ump.edu.my

Abstract: A statistical model for rainfall is useful to describe the relationship between rainfall at a given location and other weather-related variables. It is also able to provide a principled way to quantify the uncertainty that associates rainfall processes, which is crucial to the efficient design of environmental projects and to improve crop production. Various statistical models have been used for rainfall such as a right-skewed distribution including the exponential, gamma or mixed-exponential to model the rainfall intensities and Markov chain model to model rainfall occurrences.

Two models of the sum of two correlated gamma variables, namely Alouini's model and McKay distribution are studied. Alouini's model is an extension of Moschopoulos results for the sum of n correlated gamma variables. There are two forms of McKay distribution, the Type I is defined for sum of two correlated gamma variables whereas Type II is defined for difference of two correlated gamma variables. In this study, the Alouini's model and Type I of McKay distribution are compared using monthly rainfall totals to form water catchment from two meteorological stations, Hume and Beechworth within the Murray-Darling Basin. Rainfall totals during the summer season at both stations are selected based on the significant correlations. The procedure for the analysis is as follows: (1) select positive pairs set of data from two stations (2) calculate the pairwise Spearman's correlation and check if the correlation is significant (3) fit the data marginally with the gamma distribution and use maximum likelihood estimation method to estimate the parameters for the gamma distribution (4) calculate the average value of shape parameters and re-estimate the values of scale parameters (5) apply the model to generate synthetic sum of monthly rainfall totals (6) compare the sum of monthly rainfall totals between the observed and generated data using the Kolmogorov-Smirnov goodness of fit test.

The results of the analysis show that the values of mean and variance of the observed and estimated from the McKay distribution is closer than the values of mean and variance estimated from the Alouini's model. Based on the Kolmogorov–Smirnov goodness of fit test, the P-value of observed versus McKay distribution is much higher than the P-value of observed versus Alouini's model. It shows that the McKay distribution fits the data better even though both models pass the test. However, when the Alouini's model and McKay distribution are compared, the rainfall totals generated by the two models fail the test of being from the same population. Zakaria (2011) shows that the Alouini's model is suitable in modelling the sum of four correlated gamma variables and can easily be extended to more than two variables. On the other hand, the McKay model is not easily extendable to more than two variables because of extensive algebraic manipulation. Thus, in this particular example of two stations and three months, it is instructive to note that the McKay formulation can well represent the sum of individual months at the two locations, but if we wanted to represent the sum over the three months of the season, we would have to use Alouini's method. Both models are able to be used to generate synthetic sum of rainfall totals and Alouini's model can be used to model the sum of more than two correlated gamma variables. It may well be that catering for this extra flexibility is the reason that the Alouini's model does not perform so well for the two variable case. In future work, we will be comparing these two approaches with other formulations as well, such as the use of Maximum Entropy methods.

Keywords: McKay distribution; Alouini's model; correlated gamma variables; rainfall model

1 INTRODUCTION

A statistical model for rainfall is useful to describe the relationship between rainfall at a given location and other weather-related variables. It is also able to provide a principled way to quantify the uncertainty that associates rainfall processes, which is crucial to the efficient design of environmental projects and to improve crop production. Various statistical models have been used for rainfall such as a right-skewed distribution including the exponential, gamma or mixed-exponential to model the rainfall intensities and Markov chain model to model rainfall occurrences.

The gamma distribution has been used widely in diverse applications such as rainfall modelling (Wilks and Wilby (1999); Srikanthan and McMahon (2001); Rosenberg et al. (2004); Husak et al. (2007); Piantadosi et al. (2009)), wireless communication systems (Alouini et al. (2001); Holm and Alouini (2004); Paris (2011)), insurance losses (Furman (2008)) and wind farm output (Boland (2005)). Husak et al. (2007) review studies of the gamma distribution characteristics for fitting rainfall data in Africa. They highlighted why the gamma distribution is always an initial choice to model rainfall data: it fits positive data, is a good representation of rainfall distribution and is sufficiently flexible to fit various rainfall regimes from exponential to normal shape. The flexibility of the gamma distribution is due to two parameters; shape and scale, see Wilks (1990).

A common goal in model fitting is to generate synthetic data from the fitted model. The gamma distribution is also used to generate synthetic rainfall data (Srikanthan and McMahon (2001); Wilks and Wilby (1999); Rosenberg et al. (2004); Piantadosi et al. (2009)). Synthetic rainfall data generation is important as a supplement to the historical rainfall data which is always lacking in terms of temporal and spatial extent. A comprehensive review of weather generation models is given by Srikanthan and McMahon (2001) and Wilks and Wilby (1999). Piantadosi et al. (2009) applied a gamma distribution to generate synthetic monthly rainfall totals and proposed a rainfall generator algorithm (rGEN) to generate synthetic rainfall totals on different timescales (daily, monthly and yearly). Rosenberg et al. (2004) generalised the gamma density to a series of associated Laguerre polynomials to generate monthly rainfall totals and showed that the simulated data and the observed data had the same statistical characteristics.

Usually, it is fairly easy to model the independent variables. The modelling will become complicated whenever there exist correlation between the variables which add extra parameters such as correlation parameter. Kotz and Adams (1964) derived a distribution of a sum for identically correlated gamma variables. They compared two forms of distribution for the sum, approximate and exact under the assumption that the sum is also a gamma variable. Moschopoulos (1985) extended the technique developed by Kotz and Adams (1964) but he considered to the sum of n independent gamma variables with both different shape (α) and scale (β) parameters. Alouini et al. (2001) extended the work of Moschopoulos (1985) from the sum of n independent gamma variables with both different shape and scale parameters to the sum of n correlated gamma variables with a common shape but distinct scale parameters. Alouini et al. (2001) applied the sum of uncorrelated and correlated gamma variables in the field of telecommunication whereas Zakaria et al. (2012, 2013) uses the same models to model monthly rainfall totals for rainfall data in the Murray-Darling basin and peninsular Malaysia.

Two models of the sum of two correlated gamma variables, namely Alouini's model and McKay distribution are studied. Alouini's model is an extension of Moschopoulos results for the sum of n correlated gamma variables. There are two forms of McKay distribution, the Type I is defined for sum of two correlated gamma variables whereas Type II is defined for difference of two correlated gamma variables. In this study, the Alouini's model and Type I of McKay distribution are compared using monthly rainfall totals to form water catchment from two meteorological stations, Hume and Beechworth within the Murray-Darling Basin. Rainfall totals during the summer season at both stations are selected based on the significant correlations. The Kolmogorov-Smirnov goodness of fit test is used to assess the fit between the observed and generated rainfall data.

2 METHODS

Consider a set of n correlated gamma variables, $\{X_i\}_{i=1}^n$ with parameters α and β_i written as $X_i \sim G(\alpha, \beta_i)$ where the probability density function (PDF) of $\{X_i\}_{i=1}^n$ is given by

$$f_i(x_i; \alpha, \beta_i) = \frac{x_i^{\alpha-1} e^{-\frac{x_i}{\beta_i}}}{\Gamma(\alpha)\beta_i^{\alpha}} \tag{1}$$

for $x_i > 0$ and $\alpha, \beta_i > 0$. Since the rainfall data is usually non normal, the Spearman's rank correlation coefficient is used, ρ_{ij} of X_i and X_j with the formula of

$$\rho_{ij} = 1 - \frac{6 \sum_{k=1}^{n} d_{ijk}^2}{n(n^2 - 1)}$$

where the difference between the ranked pairs $R(x_{i,k})$ and $R(x_{j,k})$ is denoted by $d_{ijk} = R(x_{i,k}) - R(x_{j,k})$ and n is the number of pairs. Two models of the sum of two correlated gamma variables are described as follows, namely Alouini's model and McKay distribution. The correlated gamma variables method is employed whenever the correlations between the variables are significant.

2.1 Alouini's model

Define $Y = \sum_{i=1}^{n} X_i$ as the sum of n correlated gamma variables with the PDF Alouini et al. (2001) given by

$$f_Y(y) = \prod_{i=1}^{n} \left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{\delta_k \, y^{n\alpha+k-1} \exp\left(-y/\lambda_1\right)}{\lambda_1^{n\alpha+k} \, \Gamma(n\alpha + k)} \tag{2}$$

where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues of the matrix $A = DC$ where we assume that A is positive definite and where the minimum eigenvalue is denoted by λ_1. The matrices C and D are defined by

$$D = \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{bmatrix}_{n \times n}, \quad C = \begin{bmatrix} 1 & \sqrt{\rho_{12}} & \cdots & \sqrt{\rho_{1n}} \\ \sqrt{\rho_{21}} & 1 & \cdots & \sqrt{\rho_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\rho_{n1}} & \sqrt{\rho_{n2}} & \cdots & 1 \end{bmatrix}_{n \times n}.$$

Note that the application of the Alouini et al. (2001) method depends on knowing that A is positive definite and this condition must be tested by calculating the eigenvalues of A for each potential application. According to Alouini et al. (2001), it would be reasonable to assume that the matrix C is guaranteed to be positive definite. The coefficients of δ_k are defined by

$$\delta_{k+1} = \frac{\alpha}{k+1} \sum_{i=1}^{k+1} \left[\sum_{j=1}^{n} \left(1 - \frac{\lambda_1}{\lambda_j}\right)^i\right] \delta_{k+1-i} \tag{3}$$

for $k = 0, 1, 2 \ldots$ where $\delta_0 = 1$. This method works well and δ_k converges rapidly to zero if the terms $\left(1 - \frac{\lambda_1}{\lambda_j}\right)$ are all small.

2.2 McKay distribution

The Type I McKay distribution is suitable to be used to define the sum of two correlated gamma variables studied by Holm and Alouini (2004). The full derivation is not shown in Holm and Alouini (2004). However, Zakaria (2011) has presented the derivation of the PDF of the sum of two correlated gamma variables in order to convey some additional appreciation of the necessary analysis.

Let $X_1, X_2 \sim G(\alpha, \beta_1, \beta_2, \rho)$ and consider the sum $\Sigma = X_1 + X_2$ follows Type I McKay distribution and the PDF of the sum of two correlated gamma variables is given by

$$f_\Sigma(\sigma) = \frac{\sqrt{\pi}}{\Gamma(\alpha)} \left(\frac{\sigma}{2\widetilde{\beta}}\right)^{\alpha - \frac{1}{2}} \left[\frac{1}{\beta_1 \beta_2 (1-\rho)}\right]^{\alpha} e^{-\frac{\sigma(\beta_1+\beta_2)}{2\beta_1\beta_2(1-\rho)}} I_{\alpha-\frac{1}{2}}(\sigma\widetilde{\beta}).$$

where $\widetilde{\beta}^2 = \frac{(\beta_1-\beta_2)^2 + 4\beta_1\beta_2\rho}{4\beta_1^2\beta_2^2(1-\rho)^2}$. The shape and scale parameters are denoted by α and β, respectively. The Spearman's rank correlation coefficient, ρ is used as described in 2.1.

2.3 A case study for the sum of two correlated gamma random variables

Two meteorological stations within the Murray-Darling Basin are selected to illustrate the Alouini's model and McKay distribution for the sum of two correlated gamma variables. The stations are Hume Reservoir

(Hume) and Beechworth Composite (Beec) which are located in the upper Murray river and are approximately 72.6 km apart. The monthly rainfall totals during the summer months (December, January, February) for the period from 1922 to 2007 are chosen for this study and have been obtained from the Australian Bureau of Meteorological Station. The details are shown in Table 1.

Table 1. Geographic coordinates, elevations and mean of rainfall in summer months

Station	State	Latitude (°S)	Longitude (°E)	Elevation (m)	Mean rainfall (mm)
Hume Reservoir	NSW	36.10	147.03	184	45.95
Beechworth Composite	Vic	36.70	146.71	580	52.27

The procedure for the analysis is as follows:

1. Select positive pairs set of data from two stations

2. Calculate the pairwise Spearman's correlation for the positive pairs and check if the correlation is significant

3. Fit the data marginally with the gamma distribution and use maximum likelihood estimation method to estimate the parameters for the gamma distribution

4. Calculate the average value of shape parameters and re-estimate the values of scale parameters

5. Apply the model to generate synthetic sum of monthly rainfall totals

6. Compare the sum of monthly rainfall totals between the observed and generated data using the Kolmogorov-Smirnov goodness of fit test

Firstly, the rainfall data from each station is fitted to gamma distribution and the mean and variance are calculated, see Table 2. The value of mean for the observed data matches the estimated mean. Since the models for the sum of two correlated gamma variables requires single α value and for convenience the average value of α is used. Then, the values of scale parameters are re-estimated using a fixed average value of α. The correlation coefficient between the positive pairs of data is found to be significant at 5% significance level. The values of parameters calculated and used for both models are: $\alpha = 1.1531$, $\beta_1 = 40.5820$, $\beta_2 = 49.6806$ and $\rho = 0.8997$.

Secondly, the sum of the rainfall totals from both stations is calculated based on the positive pairs. There are 239 positive pairs in the data. Then, the synthetic rainfall totals are generated using the Alouini's model and McKay distribution of the estimated parameters. The fit between the observed and generated data is assessed using the Kolmogorov-Smirnov goodness of fit test. Based on the Kolmogorov-Smirnov goodness of fit test, if the P-value is greater than the significance level α then it indicates that there is no evidence that the two distributions are significantly different from one another. For further details, see Zakaria (2011).

Table 2. Shape and scale parameters, mean and variance of Hume and Beechworth, 1922–2007

|	Parameter		Mean (mm)		Variance	
	α	β	observed	$\alpha\beta$	observed	$\alpha\beta^2$
Hume	1.0132	46.1847	46.7950	46.7950	1983.7930	2161.2120
Beechworth	1.2930	44.3058	57.2866	57.2866	2649.3870	2538.1300

3 RESULTS AND DISCUSSION

The statistics of the observed monthly rainfall and generated data for the sum of two correlated gamma variables are shown in Table 3. The values of mean and variance of the observed and estimated from McKay distribution is closer than the values of mean and variance estimated from the Alouini's model. The observed rainfall totals at 5% and 95% percentiles are compared between the two models. Based on the 5% and 95% percentiles, the rainfall totals generated by McKay distribution give better results. Furthermore, based on the Kolmogorov–Smirnov goodness of fit test, the P-value of observed versus McKay distribution is much higher than the P-value of observed versus Alouini's model, see Table 4. It shows that the McKay distribution fits the data better even though both models pass the test. However, when the Alouini's model and McKay distribution are compared, the rainfall totals generated by the two models fail the test of being from the same population. Figure 1 shows the comparison of cumulative density functions between the observed data and generated data from the Alouini's model and McKay distribution. The comparison of probability density function and cumulative density function of the Alouini's model and McKay distribution are shown in Figure 2. Zakaria (2011) shows that the Alouini's model is suitable in modelling the sum of four correlated gamma variables and can easily be extended to more than two variables. On the other hand, Zakaria (2011) highlights that the McKay model is not easily extendable to more than two variables because of extensive algebraic manipulation. Thus, in this particular example of two stations and three months, it is instructive to note that the McKay formulation can well represent the sum of individual months at the two locations, but if we wanted to represent the sum over the three months of the season, we would have to use Alouini's method.

**Table 3**. Statistics of the observed and generated data for sum of monthly rainfall

	Observed	Alouini	McKay
Mean (mm)	104.08	116.77	102.54
Standard deviation (mm)	91.89	112.43	91.35
Median (mm)	78.20	85.00	72.00
Skewness	1.37	3.02	1.46
5% percentile	8.14	17.00	8.00
95% percentile	304.54	304.10	295.05

Table 4. Kolmogorov–Smirnov goodness of fit test

	P-value
Observed Vs. Alouini	0.1127
Observed Vs. McKay	0.5960
Alouini Vs. McKay	0.0000
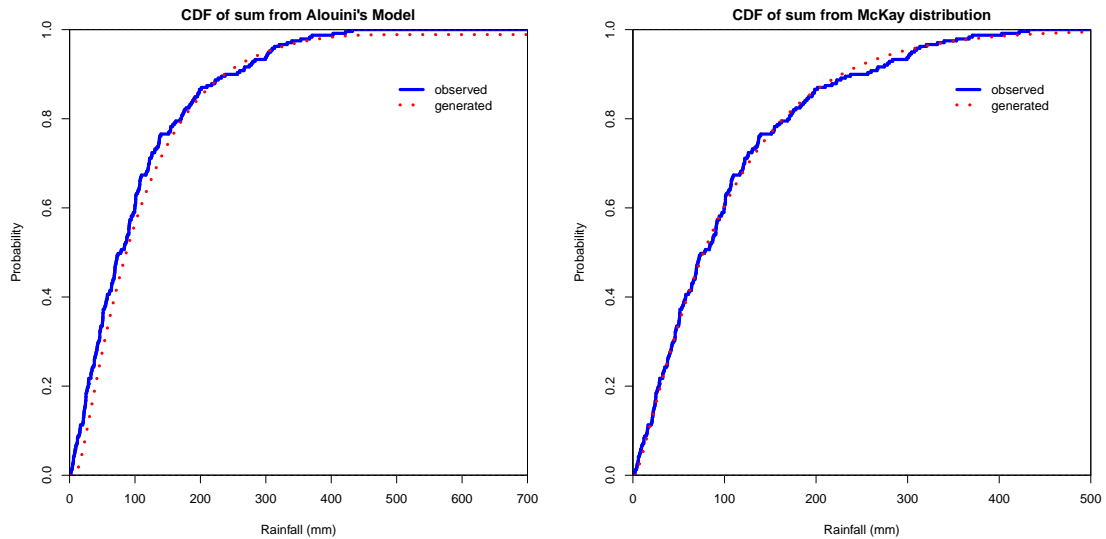
Figure 1. Observed vs generated CDF for the sum of two correlated gamma variables using the Alouini's model and McKay distribution
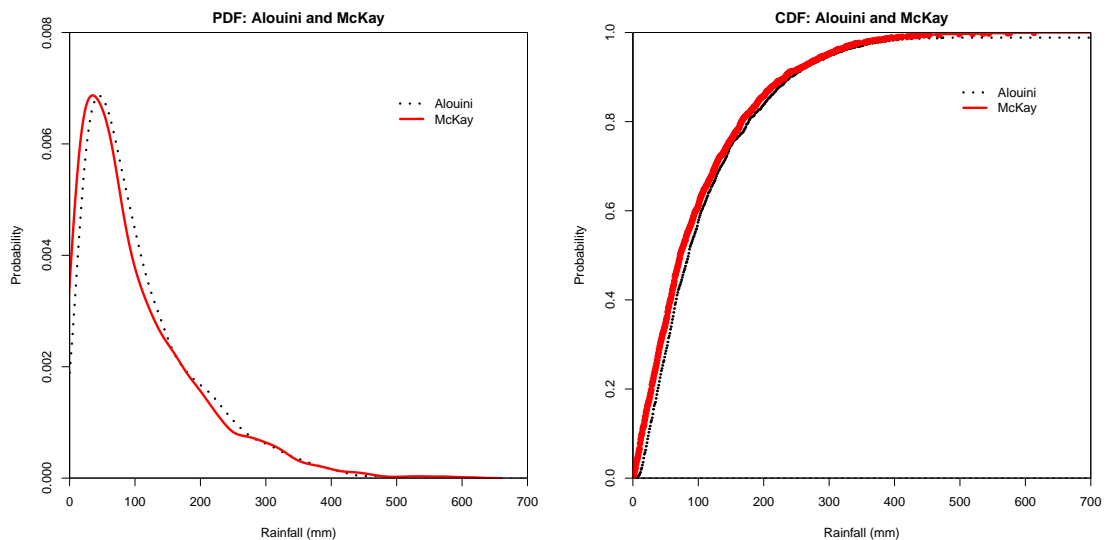
Figure 2. PDF and CDF of the sum of two correlated gamma variables for the Alouini's model and McKay distribution

4 CONCLUSIONS

This paper presents the comparison between the Alouini's model and McKay distribution to model the sum of two correlated gamma variables. Based on the Kolmogorov-Smirnov goodness of fit test, it shows that both models are suitable to model the monthly rainfall totals if the correlation between the variables is significant. However, it is found that the McKay distribution gives a better fit to the sum of two correlated gamma variables as compared to Alouini's model. Both models are able to be used to generate synthetic sum of rainfall totals and Alouini's model can be used to model the sum of more than two correlated gamma variables. It may well be that catering for this extra flexibility is the reason that the Alouini's model does not perform so well for the two variable case. Both models are also able to generate synthetic data throughout the spectrum of possibilities. In future work, we will be comparing these two approaches with other formulations as well, such

as the use of Maximum Entropy methods.

ACKNOWLEDGEMENT

REFERENCES

Alouini, M., A. Abdi, and M. Kaveh (2001). Sum of gamma variates and performance of wireless communication systems over Nakagami-fading channels. *IEEE Trans. Veh. Technol. 50*(6), 1471–1480.

Boland, J. (2005). Windfarm output variability in South Australia. In *14th International Conference on Applied Simulation and Modelling*, 15-17 July, Benalmadena, Spain.

Furman, E. (2008). On a multivariate gamma distribution. *Statistics and Probability Letters 78*, 2353–2360.

Holm, H. and M.-S. Alouini (2004). Sum and difference of two squared correlated Nakagami variates in connection with the McKay distribution. *IEEE Transactions On Communications 52*(8), 1367–1376.

Husak, G., J. Michaelsen, and C. Funk (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology 27*, 935–944.

Kotz, S. and J. Adams (1964). Distribution of sum of identically distributed exponentially correlated gamma-variables. *The Annals of Mathematical Statistics 35*(1), 277–283.

Moschopoulos, P. (1985). The distribution of the sum of independent gamma random variables. *Ann. Inst. Statist. Math.–Part A 37*, 541–544.

Paris, J. F. (2011). A note on the sum of correlated gamma random variables. *arXiv preprint arXiv:1103.0505*.

Piantadosi, J., J. Boland, and P. Howlett (2009). Simulation of rainfall totals on various time scales–daily, monthly and yearly. *Environmental Modeling and Assessment 14*(4), 431–438.

Rosenberg, K., J. Boland, and P. Howlett (2004). Simulation of monthly rainfall totals. *ANZIAM Journal 46*(E), E85–E104.

Srikanthan, R. and T. McMahon (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences 5*(4), 653–670.

Wilks, D. (1990). Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of Climate 3*(12), 1495–1501.

Wilks, D. and R. Wilby (1999). The weather generation game: A review of stochastic weather models. *Progress in Physical Geography 23*(3), 329–357.

Zakaria, R. (2011). *Mathematical modelling of rainfall in the Murray–Darling Basin*. Ph. D. thesis, University of South Australia.

Zakaria, R., P. Howlett, J. Piantadosi, J. Boland, and N. Moslim (2013). Modelling catchment rainfall using sum of correlated gamma variables. *Jurnal Teknologi 63*(2), 85–88.

Zakaria, R., P. G. Howlett, J. Piantadosi, and J. W. Boland (2012, September). Rainfall modelling using the sum of independent gamma variables. In B. Ari-Wahjoedi, R. Razali, and M. Narahari (Eds.), *American Institute of Physics Conference Series*, Volume 1482 of *American Institute of Physics Conference Series*, pp. 369–372.