Generating Synthetic Rainfall on Various Timescales - Daily, Monthly and Yearly

Piantadosi, J., J.W. Boland and P.G. Howlett

Centre for Industrial and Applied Mathematics
University of South Australia, Mawson Lakes Campus, Adelaide SA 5095, Australia.
E-Mail: julia.piantadosi@unisa.edu.au

Keywords: *Daily, monthly, yearly rainfall, correlation, copulas*

EXTENDED ABSTRACT

The main objective of this paper is to present a model for generating synthetic rainfall totals on various timescales to be applicable for a variety of uses. Many large-scale ecological and water resources models require daily, monthly and yearly rainfall data as input to the model. As historical data provides only one realisation, synthetic generated rainfall totals are needed to assess the impact of rainfall variability on water resources systems (Srikanthan 2005). Thus our preferred model should simulate rainfall for yearly, monthly, and daily periods. We believe that for water supply issues no higher resolution is needed although higher resolution would be useful in models designed to measure the risk of local flooding. The critical factors are daily, monthly and yearly totals and daily, monthly and yearly variation.

A model for generating yearly totals will be described using traditional time series methods. This model, along with a similarly constructed daily generation model by, Piantadosi et al. (2007b), will be cascaded to start with a synthetic yearly total, then generate a synthetic sequence of monthly totals (through selection from a large number of realisations) that match the yearly total, and subsequently perform a similar operation for sequences of daily totals to match the required monthly totals.

We present a new model for the generation of synthetic monthly rainfall data which we demonstrate for Parafield in Adelaide, South Australia. The rainfall for each month of the year is modelled as a non-negative random variable from a mixed distribution with either a zero outcome or a strictly positive outcome. We use maximum likelihood to find parameters for both the probability of a zero outcome and the gamma distribution that best matches the observed probability density for the strictly positive outcomes. We describe a new model that generates correlated monthly rainfall totals using a diagonal band copula with a single parameter to generate lag-1 correlated random numbers. Our model preserves the marginal monthly distributions and hence also preserves the monthly and yearly means. We show that for the particular example of Parafield the correlation between rainfall totals for successive months is not significant and so it is reasonable to assume independence. This is however not true for daily rainfall. The correlation between rainfall on successive days is certainly small but it is reasonable as suggested by Katz and Parlange (1998) to conclude that the assumption of independence is the primary reason for the low simulated monthly standard deviations. We describe a new model that generates correlated daily rainfall totals using a diagonal band copula with a single parameter to generate lag-1 correlated random numbers.

The City of Salisbury supplies recycled storm-water to local businesses on a commercial basis and it is important that they understand the full implications of the likely distribution of rainfall and the consequent impact on their ability to manage the capture, treatment and supply to consumers of recycled water. Recent work by Piantadosi (2004) used Matrix Analytic Methods to model the capture, storage and practical management of urban storm-water with stochastic input and regular demand. Howlett *et al.* (2005) describe subsequent research on the application of these methods to the optimal control of storm-water management systems. The practical value of this work has been the development by Piantadosi (2004) of computer simulations that model a sequence of small capture and storage dams on a suburban storm-water course. Because the real systems are complicated by non-essential physical features and fixed infrastructure the most effective way to study their behaviour is by constructing and running an accurate computer simulation driven by synthetic daily rainfall data. To underline the importance of the simulated input we note that suburban storage systems are relatively small and will overflow during sustained wet periods and will empty during sustained dry periods. It is important that water managers understand the level of risk associated with such occurrences. These risks are most easily calculated using repeated simulations driven by a realistic stochastic rainfall generator. Historical records of existing rainfall sequences are thought to be not sufficiently comprehensive.

1 INTRODUCTION

Many large-scale ecological and water resources models require daily, monthly and yearly rainfall data as input to the model. As historical data provides only one realisation, stochastically generated rainfall totals are necessary to assess the impact of rainfall variability on water resources systems. Synthetic rainfall data provides alternative realisations that are equally likely but have not necessarily occurred in the past (Srikanthan 2005). As a result the modelling of rainfall records has become well established over the past thirty years. In particular the gamma distribution has been used many times to model rainfall totals on wet days. Valuable general reviews on weather generators are published by Wilks and Wilby (1999) and Srikanthan and McMahon (2001).

The performance of our models is assessed by comparing the average of each characteristic from all the synthetic generations with that of the observed historical data. Different characteristics are important for different applications, for example, the synthetic daily totals should match the observed marginal daily distributions and other key medium term statistics such as the monthly means and standard deviations. Current models generally tend to perform better at the time step on which they are based, for example, a daily synthetic rainfall model will tend to simulate the daily characteristics better than the monthly and yearly characteristics, such as monthly standard deviation, (Chiew *et al.* 2005). To overcome this problem we present a rainfall generator (rGen) model that simulates rainfall at the time step on which they are based and then the models will be cascaded to start with a synthetic yearly total, then generate a synthetic sequence of monthly totals (through selection from a large number of realisations) that match the yearly total, and subsequently perform a similar operation for sequences of daily totals to match the required monthly totals.

The City of Salisbury in South Australia has constructed a diverse network of local storm-water storage systems which are being used to restore degraded wetlands and supplement existing water supplies. Salisbury lies in suburban Adelaide approximately 15 km north of the CBD. Adelaide is the capital city of South Australia. The Parafield system modelled by Piantadosi (2004) is a typical example of a storm-water storage and recycling system within the Salisbury precinct. Since the catchment is relatively small and consists mostly of paved surfaces the response time for run-off from local rainfall is at most a few hours. Thus the monthly intake of storm-water is very strongly linked to the daily rainfall. Our work is part of a broad research program to develop mathematical models that can be used to simulate the operation of such systems. Synthetic data from a suitable daily rainfall model can be used to drive the simulated operation of specific storm-water systems and hence assist in the development of optimal management strategies and risk assessment.

2 METHODS AND STUDY AREA

2.1 The rainfall generator model rGen

We present a rainfall generator (rGen) algorithm that will use yearly rainfall generated using the methods of Dick and Bowden (1973) and a new daily generation model by Piantadosi *et al.* (2007b) together with a similar model to generate synthetic monthly totals. Figure 1 shows a flowchart of the rainfall generator (rGen) algorithm. The algorithm proceeds as follows. Firstly, we estimate the parameters μ_1, σ_1^2, μ_2, σ_2^2 and p of the compound normal distribution to generate a yearly rainfall total. The choice of distribution is explained in Section 2.7. Secondly, a large number of synthetic monthly totals are generated using a 2-parameter gamma distribution. The two parameters, α and β, used to describe the gamma distribution are found using maximum likelihood estimation. Section 2.3 describes a new model for generating monthly rainfall totals. We select a sequence of monthly totals that best match the generated yearly total. The monthly rainfall totals preserve the monthly characteristics and sum to obtain the yearly total. Finally, we generate sequences of synthetic daily rainfall totals using a model developed by Piantadosi *et al.* (2007b). The rainfall for each day of the year is modelled as a non-negative random variable from a mixed distribution with either a zero outcome or a strictly positive outcome. We use maximum likelihood to find parameters for both the probability of a zero outcome and the gamma distribution that best matches the observed probability density for the strictly positive outcomes. We generate sequences of daily rainfall totals to match the required monthly totals. The synthetic daily totals preserve the daily characteristics and sum to obtain the monthly totals.

2.2 Data

Daily rainfall time series data from two locations are used for this case study. We have available 90 years of official daily rainfall records supplied by the Australian Bureau of Meteorology for Parafield, in South Australia, during the period 1901-1990. The rainfall records are measured in millimetres (mm). We use 30 years of daily rainfall records obtained from the Malaysian Meteorological Department for Mesing (1971-2000) to test our models. A summary of the available rainfall data is given in Table 1. The Table shows that the mean annual rainfall in Parafield is

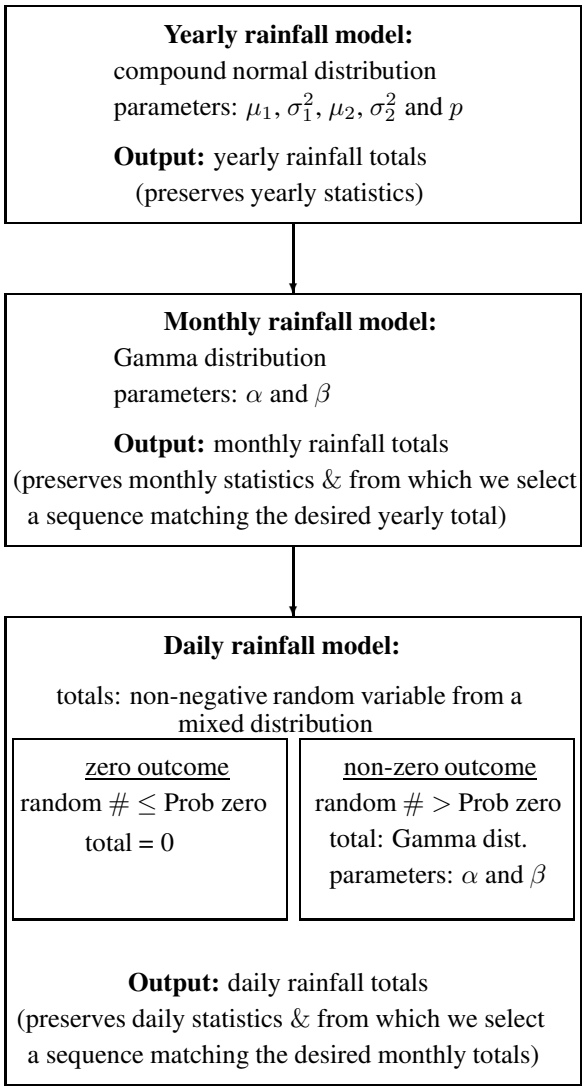

Figure 1. Flowchart of the rainfall generator rGen algorithm.

Table 1. Summary of daily rainfall data.

Location	Record length (years)	Mean annual rainfall (mm)
Parafield	90 (1901-1990)	495.46
Mesing	30 (1971-2000)	2674.94

approximately 495 mm compared to the mean annual rainfall for Mesing, approximately 2675 mm. Figure 2 shows a time series plot of the yearly totals at Parafield from 1901-1990. The plot shows a very slight decrease in yearly rainfall totals but this is not significant (with a P-value 0.793) and will not be considered in our models. Figure 3 shows the monthly average rainfall for Parafield. It can be seen that February has the lowest average monthly rainfall of 19 mm and June has the highest of 66 mm. Table 2 shows that June has the highest variability. From Figure 4, taken from Australian Bureau of Meteorology, we see the average daily rainfall for each month.

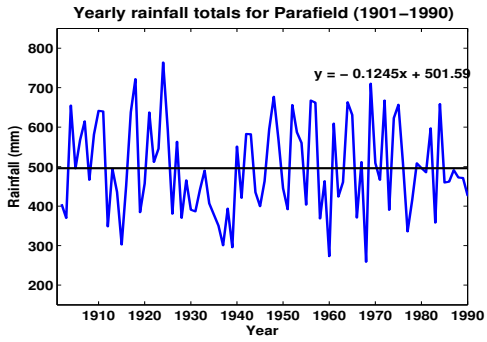

Figure 2. Time series of yearly rainfall at Parafield.

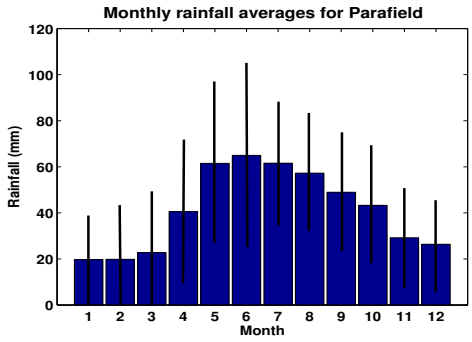

Figure 3. Monthly averages of rainfall data for Parafield; the heights of the bars are the mean monthly totals and the lines represent two standard deviations of monthly totals.

2.3 Modelling Monthly Rainfall

The Gamma distribution has been widely used for describing rainfall data, Stern and Coe (1984), Katz and Parlange (1998), Wilks and Wilby (1999), and Rosenberg *et al.* (2004). Wilks and Wilby (1999) suggest the reason for this is due to the flexible representation involving only two parameters. An extended model is explained for situations when the probability of a zero monthly total is non-zero. The monthly rainfall totals for Parafield will be modelled as non-negative random variables. For each month t of the year there is a chance that no rain will fall and so it is necessary to consider a model that allows positive probability for a zero total. To begin we divide the data into two groups; zero records and non-zero records. The probability

$$p_0 = p_0[t] = P[X[t] = 0]$$

of a zero total on month t is estimated by

$$p_0 = \frac{k}{n}$$

where we use $k = k[t]$ to denote the number of zero rainfall records and n to denote the total number of records. The gamma distribution is used to model the strictly positive component of the monthly rainfall.

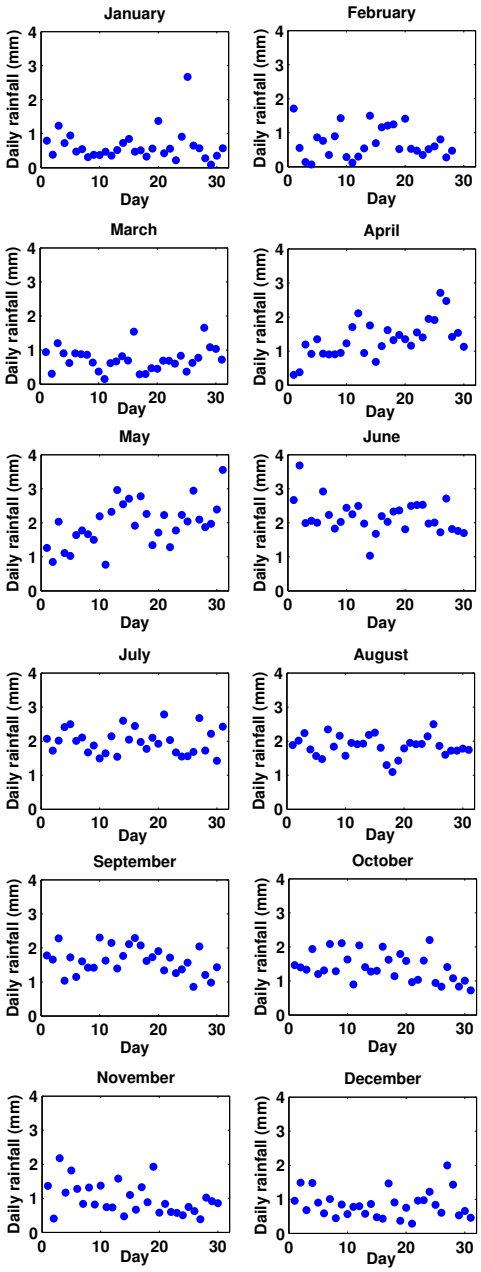

Figure 4. Daily averages of rainfall for Parafield.

The gamma distribution is defined on $(0, \infty)$ by the density function

$$p[\alpha, \beta](x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

where $\alpha > 0$ and $\beta > 0$ are parameters. The parameters $\alpha = \alpha[t]$ and $\beta = \beta[t]$ for month t will be determined from the observed non-zero records by the method of maximum likelihood. The general distribution of rainfall on the interval $[0, \infty)$ for month t can now be modelled with a cumulative distribution function (CDF) in the form

$$F[p_0, \alpha, \beta](x) = p_0 + (1 - p_0) \int_0^x p[\alpha, \beta](\xi) d\xi.$$

Let $x_a = x_a[t]$ and $x_g = x_g[t]$ be the arithmetic and geometric means of the observed non-zero values

Table 2. Monthly means and standard deviations for Parafield

	mean	standard deviation
January	19.68	19.88
February	19.83	23.24
March	22.74	25.40
April	40.54	31.83
May	61.49	35.22
June	64.96	40.19
July	61.58	26.78
August	57.27	25.89
September	48.97	25.50
October	43.29	25.92
November	29.23	21.86
December	26.37	20.09

$x_i = x_i[t]$ for month t in which case the maximum likelihood equations can be written in the form

$$\psi(\alpha) - \log_e(\alpha) + \log_e\left(\frac{x_a}{x_g}\right) = 0 \qquad (1)$$

and

$$\alpha\beta = x_a \qquad (2)$$

where the digamma function ψ is defined by the formula

$$\psi(\alpha) \stackrel{\triangle}{=} \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

for each $\alpha > 0$. The above equations can be solved easily using MATLAB or EXCEL.

2.4 Estimating the key parameters

Standard MATLAB functions were used to calculate the maximum likelihood estimates α and β for the years of monthly rainfall records at Parafield and Mesing. Guenni *et al.* (1996) present a statistical methodology to estimate the parameters of a rainfall model. The α and β values found using maximum likelihood estimation for Parafield and Mesing are shown in Table 3.

2.5 Synthetic rainfall generated by independent random numbers

Spearman's correlation coefficient is calculated for the 90 years of rainfall records at Parafield, to test if any two months are correlated. Table 4 shows the P-values for the Spearman's correlation between successive months (Dec, Jan), (Jan, Feb) and so on. The results suggest it is reasonable to assume independence, at the $\alpha = 0.05$ significance level, as only September is significant. We obtain similar results for Mesing. In our monthly model, we generate synthetic monthly totals using independent random variables. This is not the case for daily rainfall totals. The assumption

Table 3. Parameter estimates for monthly rainfall records at Parafield and Mesing.

	Parafield		Mesing	
Month	α	β	α	β
Jan	1.26	16.73	1.03	297.96
Feb	0.88	24.35	0.96	123.98
Mar	0.86	28.38	1.38	95.79
Apr	1.28	31.61	2.20	60.56
May	2.33	26.43	5.72	23.67
Jun	2.18	29.75	6.05	24.78
Jul	4.91	12.53	9.03	16.89
Aug	4.26	13.45	8.65	20.98
Sep	3.22	15.20	2.95	53.37
Oct	2.21	19.59	6.03	30.12
Nov	1.69	17.92	5.60	68.06
Dec	1.50	18.02	2.53	255.77

of independence for daily rainfall totals is incorrect so a modified model is described in Subsection 2.6 to generate correlated daily rainfall. To generate a sequence $\{x[t]\}$ of synthetic monthly rainfall we first generate realisations $\{r[t]\}$ of a sequence $\{R[t]\}$ of independent random numbers, each one uniformly distributed on the unit interval $[0, 1]$, and then use standard MATLAB functions to solve the equation

$$F[p_0[t], \alpha[t], \beta[t]](x) = r[t]$$

to find the corresponding monthly rainfall denoted by $x = x[t]$. If $r[t] < p_0[t]$ then $x[t] = 0$. The parameters $p_0[t]$, $\alpha[t]$ and $\beta[t]$ are defined by the maximum likelihood estimates from the observed monthly data.

Table 4. Spearman's correlation between successive months at Parafield

Month	P-value	Month	P-value
Jan	0.177	Jul	0.143
Feb	0.337	Aug	0.070
Mar	0.711	Sep	0.044
Apr	0.330	Oct	0.605
May	0.861	Nov	0.218
Jun	0.269	Dec	0.227

Figure 5 shows a histogram of the observed monthly totals versus the generated monthly totals using the 2-parameter gamma distribution for January and July. Similar results were obtained for Mesing.

2.6 Modelling Daily Rainfall

2.6.1 Synthetic rainfall generated by correlated random numbers using a single parameter

Piantadosi *et al.* (2007b) show that synthetic rainfall generated by independent daily random variables

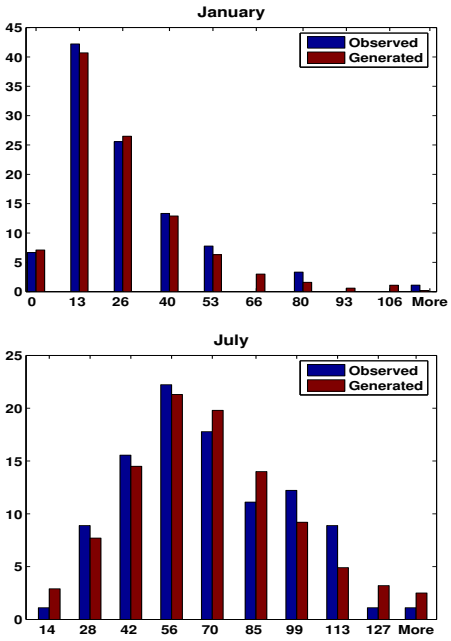

Figure 5. Histogram of observed versus generated monthly totals for January and July.

on successive days significantly underestimates the observed monthly standard deviations. The correlation between rainfall on successive days is certainly small but it is reasonable as suggested by Katz and Parlange (1998) to conclude that the assumption of independence is the primary reason for the low simulated monthly standard deviations. Piantadosi *et al.* (2007a) developed an extended model using a doubly stochastic matrix to define a copula that preserves the given marginal distributions and matches a known grade correlation coefficient so that the entropy of the doubly stochastic matrix is maximized. Cooke and Waij (1986) and Lewandowski (2005) describe a joint distribution of probability mass on the unit square $[0, 1] \times [0, 1]$, which we shall describe generically as a diagonal band copula. Copulas represent a useful approach to modelling of dependent random variables while preserving the known marginal distributions (Nelson 1999). The diagonal band copula provides a simple formula that allows the joint probability to concentrate along one or other of the diagonals while preserving a uniform marginal distribution in each individual variable. This type of copula can be used to specify a degree of positive or negative correlation between the two variables simply by specifying the width and direction of the diagonal band and at the same time the known marginal distributions can be preserved. Piantadosi *et al.* (2007b) show how this idea can be used to correlate the rainfall on different days while still allowing the non-zero totals to be modelled by a time-varying gamma distribution. We show that the generated synthetic daily data matches the observed daily and monthly statistics.

2.7 Modelling Yearly Rainfall

Preliminary data analysis indicated that yearly totals may be well described by a mixture of two normal distributions, one for dry years and one for wet years. We are concerned with the estimation of the parameters μ_1, σ_1^2, μ_2, σ_2^2 and p of the compound normal distribution with density

$$f(y) = pf_1(y) + (1-p)f_2(y)$$

where

$$f_i(y) = (2\pi\sigma_i^2)^{-1/2} \exp\left[-(y-\mu_i)^2/2\sigma_i^2\right] \quad i = 1, 2$$

when additional information is available from either $f_1(y)$ or $f_2(y)$. Suppose that n_1 observations are taken from the normal density $f_1(y)$, say, y_j, $j = 1, 2, \ldots, n_2$. Newton's method is a process yielding solutions to the likelihood function $L(\theta)$ where $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)$. Estimation by the method of moments was used to obtain trial values for the Newton iteration procedure when independent sample information is available from one of the populations. We refer the reader to Dick and Bowden (1973) for further details. The estimated parameters for Parafield are $\mu_1 = 424.30$, $\sigma_1^2 = 4551.30$, $\mu_2 = 609.14$, $\sigma_2^2 = 7342.45$ and $p = 0.398$. Figure 6 shows a histogram of the observed yearly totals versus the generated yearly totals using a compound normal distribution.

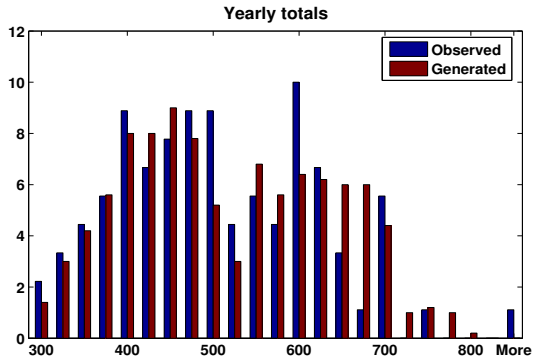

Figure 6. Histogram of observed yearly totals versus the generated yearly totals using compound normal distribution.

3 RESULTS AND DISCUSSION

We compare the performance of our models for generating monthly totals using data from two locations, Parafield rainfall data and Mesing rainfall data. The Kolmogorov-Smirnov statistic D is a particularly simple measure. It is defined as the maximum value of the absolute difference between two cumulative distribution functions (the absolute value of the area between them). Thus, for comparing one data set $S_N(x)$ to a known cumulative distribution function $P(x)$, the $K-S$ statistic is

$$D = \max_{\infty < x < \infty} |S_N(x) - P(x)|$$

where N is the number of data points. We refer the reader to Press *et al.* (1992) for further details. Table 5 shows that the K-S Test D (greatest distance between the two cumulative distributions) is not significant for each month at Parafield and Mesing. We can see that D for each month is less than the critical value 0.143 for Parafield and 0.248 for Mesing. If the K-S Test D is greater than the critical value the difference is significant. We generate any number of

Table 5. Kolmogorov-Smirnov statistic D of each month for Parafield and Mesing respectively.

	Parafield K-S Test D	Mesing K-S Test D
January	0.083	0.157
February	0.133	0.088
March	0.091	0.076
April	0.062	0.169
May	0.088	0.106
June	0.081	0.100
July	0.058	0.136
August	0.075	0.147
September	0.068	0.153
October	0.087	0.113
November	0.063	0.099
December	0.052	0.090

synthetic yearly totals. For each year, a large number of monthly sequences are generated and summed. The total that best matches the yearly total is chosen as the synthetic monthly sequence. Figure 7 shows the generation of two different monthly realisations which match the yearly total 478 mm. It is shown that the generated monthly realisations with the same yearly total are quite different. The methodology for generating monthly totals preserves the monthly statistics and matches the yearly total.

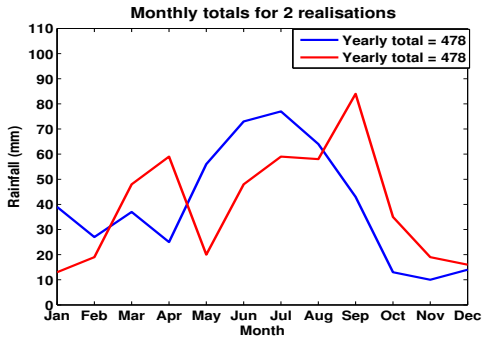

Figure 7. Generated monthly realisations matched to a generated yearly total.

A similar methodology is used to generate daily rainfall totals that match the monthly totals. For each month, a large number of daily rainfall totals are generated and summed. The daily totals that best match the monthly total is selected as the synthetic daily realisation. Figure 8 shows the generation of two different daily realisations which match the monthly totals for February (19 mm) and July (59 mm). Once again we can see the generated realisations are quite different although the monthly totals are the same. This methodology for generating daily rainfall totals preserves the daily statistics and matches the monthly totals. Figures 7 and 8 illustrate the principle under

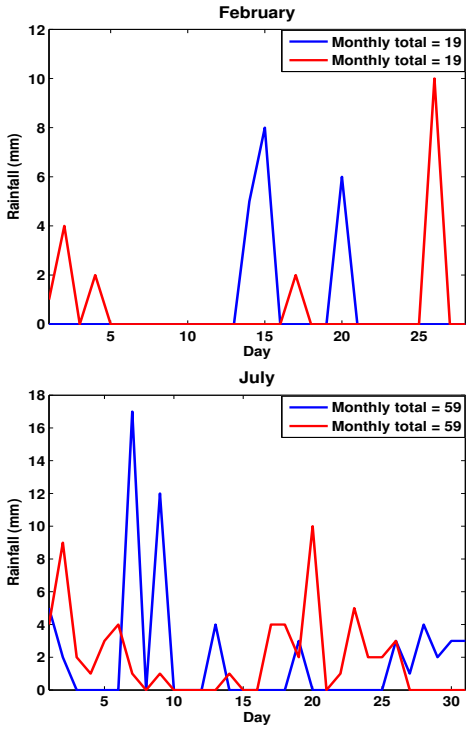

Figure 8. Generated daily realisations matched to a generated monthly total (February and July) .

which we base our procedure. Any sequence of monthly totals that give the same yearly total are equally likely. Thus we can select any of these realisations as a possible sequence for the synthetic year. The same principle is utilised for selecting a sequence of daily totals to match a pre-determined monthly total. In this way, we ensure that the long-term statistics are preserved at all three time scales.

4 REFERENCES

Australian Bureau of Meteorology (2004), Bureau of Meteorology http://www.bom.gov.au.

Chiew, F.H.S., R. Srikanthan, A.J. Frost and E.G.I Payne (2005), Reliability of daily and annual stochastic rainfall data generated from different data lengths and data characteristics, *MODSIM2005*, Melbourne, Dec. 2005, 1223–1229.

Cooke, R. M. and R. Waij (1986), Monte Carlo sampling for generalized knowledge dependence with application to human reliability, *Risk Analysis*, 6, 335–343.

Dick, N.P. and D.C. Bowden (1973), Maximum likelihood estimation for mixtures of two normal distributions, *Biometrics*, 29, 781–790.

Guenni, L., M.F. Hutchinson, W. Hogarth, C.W. Rose and R. Braddock (1996), A model for seasonal variation of rainfall at Adelaide and Turen, *Ecological Modelling*, 85, 203–217.

Howlett, P.G., J. Piantadosi, and C.E.M. Pearce (2005), Analysis of a practical control policy for water storage in two connected dams, *Cont. opt.*. V. Jeyakumar and A. Rubinov (eds.), Springer-Verlag, NY, 99, 433–450.

Katz, R.W. and M.B. Parlange (1998), Overdispersion phenomenon in stochastic modelling of precipitation, *J. Climate*, 11, 591–601.

Lewandowski, D. (2005), Generalized diagonal band copulas, *Insurance Math. and Econ.*, 37, 49–67.

Malaysian Meteorological Department (2007), Jabatan Perkhidmatan Kajicuaca Malaysia.

Nelson, R.D. (1999), An introduction to copulas, Springer Lecture Notes in Statistics, New York.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vettering (1992), Numerical recipes in C: The art of scientific computing, Cambridge Press, NY.

Piantadosi, J. (2004), Optimal policies for storage of urban stormwater, PhD Thesis, University of South Australia.

Piantadosi, J., P.G. Howlett and J.W. Boland (2007a), Matching the grade correlation coefficient using a copula with maximum disorder, *J.I.M.O.*, 3(2), 305–312.

Piantadosi, J., P.G. Howlett and J.W. Boland (2007b), A new model for correlated daily rainfall, (in preparation).

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

Srikanthan, R. (2005), Stochastic Generation of Daily Rainfall Data, *MODSIM2005*, Melbourne, 12-15 Dec. 2005, 1915–1921.

Srikanthan, R. and T. A. McMahon (2001), Stochastic generation of annual, monthly and daily climate data: A review, *Hydr. and Earth Sys. Sci.*, 5(4), 633–670.

Stern, R. D. and R. Coe (1984), A model fitting analysis of daily rainfall, *J. Roy. Statist. Soc. A*, 147, Part 1, 1–34.

Wilks, D.S. and R.L.Wilby (1999), The weather generation game: a review of stochastic weather models, *Prog. Phys. Geog.*, 23(3), 329–357.