19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Generation of simulated rainfall data at different time-scales

Julia Piantadosi [a], **Phil Howlett** [b], **Jonathan Borwein** [c], **John Henstridge** [d]

[a] *Research Fellow, Centre for Industrial and Applied Mathematics, Scheduling and Control Group, University of South Australia, Mawson Lakes, SA 5095.*

[b] *Emeritus Professor, Industrial and Applied Mathematics, Scheduling and Control Group, University of South Australia, Mawson Lakes, SA 5095.*

[c] *Laureate Professor and Director Centre for Computer Assisted Research Mathematics and its Applications (CARMA), University of Newcastle, Callaghan, NSW 2308, Australia. Distinguished Professor, King Abdulaziz University, Jeddah 80200, Saudi Arabia.*

[d] *Managing Director and Principal Consultant Statistician, Data Analysis Australia Pty Ltd, Adjunct Professor of Statistics, University of Western Australia, 97 Broadway, Nedlands WA, 6009.*

Email: Julia.Piantadosi@unisa.edu.au

Abstract: We desire to generate monthly rainfall totals for a particular location in such a way that the statistics for the simulated data match the statistics for the observed data. We are especially interested in the accumulated rainfall totals over several months. We propose two different ways to construct a joint rainfall probability distribution that matches the observed grade correlation coefficients and preserves the known marginal distributions. Both methods use multi-dimensional checkerboard copulas. In the first case we construct a copula of maximum entropy and in the second case we use a copula derived from a multi-variate normal distribution. Finally we simulate monthly rainfall totals at a particular location and compare the results.

Keywords: Copulas, Maximum entropy, multi-variate normal distribution, grade correlation

1 MODELLING ACCUMULATED RAINFALL

It has been usual to model both short-term and long-term rainfall accumulations at a specific location by a gamma distribution (Wilks and Wilby, 2011; Srikanthan and McMahon, 2004; Fowler et al., 2005; Hasan and Dunn, 2010). Some authors (Withers and Nadarajah, 2011; Katz and Parlange, 1998) have, however, observed that simulations in which monthly rainfall totals are modelled as mutually independent gamma random variables generate accumulated bi-monthly, quarterly and yearly totals with much lower variance than the observed accumulations. We surmise that the variance of the generated totals will be increased if the model includes an appropriate level of positive correlation between individual monthly totals. More generally, the problem we address is *how to construct a joint probability distribution which preserves the known marginal distributions and matches the observed grade correlation coefficients*. We propose two alternative methods using multi-dimensional copulas.

1.1 Multi-dimensional copulas

An m-dimensional *copula* where $m \geq 2$, is a continuous, m-increasing cumulative probability distribution $C : [0,1]^m \mapsto [0,1]$ on the unit m-dimensional hyper-cube with uniform marginal probability distributions. If $F_r : \mathbb{R} \mapsto [0,1]$ is a prescribed continuous distribution for the real valued random variable X_r for each $r = 1, \ldots, m$ then the function $G : \mathbb{R}^m \mapsto [0,1]$ defined by

$$G(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_m(x_m))$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$ is a joint probability distribution for the vector-valued random variable $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ with the marginal distribution for X_r defined by F_r for each $r = 1, \ldots, m$. The joint density $g : \mathbb{R}^m \mapsto [0, \infty)$ is defined almost everywhere and is given by the formula

$$g(\boldsymbol{x}) = c(F_1(x_1), \ldots, F_m(x_m)) f_1(x_1) \cdots f_m(x_m)$$

where $c : [0,1]^m \mapsto [0, \infty)$ is the density for the joint distribution defined by C and where $f_r : \mathbb{R} \mapsto [0, \infty)$ for each $r = 1, \ldots, m$ are the densities for the prescribed marginal distributions. If we define $U_r = F_r(X_r)$ for each $r = 1, \ldots, m$ then each U_r is uniformly distributed on $[0,1]$ and the copula C describes the distribution of the vector valued random variable $\boldsymbol{U} = (U_1, \ldots, U_m)^T$. The *grade correlation coefficients* for $\boldsymbol{X}$ are defined by

$$\begin{aligned} \rho_{r,s} &= \frac{E[(F_r(X_r) - 1/2)(F_s(X_s) - 1/2)]}{\sqrt{E[(F_r(X_r) - 1/2)^2] \cdot E[(F_s(X_s) - 1/2)^2]}} \\ &= \frac{E[(U_r - 1/2)(U_s - 1/2)]}{\sqrt{E[(U_r - 1/2)^2] \cdot E[(U_s - 1/2)^2]}} \\ &= 12E[U_r U_s] - 3 \end{aligned}$$

for each $1 \leq r < s \leq m$. The grade correlation coefficients for $\boldsymbol{X}$ are simply the correlations for $\boldsymbol{U}$. The *entropy* for the copula C is defined by

$$J(C) = (-1) \int_{[0,1]^m} c(\boldsymbol{u}) \log_e c(\boldsymbol{u}) \, d\boldsymbol{u}$$

where $\boldsymbol{u} = (u_1, \ldots, u_m)^T \in [0,1]^m$. The entropy $J(C)$ measures the inherent disorder of the distribution. The most disordered copula is the one with $c(\boldsymbol{u}) = 1$ for all $\boldsymbol{u} \in [0,1]^m$ for which $J(C) = 0$.

We introduce two special copulas which we will use to model monthly rainfall. For the first method we use the checkerboard copula of maximum entropy proposed by Piantadosi et al. (2007, 2010). For the second method we use a copula defined by a multi-variate normal distribution.

1.2 Checkerboard copulas

An m-dimensional *checkerboard* copula is a distribution with a corresponding density defined almost everywhere by a step function on an m-uniform subdivision of the hyper-cube $[0,1]^m$. Any continuous copula can be uniformly approximated by a checkerboard copula.

Let $n \in \mathbb{N}$ be a natural number and let $\boldsymbol{h}$ be a non-negative m–dimensional hyper–matrix given by $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ where $\ell = n^m$ and $\boldsymbol{i} \in \{1, \ldots, n\}^m$ with $h_{\boldsymbol{i}} \in [0, 1]$. Define the *marginal sums* $\sigma_r : \{1, \ldots, n\} \mapsto \mathbb{R}$ by the formulae

$$\sigma_r(i_r) = \sum_{j \neq r, i_j \in \{1,2,\ldots,n\}} h_{\boldsymbol{i}}$$

for each $r = 1, 2, \ldots, m$. If $\sigma_r(i_r) = 1$ for all $i_r \in \{1, 2, \ldots, n\}$ and all $r = 1, 2, \ldots, m$ then we say that $\boldsymbol{h}$ is *multiply stochastic*. Define the partition $0 = a(1) < a(2) < \cdots < a(n) < a(n+1) = 1$ of the interval $[0, 1]$ by setting $a(k) = (k-1)/n$ for each $k = 1, \ldots, n+1$ and define a step function $c_{\boldsymbol{h}} : [0, 1]^m \mapsto \mathbb{R}$ almost everywhere by the formula

$$c_{\boldsymbol{h}}(\boldsymbol{u}) = n^{m-1} \cdot h_{\boldsymbol{i}} \quad \text{if} \quad \boldsymbol{u} \in I_{\boldsymbol{i}} = \bigtimes_{r=1}^{m} (a(i_r), a(i_r + 1))$$

for each $\boldsymbol{i} = (i_1, \ldots, i_m) \in \{1, 2, \ldots, n\}^m$. Now the step function $c_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, \infty)$ is the density for a corresponding checkerboard copula $C_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, 1]$.

1.3 Copulas of maximum entropy

If $n \in \mathbb{N}$ is sufficiently large then Piantadosi et al. (2007, 2010) showed that $\boldsymbol{h}$ could be chosen in such a way that the known grade correlations were imposed and the entropy of the hyper-matrix was maximized. The corresponding checkerboard copula $C_{\boldsymbol{h}}$ is the most *disordered* or *least prescriptive* such copula and is defined by the hyper-matrix $\boldsymbol{h}$ that solves the following problem.

Problem 1 (The primal problem). *Find the hyper-matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ to maximize the entropy*

$$J(\boldsymbol{h}) = (-1) \left[\frac{1}{n} \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1) \log_e n \right] \tag{1.1}$$

subject to the constraints

$$\sum_{j \neq r,\, i_j \in \{1,\ldots,n\}} h_{\boldsymbol{i}} = 1 \tag{1.2}$$

for all $i_r \in \{1, \ldots, n\}$ and each $r = 1, \ldots, m$ and

$$h_{\boldsymbol{i}} \geq 0 \tag{1.3}$$

for all $\boldsymbol{i} \in \{1, \ldots, n\}^m$ and the additional grade correlation coefficient constraints

$$12 \left[\frac{1}{n^3} \cdot \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} (i_r - 1/2)(i_s - 1/2) \right] - 3 = \rho_{r,s} \tag{1.4}$$

for $1 \leq r < s \leq m$ where $\rho_{r,s}$ is known for all $1 \leq r < s \leq m$.

Piantadosi et al. (2010) noted that the problem is well posed. Nevertheless, it is not easy to compute a numerical solution directly. In fact it is much easier to solve the problem using the theory of Fenchel duality. We refer to the paper by Piantadosi et al. (2010) for details of the solution procedure.

1.4 Multi-variate normal copulas

The m-dimensional normal distribution $\varphi : \mathbb{R}^m \to [0, \infty)$ for the vector-valued random variable $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^T \in \mathbb{R}^m$ with unit normal marginal distributions is defined by the density

$$\varphi(\boldsymbol{z}) = \frac{1}{(2\pi)^{m/2} (\det \Sigma)^{1/2}} \exp \left[-\frac{1}{2} \boldsymbol{z}^T \Sigma^{-1} \boldsymbol{z} \right]$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)^T \in \mathbb{R}^m$ and where $\Sigma = E[\boldsymbol{Z}\boldsymbol{Z}^T] = [\cos\theta_{r,s}] \in [-1,1]^{m\times m}$ is the correlation matrix. The marginal distributions for Z_r are standard unit normal distributions (Wang, 2006) given by

$$\Phi(z_r) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{z_r} \exp\left[-\frac{\zeta_r^2}{2}\right] d\zeta_r.$$

If we define $U_r = \Phi(Z_r)$ for each $r = 1, 2, \ldots, m$ then the random variables U_r are uniformly distributed on the interval $[0,1]$ and the joint density $c : [0,1]^m \to [0,\infty)$ defined by

$$c(\boldsymbol{u}) = \frac{\varphi(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m))}{\Phi'(\Phi^{-1}(u_1)) \cdots \Phi'(\Phi^{-1}(u_m))}$$

is the density for a corresponding m-dimensional *normal copula* $C : [0,1]^m \to [0,1]$. We note from Wang (2006) that for any $1 \le r < s \le m$ the marginal distribution of $(Z_r, Z_s)^T$ is a bi-variate normal distribution with correlation defined by the relevant sub-matrix of Σ. To find $\theta = \theta_{r,s}$ we solve the equation $f(\theta) = \rho_{r,s}$ where

$$f(\theta) = \frac{6}{\pi \sin\theta} \int_{\mathbb{R}^2} \Phi(z_r)\Phi(z_s) \exp\left[-\frac{1}{2\sin^2\theta}\left(z_r^2 - 2\cos\theta z_r z_s + z_s^2\right)\right] dz_r dz_s - 3$$

and where $\rho_{r,s}$ is the desired grade correlation coefficient for each $1 \le r < s \le m$.

For the purpose of simulation and to enable a direct comparison with the previous method we have preferred to approximate the multi-variate normal copula by a checkerboard copula defined by a hyper-matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ determined by the formula

$$h_{\boldsymbol{i}} = n \int_{I_{\boldsymbol{i}}} c(\boldsymbol{u}) d\boldsymbol{u}$$

for each $\boldsymbol{i} \in \{1, \ldots, n\}^m$. Finally the step function $c_{\boldsymbol{h}} : [0,1]^m \mapsto \mathbb{R}$ and the corresponding copula $C_{\boldsymbol{h}} : [0,1]^m \mapsto [0,1]$ are defined in the usual manner. We refer to this copula as a *normal checkerboard copula*. We use the formula (1.4) to calculate the grade correlation coefficients and choose $\theta_{r,s}$ to match the observed values $\rho_{r,s}$ for each $1 \le r < s \le m$. These calculations can be done separately for each $1 \le r < s \le m$ using the relevant marginal bi-variate normal copula.

2 MONTHLY RAINFALL DATA FOR SYDNEY

We used official monthly rainfall records supplied by the Australian Bureau of Meteorology for Sydney, NSW, Australia, for the 150 year period 1859-2008. Table 1 shows the monthly statistics. The rainfall is measured in millimetres (mm).

Table 1. Monthly means (m) and standard deviations (s) for Sydney

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
m	103	118	130	126	103	131	98	82	70	77	84	78
s	76	110	103	112	111	116	82	84	60	66	76	63

Table 2 shows the grade correlation coefficients for all monthly pairs. The distributions appear to be weakly correlated. The correlation for (Oct,Nov) is significant at the 0.01 level (2-tailed) and the correlations for (Jan,Feb), (Jan,Apr), (Jan,Oct), (Mar,Jun), (Apr,May), (Jun,Sep) are significant at the 0.05 level (2-tailed). The significant correlations are shown in bold print.

2.1 Modelling individual monthly rainfall totals for Sydney

There are no observed zero rainfall totals and the distributions for individual months can be modelled effectively using a gamma distribution (Katz and Parlange, 1998; Piantadosi et al., 2009; Rosenberg

Table 2. Grade correlation coefficients for all monthly pairs

	Ja	Fe	Mr	Ap	Ma	Jn	Jl	Au	Se	Oc	No	De
Ja		**.18**	-.06	**-.19**	-.01	-.02	-.02	.13	.09	**-.16**	.05	-.04
Fe	**.18**		-.03	-.08	-.09	.05	-.01	.10	.09	-.05	.08	-.07
Mr	-.06	-.03		.11	.04	**.19**	-.14	-.15	-.12	.15	-.05	-.01
Ap	**-.19**	-.08	.11		**.18**	.05	.13	.12	-.08	.11	.09	-.03
Ma	-.01	-.09	.04	**.18**		.05	-.02	-.05	-.08	-.07	.05	-.06
Jn	-.02	.05	**.19**	.05	.05		-.04	-.07	**-.17**	.02	.05	-.05
Jl	-.02	-.01	-.14	.13	-.02	-.04		.11	.12	.08	-.08	-.02
Au	.13	.10	-.15	.12	-.05	-.07	.11		.13	.13	.12	-.09
Se	.09	.09	-.12	-.08	-.08	**-.17**	.12	.13		.04	.07	-.01
Oc	**-.16**	-.05	.15	.11	-.07	.02	.08	.13	.04		**.22**	-.03
No	.05	.08	-.05	.09	.05	.05	-.08	.12	.07	**.22**		.08
De	-.04	-.07	-.01	-.03	-.06	-.05	-.02	-.09	-.01	-.03	.08	

et al., 2004; Srikanthan and McMahon, 2004; Stern and Coe, 1984; Wilks and Wilby, 2011). The gamma distribution is defined on $(0, \infty)$ by the formula

$$F[\alpha, \beta](x) = \int_0^x \frac{\xi^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp(-\xi/\beta)\, d\xi$$

where $\alpha > 0$ and $\beta > 0$ are parameters. The parameters $\alpha = \alpha[t]$ and $\beta = \beta[t]$ for month t were determined by the method of maximum likelihood. The calculated values are

$$\begin{aligned} \boldsymbol{\alpha} &= (1.817,\ 1.359,\ 1.741,\ 1.333,\ 1.258,\ 1.338,\ 1.202,\ 1.051,\ 1.412,\ 1.468,\ 1.461,\ 1.777) \\ \boldsymbol{\beta} &= (56.40,\ 86.75,\ 74.60,\ 94.70,\ 95.97,\ 97.64,\ 81.56,\ 78.12,\ 49.33,\ 52.31,\ 57.29,\ 43.92). \end{aligned}$$

We consider further the months of October and November where the grade correlation is the most significant. When we generate simulated rainfall independently from the respective gamma distributions the variance of the total is much less than the observed variance. The details are shown in Table 3. The observed value of $\rho \approx 0.22$ strongly suggests we should seek a model using a correlated joint distribution.

2.2 Simulating the total rainfall using a bi-variate checkerboard copula

Suppose we have obtained a checkerboard copula $C_{\boldsymbol{h}}$ defined by a matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ where $\ell = n^2$ and $\boldsymbol{i} = (i, j) \in \{1, \ldots, n\}^2$ on a uniform partition $\{I_{\boldsymbol{i}}\}$ of the unit square. Simulated data for monthly rainfall pairs may be generated as follows. Define an order for the indices $\boldsymbol{i} = (i, j)$ by saying that $(i, j) \prec (p, q)$ if $i \le p$ and $j < q$. For each pseudo-random number $r \in (0, 1)$ select the interval $I_{pq} = (a(p), a(p+1)) \times (a(q), a(q+1))$ if

$$\sum_{(i,j)\prec(p,q)} h_{ij} \;<\; nr \;<\; \left[\sum_{(i,j)\prec(p,q)} h_{ij}\right] + h_{pq}.$$

Once the interval I_{pq} has been selected the precise position of the pseudo-random point $(u_r, v_r) \in I_{pq}$ and the corresponding rainfall pair is fixed by generating two more (independent) pseudo-random numbers $(s_r, t_r) \in (0, 1)^2$ and setting

$$(u_r, v_r) = \left(\frac{(p-1) + s_r}{n}, \frac{(q-1) + t_r}{n}\right) \quad \text{and} \quad (x_r, y_r) = \left(F_x^{-1}(u_r), F_y^{-1}(v_r)\right)$$

where F_x and F_y are the given marginal distributions.

The fitted bi-variate checkerboard copula of maximum entropy. We set $\rho = 0.2169$ and calculate

$$\boldsymbol{h} \approx \begin{bmatrix} 0.35818 & 0.28072 & 0.21037 & 0.15073 \\ 0.28072 & 0.26668 & 0.24223 & 0.21037 \\ 0.21037 & 0.24223 & 0.26668 & 0.28072 \\ 0.15073 & 0.21037 & 0.28072 & 0.35818 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

where $\boldsymbol{h} = [h_{ij}]$. The entropy is given by $J(\boldsymbol{h}) \approx -0.02728$.

The fitted bi-variate normal copula. We choose $\theta = 1.29893$ to give $\rho \approx 0.2169$ and calculate

$$\boldsymbol{h} \approx \begin{bmatrix} 0.36575 & 0.27132 & 0.21489 & 0.14804 \\ 0.27132 & 0.26468 & 0.24911 & 0.21489 \\ 0.21489 & 0.24911 & 0.26468 & 0.27132 \\ 0.14804 & 0.21489 & 0.27132 & 0.36575 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

where $\boldsymbol{h} = [h_{ij}]$. The entropy is given by $J(\boldsymbol{h}) \approx -0.02759$.

Figures 1 (a) and (b) show scatter plots for the synthetic rainfall pairs generated by the respective checkerboard copulas and histograms for the corresponding marginals.

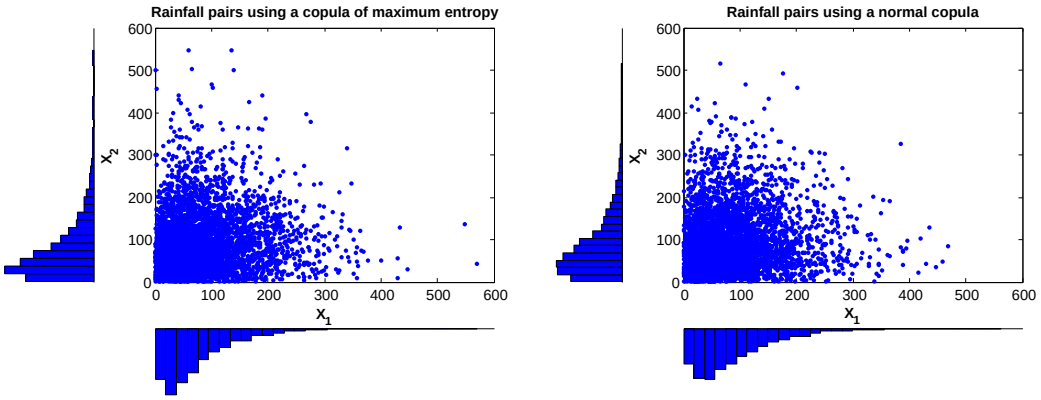

Figure 1. Generated rainfall pairs using (a) a copula of maximum entropy and (b) a normal copula.

2.3 Comparison of the results

The variance of the synthetic rainfall totals generated by the independent model is significantly less than the variance for the observed sums. The totals generated using either the copula of maximum entropy or the normal copula provide a much better match. See Table 3 for details. We apply the Kolmogorov-Smirnov goodness of fit test to analyse the results. The P-values for the copula of maximum entropy and for the normal copula are both greater than 0.05 and so there is no significant difference between the observed and generated totals at the 5% significance level.

Table 3. Comparison of the three different models

	Mean (mm)	Variance
Observed	160.488	10830.299
Independent model	160.451	8732.729
Maximum entropy checkerboard copula	160.900	10742.020
Normal checkerboard copula	161.661	10648.118

3 CONCLUSIONS

Our preliminary work shows that the variance of the generated bi-monthly rainfall totals can be significantly increased by using a copula that allows us to incorporate the observed correlation. We have used two different copulas. The rationale for using a copula of maximum entropy was a desire to avoid unwarranted assumptions about the unobserved statistics. Likewise, the corresponding bi-variate distribution is the most disordered distribution that preserves the prescribed marginal distributions and matches the observed grade correlation.

In comparing the two methods there is little to distinguish them. The normal checkerboard copula turns out to be very close to the maximum entropy checkerboard copula of the same size in all of the two-dimensional examples we considered, irrespective of the number of subdivisions. In two dimensions, the numerical calculations for each method are of similar complexity and can be implemented using standard MATLAB packages. For higher dimensions, the recent paper by Piantadosi et al. (2010) shows that the calculations required for the maximum entropy checkerboard copula are feasible. It would seem that the same should be true for the normal checkerboard copula but preliminary calculations of the relevant probability integrals with the MATLAB package *triplequad* did not yield sufficiently accurate results. More work is still required to find a suitable calculation procedure.

REFERENCES

Fowler, H., C. Kilsby, P. O'Connell, and A. Burton (2005). A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. *J. Hydrol. 308(1-4)*, 50–60.

Hasan, M. M. and P. K. Dunn (2010). Two tweedie distributions that are near optimal for modelling monthly rainfall in australia. *International J Climatology DOI: 10.1002/joc.2162*, 50–60.

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

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

Piantadosi, J., P. Howlett, and J. Boland (2007). Matching the grade correlation coefficient using a copula with maximum disorder. *Journal of Industrial and Management Optimization 3(2)*, 305–312.

Piantadosi, J., P. Howlett, and J. Borwein (2010). Copulas with maximum entropy. *Optimization Letters DOI: 10.1007/s11590-010-0254-2*, 431–438.

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

Srikanthan, R. and T. A. McMahon (2004). Stochastic generation of annual, monthly and daily climate data: A review. *Hydr. and Earth Sys. Sci. 5(4)*, 633–670.

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

Wang, R. (2006). Conditional and marginal of multivariate gaussian, http://fourier.eng .hmc.edu/e161/lectures/gaussianprocess/node5.html. *J. Roy. Statist. Soc. A*, 1–34.

Wilks, D. S. and R. L. Wilby (2011). The weather generation game: a review of stochastic weather models. *Prog. Phys. Geog. 23(3)*, 329–357.

Withers, C. S. and S. Nadarajah (2011). On the compound poisson–gamma distribution. *Kybernetika 47(1)*, 15–37.