Jurnal Teknologi

Full paper

Modelling Catchment Rainfall Using Sum of Correlated Gamma Variables

Zakaria, R.[a*], Howlett, P. G.[b], Piantadosi, J.[b], Boland, J. W.[b], Moslim, N. H.[a]

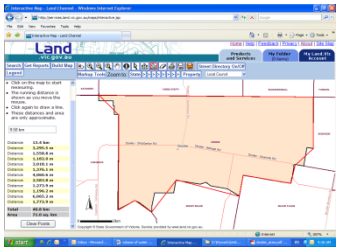
[a]*Faculty of Industrial Sciences and Technology, University Malaysia Pahang, Lebuhraya Tun Razak 26300 Gambang, Kuantan Pahang, Malaysia*
[b]*School of Mathematics and Statistics, University of South Australia, Mawson Lakes, 5095, South Australia, Australia*

*Corresponding author: roslinazairimah@ump.edu.my

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract



Abstract

One of the major difficulties in simulating rainfall is the need to accurately represent rainfall accumulations. An accurate simulation of monthly rainfall should also provide an accurate simulation of yearly rainfall by summing the monthly totals. A major problem in this regard is that rainfall distributions for successive months may not be independent. Thus the rainfall accumulation problem must be represented as the summation of dependent random variables. This study is aimed to show if the statistical parameters for several stations within a particular catchment is known, then a weighted sum is used to determine a rainfall model for the entire local catchment. A spatial analysis for the sum of rainfall volumes from four selected meteorological stations within the same region using the monthly rainfall data is conducted. The sum of n correlated gamma variables is used to model the sum of monthly rainfall totals from four stations when there is significant correlation between the stations.

Keywords: Rainfall model; correlated gamma; catchment rainfall

Abstrak

Salah satu masalah utama dalam simulasi hujan adalah untuk menunjukkan pengumpulan hujan dengan tepat. Simulasi hujan bulanan yang tepat juga menghasilkan simulasi hujan tahunan yang tepat dengan menjumlahkan jumlah hujan bulanan. Satu masalah utama dalam hal ini adalah bahawa taburan hujan bulanan yang berturutan mungkin bersandar diantara satu sama lain. Oleh itu, masalah pengumpulan hujan mesti diwakili sebagai hasil tambah pemboleh ubah rawak bersandar. Kajian ini bertujuan menunjukkan jika parameter statistik untuk beberapa stesen dalam kawasan tadahan tertentu yang diketahui, maka jumlah wajaran digunakan untuk menentukan model hujan untuk keseluruhan kawasan tadahan tempatan. Satu analisis ruang bagi jumlah hujan daripada empat stesen meteorologi terpilih dalam kawasan yang sama dengan menggunakan data hujan bulanan dijalankan. Hasil tambah *n* pemboleh ubah yang berkolerasi gamma digunakan dalam memodelkan hasil tambah jumlah hujan secara bulanan daripada empat stesen berdasarkan kepada nilai kolerasi yang signifikan diantara stesen-stesen.

Kata kunci: Model hujan; gamma berkolerasi; kawasan tadahan hujan

1.0 INTRODUCTION

A search for the best rainfall model is always a challenging task. The rainfall data can be used as discrete form or continuous form. In the discrete form, the study is based on rainfall occurrences whereas in the continuous form, it is using the rainfall amounts. To model rainfall occurrences, a Markov chain model has been used widely in many application such as climatology[1-3]. To model rainfall occurrences, a gamma distribution is always a popular choice. The gamma distribution is chosen because it is suitable for modelling continuous variables that are always positive and have skewed distributions like rainfall amounts. It is also flexible as it involves two parameters, shape and scale[4-5]. In the previous studies, gamma distribution is used widely in hydrology and climatology to model rainfall[6-7], food and drought[4] and wind farm output[8]. The gamma distribution is also used to generate synthetic rainfall data.[9-12] Synthetic rainfall data generation is important as a supplement to the historical rainfall data which is always lacking in terms of temporal and spatial details. A comprehensive review of weather generation models is given by Srikanthan and McMahon[11] and Wilks and Wilby[12].

In this study, an extended form of gamma distribution, namely the sum of *n* correlated gamma variables is used to model the accumulation of monthly rainfall totals from four meteorological stations. The sum of *n* correlated gamma variables has been used by Alouini *et al.*[13] in the field of telecommunications. Originally, the model is introduced by Moschopoulos[14]. Tran and Sesay[15] and Chen[16] also apply the sum of *n* correlated gamma variables in the same field. Chen[16] extends the work of Alouini *et al.*[13] to improve the process of deriving the probability density function of the signal-to-noise ratio (SNR) explicitly. In the model, Alouini *et al.*[13] used a single shape

parameter α with different scale parameters β. The basis of this model came from Moschopoulos[14], in which for the sum of n independent random variables and with both different shape and scale parameters. The model is able to simulate the synthetic yearly rainfall by summing the monthly totals.

2.0 AREA OF STUDY

Murray-Darling Basin is situated at the south-east of Australia and is the most important river system that gives life to most part of Australia. Four meteorological stations within the Murray-Darling Basin are selected, namely Hume Reservoir, Beechworth Composite, Dookie and Rutherglen Research. The details are shown in Table 1. The monthly rainfall data is used based on completeness and the years of data selected are between 1922 and 2008. The summary of yearly rainfall means and standard deviations is given in Table 1. Of all the selected stations, Beechworth has the highest yearly mean of rainfall (82.96 mm) and Dookie has the lowest yearly mean of rainfall (46.32 mm). The catchment areas (km^2) are approximated using the interactive map software from the Victorian Government land services and spatial information website, see (http://www.land.vic.gov.au). An example on how to use the interactive map is shown for Dookie, see Figure 1.

Table 1 Geographic coordinates, elevations and yearly rainfall for selected meteorological stations in the Murray-Darling Basin

Station	State	Latitude (° S)	Longitude (° E)	Elevation (m)	Yearly (mm)	
					Mean	Sd
Hume Reservoir	NSW	36.10	147.03	184	58.11	40.83
Beechworth Composite	Vic	36.70	146.71	580	82.96	52.78
Dookie	Vic	36.37	145.70	185	46.32	34.87
Rutherglen	Vic	36.10	146.51	175	49.03	35.91

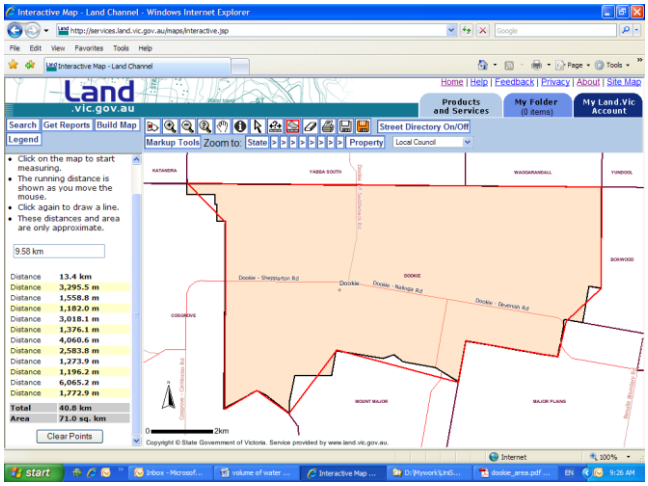


Figure 1 Approximate area of Dookie catchment

Since the rainfall data is usually skewed, the use of Spearman's rank correlation coefficient is more appropriate for determining correlation between rainfall totals at separate locations. The formula of the Spearman's rank correlation coefficient is calculated based on the ranks of each data element. This is a nonparametric measure. Consider two sets of data, $\{X_i\}$ and $\{Y_i\}$, for $i=1,\ldots,n$. First, we rank both sets of data separately as $R(x_i)$ and $R(y_i)$, for $i=1,\ldots,n$ from the smallest to the largest. Then, the formula for the Spearman's rank correlation coefficients r_{ij} of X_i and X_j is

$$r_{ij}=1-\frac{6\sum_{k=1}^{n}d_k^2}{n\left(n^2-1\right)} \quad (2)$$

where $d_k=R(x_k)-R(y_k)$ is the difference between the ranked pairs and n is the number of pairs.

3.0 METHODOLOGY

3.1 Modelling Monthly Rainfall Totals

A gamma distribution is used to model the monthly rainfall totals for each selected meteorological station. The probability density function of gamma distribution with two parameters, α and β denote the shape and scale parameters, respectively is given by

$$f_X(x;\alpha,\beta)=\frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \text{ for } x>0 \text{ and } \alpha,\beta>0 \quad (1)$$

where X is a random variable of gamma distribution denoted by $X\sim G(\alpha,\beta)$. For convenience, the maximum likelihood estimation (MLE) method is used to estimate both parameters. The MLE method maximises the logarithm of the likelihood function of the gamma distribution by differentiating the log likelihood function with respect to the parameters of the distribution.

3.2 Modelling the Sum of Correlated Gamma Variables

Using Alouini *et al.*[13] approach, consider a set of n correlated gamma variables, $\{X_i\}_{i=1}^{n}$ with parameters α and β_i written as $X_i\sim G(\alpha,\beta_i)$. The correlation ρ_{ij} of X_i and X_j are $\rho_{ij}=\frac{\text{cov}(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$ where $0<\rho_{ij}<1$ for $i,j=1,2,\ldots,n$. The probability density function of the sum of n correlated gamma variables, $Y=\sum_{i=1}^{n}X_i$ is

$$f_Y(y)=\prod_{i=1}^{n}\left(\frac{\lambda_1}{\lambda_i}\right)^{\alpha}\sum_{k=0}^{\infty}\frac{\delta_k y^{n\alpha+k-1}\exp(-y/\lambda_1)}{\lambda_1^{n\alpha+k}\Gamma(n\alpha+k)} \quad (3)$$

where $\{\lambda_i\}_{i=1}^{n}$ are eigenvalues of matrix $A=DC$ and $\lambda_1=\min_n\{\lambda_n\}$. The matrix D is a diagonal matrix of $\{\beta_i\}_{i=1}^{n}$ and matrix C is a positive definite matrix formed from the correlations ρ_{ij} presented by

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

The coefficients of δ_k are calculated using

$$\delta_{k+1}=\frac{\alpha}{k+1}\sum_{i=1}^{k+1}\left[\sum_{j=1}^{n}\left(1-\frac{\lambda_1}{\lambda_j}\right)^{i}\right]\delta_{k+1-i} \quad (4)$$

for $k=0,1,2,\ldots$ where $\delta_0=1$. Zakaria[17] has shown that matrix C is not guaranteed to be positive definite. Therefore, it is necessary to check the eigenvalues of C before applying Alouini *et al.*[13] method.

3.3 Goodness of Fit Tests

In this study, two types of nonparametric goodness of fit tests are applied, namely Kolmogorov-Smirnov and Anderson-Darling to assess the fit between the empirical distribution and the specified distribution. Both goodness of fit tests used are based on the cumulative distribution functions. For the Kolmogorov-Smirnov goodness of fit test, the test statistic is

$$D_{1,2}=\max_x\left|F_1(x)-F_2(x)\right| \quad (5)$$

where $F_1(x)$ and $F_2(x)$ are the cumulative distribution functions of the observed and generated data. The value of D determines the absolute maximum difference between the CDF of the observed and generated data. For the Anderson-Darling goodness of fit test, the test statistic is

$$A_{nm}^2=\frac{1}{nm}\sum_{i=1}^{N-1}\frac{(M_iN-ni)^2}{i(N-i)} \quad (6)$$

where M_i is the count of X less than or equal to the ith smallest in the pooled sample. The cumulative distribution functions of x and y are represented by $F_n(x)$ and $G_m(x)$, respectively. For both goodness of fit tests, the P-values are compared with the significance level of $\alpha=0.05$. If the P-value is greater than $\alpha=0.05$, it indicates that there is no evidence that the observed and the generated distributions are significantly different from each other.

4.0 RESULTS AND CONCLUSIONS

The analysis of rainfall model of four meteorological stations in the Murray-Darling Basin involves two stages. In the first stage, each data set is fitted to the gamma distribution and the shape and scale parameters are estimated using the MLE method. The values of α and β associated with the volume are calculated and shown in Table 2. Then, the Spearman correlation coefficient is calculated and found to be significant at 1% significant level with correlation coefficients greater than 0.8 (see Table 3). Once the correlation is found to be significant, the sum of n correlated gamma variables is applied to model the sum of volumes of rainfall totals to form a catchment. The sum of n correlated gamma variables requires a single value of α. Hence, the average of the α is used and the values of β are re-estimated, see Table 4. Further calculations and detail procedures are described in Zakaria[17].

The graphical representation of the probability density function (PDF) and cumulative distribution function (CDF) are used together with the statistical tests to evaluate the goodness of fit between the observed and generated data. For the analysis of the sum of four correlated gamma variables, the P-values from the Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests are 0.3411 and 0.1018, respectively. The P-values show that the observed and generated data are not significantly different at 5% significance level, refer to Table 5. Figure 2 supports the result. In conclusion, the sum of n correlated gamma model is suitable for use in modelling the sum of rainfall volumes from the nearby stations within the same region which have significant correlation to obtain the total volume of a catchment. The methodology and output of the study described above will provide input for the rainfall-runoff model and can be used in generating synthetic rainfall data.

Table 2 Estimated parameters, means and variances of the observed and formulae from gamma distribution

Stations	Parameters		Mean (m^3)		Variance	
	α	β	observed	$\alpha\beta$	observed	$\alpha\beta^2$
Hume	1.67	6419000	10693000	10693000	573922000	686335000
Beechworth	1.84	3266000	6017000	6017000	165422000	196544000
Dookie	1.53	4863000	7435000	7435000	311788000	361600000
Rutherglen	1.58	3968000	6268000	6268000	210024000	248702000

Table 3 Spearman's rank correlation coefficients between the stations

Stations	Beecworth	Dookie	Rutherglen
Hume	0.90	0.84	0.91
Beechworth		0.87	0.91
Dookie			0.88

Table 4 Re-estimated values of β for $\bar{\alpha} = 1.65$

	Hume	Beechworth	Dookie	Rutherglen
β_i	6464000	3586000	4495000	3789000

Table 5 P-values of goodness of fit tests for the sum of four correlated gamma variables from four stations

	Kolmogorov-Smirnov	Anderson-Darling
Observed vs Generated	0.3411	0.1018

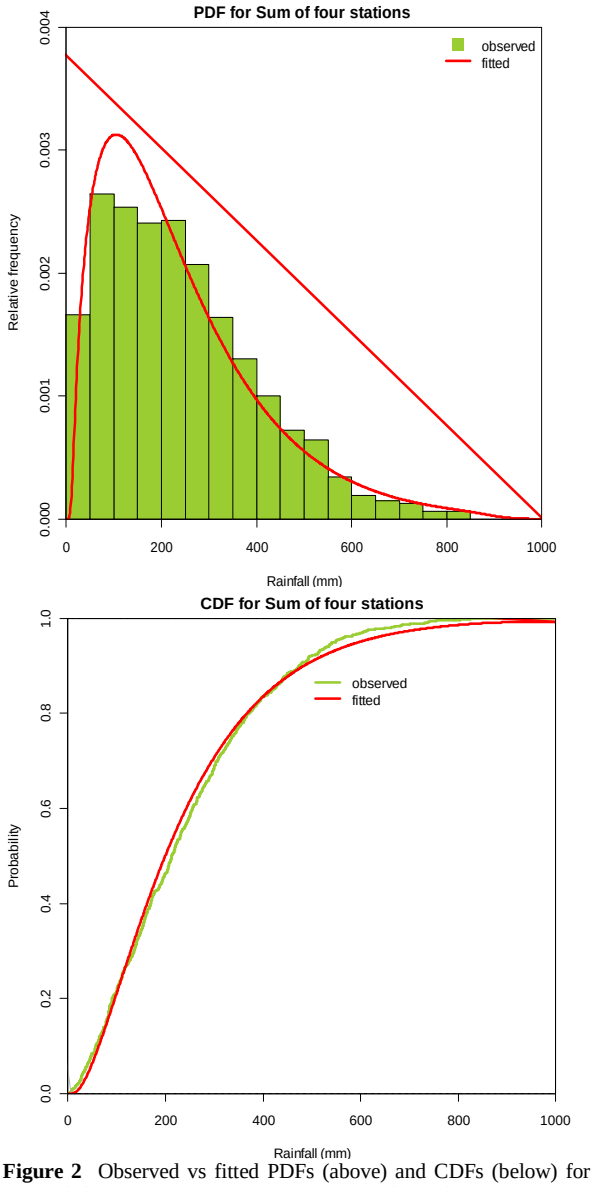


Figure 2 Observed vs fitted PDFs (above) and CDFs (below) for the sum of four correlated gamma variables from four stations Hume, Beechworth, Dookie and Rutherglen)

Acknowledgements

The authors are thankful to the Australian Bureau of Meteorological Station for providing the rainfall data. This study was supported by the Universiti Malaysia Pahang (Grant No: RDU120101).

References

[1] K. R. Gabriel, J. Neumann. 1962. A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*.

[2] S. Deni, A. Jemain, K. Ibrahim. 2009. Fitting Optimum Order of Markov Chain Models for Daily Rainfall Occurrences in Peninsular Malaysia. *Theoretical and Applied Climatology*. 97: 109–121.

[3] P. Dash. 2012. A Markov Chain Modeling of Daily Precipitation Occurences of Odisha. *International Journal of Advanced Computer and Mathematical Sciences*. 3(4): 482–486.

[4] G. Husak, J. Michaelsen, 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.

[5] D. Wilks. 1990. Maximum Likelihood Estimation for the Gamma Distribution Using Data Containing Zeros. *Journal of Climate*. 3(12): 1495–1501.

[6] H. Aksoy. 2000. Use of Gamma Distribution in Hydrological Analysis. *Turk J Engin Environ Sci*. 24: 419–428.

[7] J. Suhaila, A. Jemain. 2007. Fitting Daily Rainfall Amount in Peninsular Malaysia Using Several Types of Exponential Distributions. *Journal of Applied Sciences Research*. 3(10): 1027–1036.

[8] J. Boland. 15-17 July 2005. Windfarm Output Variability in South Australia. *In 14th International Conference on Applied Simulation and Modelling*. Benalmadena, Spain.

[9] J. Piantadosi, J. Boland, P. Howlett. 2009. Simulation of Rainfall Totals on Various Time Scales-daily, Monthly and Yearly. *Environmental Modeling and Assessment*. 14(4): 431–438.

[9] K. Rosenberg, J. Boland, P. Howlett. 2004. Simulation of Monthly Rainfall Totals. *ANZAM Journal*. 46(E): E85–E104.

[10] R. Srikanthan, T. McMahon. 2001. Stochastic Generation of Annual, Monthly and Daily Climate Data: A Review. *Hydrology and Earth System Sciences*. 5(4): 653–670.

[11] D. Wilks, R. Wilby. 1999. The Weather Generation Game: A Review of Stochastic Weather Models. *Progress in Physical Geography*. 23(3): 329–357.

[12] M. Alouini, A. Abdi, 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.

[13] P. Moschopoulos. 1985. The Distribution of the Sum of Independent Gamma Random Variables. *Ann. Inst. Statist. Math.-Part A*. 37: 541–544.

[14] T. Tran, A. Sesay. 2007. Sum of Arbitrarily Correlated Gamma Variates and Performance of Wireless Communication Systems Over Nakagami-*m* Fading Channels. *IET Communications*. 1(6): 1133–1137.

[15] J. Chen. 2006. Performance Analysis of MC-CDMA Communication Systems Over Nakagami-*m* Environments. *J Marine Sci Technol*. 14(1): 58–63.

[16] R. Zakaria. 2011. *Mathematical Modeling of Rainfall in the Murray-Darling Basin*. University of South Australia: Ph.D. Thesis.