Modelling rainfall amounts using mixed-gamma model for Kuantan district

Roslinazairimah Zakaria and Nor Hafizah Moslim

Citation: AIP Conference Proceedings **1842**, 020001 (2017); doi: 10.1063/1.4982831
View online: http://dx.doi.org/10.1063/1.4982831
View Table of Contents: http://aip.scitation.org/toc/apc/1842/1
Published by the American Institute of Physics

Modelling Rainfall Amounts using Mixed-Gamma Model for Kuantan District

Roslinazairimah Zakaria[1,a)] and Nor Hafizah Moslim[1,b)]

[1] *Faculty of Industrial Sciences and Technology, University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia.*

[a)] Corresponding author: roslinazairimah@ump.edu.my

[b)] fizahm@ump.edu.my

Abstract. An efficient design of flood mitigation and construction of crop growth models depend upon good understanding of the rainfall process and characteristics. Gamma distribution is usually used to model nonzero rainfall amounts. In this study, the mixed-gamma model is applied to accommodate both zero and nonzero rainfall amounts. The mixed-gamma model presented is for the independent case. The formulae of mean and variance are derived for the sum of two and three independent mixed-gamma variables, respectively. Firstly, the gamma distribution is used to model the nonzero rainfall amounts and the parameters of the distribution (shape and scale) are estimated using the maximum likelihood estimation method. Then, the mixed-gamma model is defined for both zero and nonzero rainfall amounts simultaneously. The formulae of mean and variance for the sum of two and three independent mixed-gamma variables derived are tested using the monthly rainfall amounts from rainfall stations within Kuantan district in Pahang Malaysia. Based on the Kolmogorov-Smirnov goodness of fit test, the results demonstrate that the descriptive statistics of the observed sum of rainfall amounts is not significantly different at 5% significance level from the generated sum of independent mixed-gamma variables. The methodology and formulae demonstrated can be applied to find the sum of more than three independent mixed-gamma variables.

INTRODUCTION

Most hydrological processes such as flood mitigation, water resources or environmental studies require rainfall data as the key input. The studies may be possible using the available historical rainfall data, but there is often issues highlighted regarding the record length, temporal resolution and spatial coverage. Hence, rainfall models are required to solve water related problems including temporal and spatial analyses.

Various statistical models such as gamma model has been widely used to model rainfall amounts and Markov model for modelling rainfall occurrence. In most cases, the rainfall data is positively skewed distribution. Hence, the gamma distribution is chosen due to its simplicity with only two parameters, shape and scale, [1, 2, 3, 4, 5, 6, 7, 11]. Exponential and Weibull distributions are also among the popular choice to model rainfall amounts, [2, 9].

A mixed-gamma distribution which considers both zero and nonzero rainfall is the focus of this study. Piantadosi et al. [10] use the mixed-gamma distribution to generate synthetic rainfall data on various time scales. Whereas, Rosenberg [8] uses a different type of mixed-gamma distribution associated with Laguerre polynomials.

The study is aimed to derive formulae of mean and variance for the sum of three independent mixed-gamma variables using the mixed-gamma model. The generation of synthetic rainfall amounts is presented using the mixed-gamma distribution. The Kolmogorov-Smirnov goodness of fit test is applied to assess the fit between the historical and generated rainfall amounts.

The 3rd ISM International Statistical Conference 2016 (ISM-III)
AIP Conf. Proc. 1842, 020001-1–020001-8; doi: 10.1063/1.4982831
Published by AIP Publishing. 978-0-7354-1512-6/$30.00

Gamma Distribution

A gamma distribution is commonly used to fit data with values distributed on the interval $(0, \infty)$. If a random variable, X has the gamma distribution with shape parameter α and scale parameter β, then X has probability density function (PDF) given by

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^{\alpha}} \tag{1}$$

for $x > 0$ and $\alpha, \beta > 0$. In this study, X represents the random variable of monthly rainfall amounts. The shape parameter controls the shape of the rainfall distribution and the scale parameter determines the variation of the rainfall data. Maximum likelihood estimation method (MLE) is used to estimate both the shape and scale parameters. The MLE method determines the set of parameters which maximize the likelihood function. The parameters are obtained by differentiating the log likelihood function with respect to the parameters of the distribution. The logarithm of the likelihood function is as follows:

$$\ln L = -N\ln\Gamma(\alpha) - N\alpha\ln\beta + (\alpha-1)\Sigma\ln x - \frac{\Sigma x}{\beta}. \tag{2}$$

The mean and variance of the gamma distribution are given by $\alpha\beta$ and $\alpha\beta^2$, respectively.

Mixed-Gamma Distribution

The gamma probability density function (PDF) in Equation (1) is defined on $x \in (0, \infty)$, for nonzero rainfall amounts. A mixed-gamma distribution is defined on $x \in [0, \infty)$, for zero and nonzero rainfall amounts [8]. The probability of zero rainfall amounts is calculated by

$$p_0 = P(X = 0) = \frac{n}{N}$$

where n is the count for zero rainfall amounts and N is the total count of the rainfall data. The cumulative distribution function (CDF) for the mixed-gamma model is given by

$$\begin{aligned} F(x) &= P[X \le x] \\ &= P[X = 0] + P[0 < X \le x] \\ &= p_0 + (1 - p_0)F_{\alpha,\beta}(x). \end{aligned} \tag{3}$$

Hence, the PDF for the mixed–gamma model is written as

$$f(x) = (1 - p_0)f(x;\alpha,\beta).$$

The mean is denoted by

$$\begin{aligned} E[X] &= p_0 \cdot 0 + (1 - p_0)E[X_{\alpha,\beta}] \\ &= (1 - p_0)\alpha\beta \end{aligned} \tag{4}$$

where $X_{\alpha,\beta} \sim G(\alpha,\beta)$ and the variance is given by

$$\begin{aligned} V[X] &= E[X^2] - (E[X])^2 \\ &= \alpha^2\beta^2(p_0 - p_0^2) + \alpha\beta^2(1 - p_0) \end{aligned} \tag{5}$$

where

$$\begin{aligned} E[X^2] &= p_0 \cdot 0^2 + (1 - p_0)E[X_{\alpha,\beta}^2] \\ &= (1 - p_0)\alpha(\alpha+1)\beta^2. \end{aligned}$$

If X is a random variable with mixed-gamma distribution and parameters p_0 for the probability of zero, α for the shape and β for the scale then we will write $X \sim G(\alpha, \beta, p_0)$. If $p_0 = 0$ then we write $G(\alpha, \beta, 0) = G(\alpha, \beta)$.

Sum of Independent Mixed-Gamma Variables

In this study, the formulae of mean and variance of the sum of two and three independent mixed-gamma variables are derived as follows.

Sum of Two Independent Mixed-Gamma Variables

Consider two independent mixed-gamma random variables X and Y, both on $[0, \infty)$ with densities defined and denoted by $X \sim G(\alpha_1, \beta, p_{01})$ and $Y \sim G(\alpha_2, \beta, p_{02})$ where we assume both have the same scale parameters ($\beta_1 = \beta_2 = \beta$). The densities for the sum of two independent mixed-gamma variables X, Y denoted by $S = X + Y$ where

$$\begin{aligned} F(x) &= p_{01} + (1 - p_{01}) F_{\alpha_1,\beta}(x) \\ G(y) &= p_{02} + (1 - p_{02}) F_{\alpha_2,\beta}(y). \end{aligned}$$

Hence, the CDF for the sum of two independent mixed-gamma variables is

$$\begin{aligned} H(s) &= P[X + Y \leq s] \\ &= p_{01}p_{02} + p_{01}(1 - p_{02})F_{\alpha_2,\beta}(s) + p_{02}(1 - p_{01})F_{\alpha_1,\beta}(s) \\ &\quad +(1 - p_{01})(1 - p_{02})F_{\alpha_1+\alpha_2,\beta}(s). \end{aligned} \tag{6}$$

The mean of the sum is calculated using the result obtained in Equation (4) and is given by

$$\begin{aligned} E[S] &= p_{01}p_{02} \cdot 0 + p_{01}(1 - p_{02})\alpha_2\beta + p_{02}(1 - p_{01})\alpha_1\beta \\ &\quad +(1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2)\beta \\ &= \beta[\alpha_1(1 - p_{01}) + \alpha_2(1 - p_{02})]. \end{aligned}$$

The variance $V[S]$ of the sum is

$$\begin{aligned} V[S] &= E[S^2] - (E[S])^2 \\ &= \beta^2[\alpha_1(\alpha_1 p_{01} + 1)(1 - p_{01}) + \alpha_2(\alpha_2 p_{02} + 1)(1 - p_{02})] \end{aligned}$$

where $E[S^2]$ is calculated using the result from Equation (5) and is shown as follows.

$$\begin{aligned} E[S^2] &= \beta^2 \left[\alpha_1(\alpha_1 + 1)(1 - p_{01}) + \alpha_2(\alpha_2 + 1)(1 - p_{02}) \right. \\ &\quad \left. +2\alpha_1\alpha_2(1 - p_{01})(1 - p_{02})\right]. \end{aligned}$$

If $\alpha_1 = \alpha_2 = \alpha$, then the mean is

$$E[S] = \alpha\beta[2 - p_{01} - p_{02}] \tag{7}$$

and

$$E[S^2] = \alpha^2\beta^2(4 - 3p_{01} - 3p_{02} + 2p_{01}p_{02}) + \alpha\beta^2(2 - p_{01} - p_{02}).$$

Hence, the variance is

$$V[S] = \alpha^2\beta^2(p_{01} + p_{02} - p_{01}^2 - p_{02}^2) + \alpha\beta^2(2 - p_{01} - p_{02}). \tag{8}$$

Sum of Three Independent Mixed-Gamma Variables

Similar procedures are employed to derive the PDFs and to obtain the formulae of mean and variance for the sum of three independent mixed-gamma variables. Consider three independent mixed-gamma variables defined on $[0, \infty)$ denoted by $X \sim G(\alpha_1, \beta, p_{01})$, $Y \sim G(\alpha_2, \beta, p_{02})$ and $Z \sim G(\alpha_3, \beta, p_{03})$ where the shape parameters are distinct and with the same scale parameters ($\beta_1 = \beta_2 = \beta_3 = \beta$). To derive the density for the sum of three independent mixed-gamma variables X, Y and Z, let $S = X + Y + Z$ denote the sum where the CDF for each mixed-gamma variables are given by

$$\begin{aligned} F(x) &= p_{01} + (1 - p_{01}) F_{\alpha_1,\beta}(x) \\ G(y) &= p_{02} + (1 - p_{02}) F_{\alpha_2,\beta}(y) \\ H(z) &= p_{03} + (1 - p_{03}) F_{\alpha_3,\beta}(z). \end{aligned}$$

If $X_1 \sim G(\alpha_1,\beta)$ and $X_2 \sim G(\alpha_2,\beta)$ then $X_1 + X_2 \sim G(\alpha_1 + \alpha_2,\beta)$. Therefore, $X_1 + X_2 \sim G(\alpha_1 + \alpha_2,\beta)$ and $X_3 \sim G(\alpha_3,\beta)$ it follows from the previous argument that $(X_1 + X_2) + X_3 \sim G((\alpha_1 + \alpha_2) + \alpha_3,\beta)$. Hence the CDF for the sum of three independent mixed-gamma variables is

$$\begin{aligned} I(s) = \; & p_{01}p_{02}p_{03} + p_{01}p_{02}(1 - p_{03})F_{\alpha_3,\beta}(s) \\ & + p_{01}p_{03}(1 - p_{02})F_{\alpha_2,\beta}(s) + p_{02}p_{03}(1 - p_{01})F_{\alpha_1,\beta}(s) \\ & + p_{01}(1 - p_{02})(1 - p_{03})F_{\alpha_2+\alpha_3,\beta}(s) \\ & + p_{02}(1 - p_{01})(1 - p_{03})F_{\alpha_1+\alpha_3,\beta}(s) \\ & + p_{03}(1 - p_{01})(1 - p_{02})F_{\alpha_1+\alpha_2,\beta}(s) \\ & + (1 - p_{01})(1 - p_{02})(1 - p_{03})F_{\alpha_1+\alpha_2+\alpha_3,\beta}(s). \end{aligned} \tag{9}$$

The mean of the sum is calculated using the result obtained in Equation (4), and is given by

$$\begin{aligned} E[S] = \; & \beta \left[p_{01}p_{02}(1 - p_{03})\alpha_3 + p_{01}p_{03}(1 - p_{02})\alpha_2 + p_{02}p_{03}(1 - p_{01})\alpha_1\right. \\ & + p_{01}(1 - p_{02})(1 - p_{03})(\alpha_2 + \alpha_3) + p_{02}(1 - p_{01})(1 - p_{03})(\alpha_1 + \alpha_3) \\ & + p_{03}(1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2) \\ & \left. + (1 - p_{01})(1 - p_{02})(1 - p_{03})(\alpha_1 + \alpha_2 + \alpha_3)\right]. \end{aligned}$$

Similarly, $E[S^2]$ is calculated using the result from Equation (5),

$$\begin{aligned} E[S^2] = \; & \beta^2[p_{01}p_{02}(1 - p_{03})\alpha_3(\alpha_3 + 1) + p_{01}p_{03}(1 - p_{02})\alpha_2(\alpha_2 + 1) \\ & + p_{02}p_{03}(1 - p_{01})\alpha_1(\alpha_1 + 1) \\ & + p_{01}(1 - p_{02})(1 - p_{03})(\alpha_2 + \alpha_3)(\alpha_2 + \alpha_3 + 1) \\ & + p_{02}(1 - p_{01})(1 - p_{03})(\alpha_1 + \alpha_3)(\alpha_1 + \alpha_3 + 1) \\ & + p_{03}(1 - p_{01})(1 - p_{02})(\alpha_1 + \alpha_2)(\alpha_1 + \alpha_2 + 1) \\ & + (1 - p_{01})(1 - p_{02})(1 - p_{03})(\alpha_1 + \alpha_2 + \alpha_3)(\alpha_1 + \alpha_2 + \alpha_3 + 1)]. \end{aligned}$$

If $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$, then the mean is

$$E[S] = \alpha\beta[3 - (p_{01} + p_{02} + p_{03})]. \tag{10}$$

Hence, the variance of the sum of three independent mixed-gamma variables is

$$\begin{aligned} V[S] = \; & \alpha^2\beta^2(p_{01} + p_{02} + p_{03} - p_{01}^2 - p_{02}^2 - p_{03}^2) \\ & + \alpha\beta^2(3 - p_{01} - p_{02} - p_{03}). \end{aligned} \tag{11}$$

The reader is referred to Zakaria [11] for further details of the derivation of means and variances of sum of two and three independent mixed-gamma variables.

CASE STUDY

To illustrate the methodology discussed, the monthly rainfall data from six rainfall stations located within Kuantan District are selected. Kuantan city is located in the east coast of peninsular Malaysia and facing the South China Sea. The detailed of the location is given in Fig. 1 and Table 1. Table 1 also shows the percentage of missing values in the

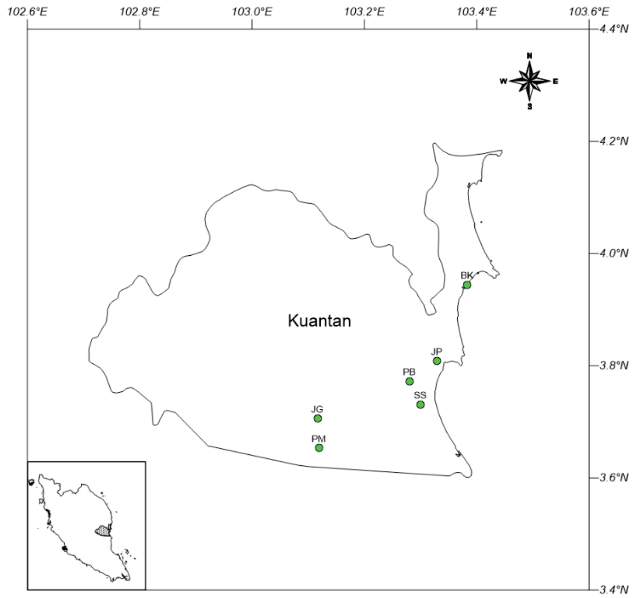

FIGURE 1. Map of selected rainfall stations in Kuantan District, Pahang.

TABLE 1. Location of selected rainfall stations in Kuantan District, Pahang

Station Code	Station Name	Latitude	Longitude	% of Missing Value
BK	Balok Kuantan	3.9444°	103.3833°	2.9
JG	JKR Gambang	3.7056°	103.1167°	12.1
SS	Kg Sg Soi	3.7306°	103.3000°	3.7
JP	Pej. JPS Pahang	3.8083°	103.3292°	11.1
PB	Paya Besar Kuantan	3.7722°	103.2806°	11.7
PM	Kg Pulau Manis	3.6528°	103.1194°	5.5

rainfall data which is less than 15%. The distances between the stations are within 5 km and 45 km apart. The original rainfall data chosen is from the hourly data aggregated monthly for the period 2009–2014.

Before conducting further analysis, the positive rainfall amounts from each station is fitted to the gamma distribution. The gamma distribution is only applicable for nonzero data. The MLE method is then used to estimate the shape (α) and scale (β) parameters. The summary of the estimated parameters are shown in Table 2. From Table 2, it is noted that the values of α and β from Balok (BK), JPS Pahang (JP) and Paya Besar (PB) stations are quite similar. Hence, these stations are chosen to illustrate the methodology of sum of independent mixed–gamma model. The synthetic nonzero monthly rainfall amounts are generated marginally using gamma distribution for each pair of α and β parameters ($\alpha_i, \beta_i;\ i = 1, 2, 3$), respectively. The descriptive statistics of mean and variance between the observed and generated data and using formulae based on gamma distribution (mean=$\alpha\beta$; variance=$\alpha\beta^2$) are compared and shown in Table 3.

The mixed–gamma model as described in Equation (3) is defined for zero and nonzero data. The mixed–gamma model is still based on the gamma distribution plus the probability of zero rainfall. Similar procedure is followed to generate the synthetic rainfall amounts which include zero and nonzero values. The synthetic rainfall amounts for each selected station are generated marginally based on the values of p_{0i}, α_i and β_i for $i = 1, 2, 3$, respectively as given in Table 2. The mean and variance for the generated data from mixed–gamma distribution and using formulae given by Equation (4) and (5) are compared as shown in Table 3. In conclusion, the values of mean from both models are reasonably close to the observed mean, however the values of variance are sometimes overestimated or underestimated the observed values. When there is zero data exist, the formula of mean for mixed-gamma agrees with the observed mean. Otherwise, the observed mean matches the mean of gamma formula since it defines for the nonzero data.

TABLE 2. Estimated shape, scale parameters and probability of zero rainfall for stations in Kuantan district, 2009-2014

Station	α	β	p_0
BK	1.2046	197.3841	0.0000
JG	2.1747	111.0202	0.0000
SS	1.4074	179.5236	0.0000
JP	1.2691	181.0682	0.0972
PB	1.1709	192.8375	0.0972
PM	1.9032	105.7709	0.0139

TABLE 3. Comparison of mean (mm) and variance for the observed, formula and generated monthly rainfall data using gamma and mixed-gamma distributions

			Gamma		Mixed-gamma	
Value	Station	Observed	Generated	Formula	Generated	Formula
Mean	BK	237.77	245.92	237.77	236.53	237.77
	JP	207.45	226.40	229.79	212.98	207.46
	PB	203.84	229.89	225.80	205.93	203.85
Variance	BK	55691.08	54221.93	46932.71	45992.26	46931.80
	JP	58724.84	40690.69	41607.51	43680.05	42197.76
	PB	54932.92	42972.87	43542.25	44241.40	43783.05

Mean and Variance for Sum of Independent Mixed-gamma Distributions

Formulae of mean and variance for sum of two independent mixed-gamma distributions are presented by Equation 7 and 8, respectively. Similarly, for sum of three independent mixed-gamma distributions, the mean and variance are given by Equation 10 and 11, respectively. The formulae are tested using data from three stations (Balok, JPS Pahang and Paya Besar).

TABLE 4. Estimated parameters, mean (mm) and variance for sum of two and three stations using formulae from gamma distribution

	Estimated parameters			Mean	Variance
	p_0	α	β	$\alpha\beta$	$\alpha\beta^2$
Sum of Two	0.0000	1.1362	361.9840	411.29	148880.50
Sum of Three	0.0000	1.5419	420.9531	649.06	273225.40

Formulae of Mean and Variance for Sum of Two Independent Mixed-gamma Distributions

To demonstrate the formula of sum of two independent mixed-gamma variables, using data from JPS Pahang and Paya Besar stations, firstly construct the sum of the stations by adding the corresponding rainfall data. Table 4 shows the estimated parameters including shape, scale and probability of zero of gamma distribution. The generated sum is formed from the sum of two individual generated set of the mixed-gamma distribution. In each generated set, the average of shape parameter ($\overline{\alpha} = (\alpha_1 + \alpha_2)/2 = 1.2200$), the average of scale parameter ($\overline{\beta} = (\beta_1 + \beta_2)/2 = 186.9529$) and the respective probability of zero rainfall ($p_1 = 0.0972, p_2 = 0.0972$) are used. Table 5 shows the comparison between the mean and variance of sum of two stations for the observed data, generated data and using formula of mixed-gamma distribution.

The values of mean are almost similar between observed, generated and calculated using formula. However, the variances are underestimated for the generated and using formula as compared to the observed value. Based on the Kolmogorov-Smirnov goodness of fit test, P-value is greater than 0.05 significance level, indicates that there is no significant difference between the observed and the generated data.

TABLE 5. Compare mean (mm) and variance of sum between two stations using mixed-gamma distribution

	Observed	Generated	Formula
Mean	411.29	400.52	411.83
Variance	210495.33	84238.57	86122.01
P-value	0.4310		-

Formulae of Mean and Variance for Sum of Three Independent Mixed-gamma Distributions

Similar procedure is applied to test the formulae for sum of three independent mixed-gamma variables. In Table 6, the results obtained indicate that the mean are reasonably similar. However, the variance from the generated data and using formula are underestimated from the observed value. The P-value of the Kolmogorov-Smirnov goodness of fit test is greater than 0.05 significance level. Hence, the observed and the generated data are not significantly different for both data.

TABLE 6. Compare mean (mm) and variance of sum between three stations using mixed-gamma distribution

	Observed	Generated	Formula
Mean	649.06	638.78	649.07
Variance	349222.37	133660.53	132995.00
P-value	0.05335		-

CONCLUSION

In this study, the formulae of mean and variance for the sum of two and three independent mixed-gamma variables are derived. The formulae for the simplest case when the shape and scale are common are also presented. For illustration, the formulae are tested using data from the stations within Kuantan district. The estimated values of mean are almost similar for both cases. However, the estimated values of variance is always underestimated, which is due to correlation between the variables. In general, the results show that the generated sum of two and three are not significantly different from the observed sum based on the Kolmogorov-Smirnov goodness of fit test. Hence, the methodology can be applied to sum of more than three independent mixed-gamma variables and also be used to generate synthetic rainfall data which able to solve problem in missing data. In future, we will consider rainfall model which include the correlation parameter for better results.

ACKNOWLEDGMENTS

The research is partly supported by University Malaysia Pahang. We thank Mr Muhammad Az-zuhri for preparing the Kuantan map.

REFERENCES

[1] N. T. Ison, A. M. Feyerherm and B. L. Dean, Journal of Applied Meteorology, **10:419-428**, (1971).
[2] D. A. Woolhiser and J. Roldan, Water Resources Research, **22(6):965-978**, (1982).
[3] R. W. Kats and M. B. Parlange, Journal of Climate, **11(4):591-601**, (1998).
[4] D. S. Wilks and R. L. Wilby, Progress in Physical Geography, **23(3):329-357**, (1998).
[5] H. Aksoy, Turk J Engin Environ Sci, **24:419-428**, (2000).
[6] R. Srikanthan and T.A. McMahon, Hydrology and Earth System Sciences, **5(4):653-670**, (2001).
[7] K. Rosenberg, J. W. Boland and P. G. Howlett, ANZIAM Journal, **46(E):85-104**, (2004).

[8] K. Rosenberg, Stochastic modelling of rainfall and generation of synthetic rainfall data at Mawson Lakes, Ph.D thesis, University of South Australia, (2004).
[9] K. Sharda and P.K. Das, Agricultural Water Management, **76(2):120-138**, (2005).
[10] J. Piantadosi, J. W. Boland and P. G. Howlett, Environmental Modeling and Assessment, **14(4):431-438**, (2009).
[11] R. Zakaria, Mathematical modelling of rainfall in the Murray-Darling Basin, Ph.D thesis, University of South Australia, (2011).