Maximum entropy methods for generating simulated rainfall

Julia Piantadosi[*], Phil Howlett[†], Jonathan Borwein[‡], John Henstridge[§]

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 use the theory of Fenchel duality to 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 using each method and analyse the statistical behaviour of the corresponding quarterly accumulations.

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 [16, 11, 3, 4]. Some authors [14, 5] 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. It is reasonable to 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. We use a typical case study to show that this is indeed the case. 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 ways in which this could be done. Both methods use 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

[*]Research Fellow, Centre for Industrial and Applied Mathematics, Scheduling and Control Group, University of South Australia, Mawson Lakes, SA 5095. Email: julia.piantadosi@unisa.edu.au.

[†]Emeritus Professor, Industrial and Applied Mathematics, Scheduling and Control Group, University of South Australia, Pooraka, SA 5095. Email: phil.howlett@unisa.edu.au.

[‡]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. Email: jonathan.borwein@newcastle.edu.au.

[§]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: john@daa.com.au.

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 $r = 1, 2, \ldots, m$. 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 $r = 1, 2, \ldots, m$ are the densities for the prescribed marginal distributions. If related real valued random variables $U_r = F_r(X_r)$ are defined for each $r = 1, 2, \ldots, m$ 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 $1 \leq r < s \leq m$. Thus, the grade correlation coefficients for $\boldsymbol{X}$ are simply the correlations for $\boldsymbol{U}$. The *entropy* for the copula C with density 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 $J(C)$ of the copula 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.* [7, 9]. 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. For each fixed $n \in \mathbb{N}$ we will consider a subdivision of the interval $[0,1]$ into n equal length subintervals and a corresponding m–uniform subdivision of the unit hyper–cube $[0,1]^m$ into n^m congruent hyper–cubes. Consider an elementary checkerboard copula $C = C_{\boldsymbol{h}}$ with density $c = c_{\boldsymbol{h}}$ in the form of a step function defined by an m–dimensional hyper–matrix $\boldsymbol{h}$ such that the density takes a constant non–negative value on each hyper–cube of the subdivision. If $\boldsymbol{h}$ is a non–negative

m–dimensional hyper–matrix given by $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ where $\boldsymbol{i} = (i_1, \ldots, i_m) \in \{1, \ldots, n\}^m$ and $\ell = n^m$ then the grade correlation coefficients for $C_{\boldsymbol{h}}$ are given by

$$\rho_{r,s} = 12 \left[\frac{1}{n^3} \sum_{\boldsymbol{i} \, \in \, \{1,\ldots,n\}^m} h_{\boldsymbol{i}} (i_r - 1/2)(i_s - 1/2) \right] - \, 3 \tag{1.1}$$

and the *entropy* of $\boldsymbol{h}$ is given by

$$J(\boldsymbol{h}) = (-1) \left[\frac{1}{n} \sum_{\boldsymbol{i} \, \in \, \{1,\ldots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1) \log_e n \right]. \tag{1.2}$$

If $n \in \mathbb{N}$ is sufficiently large then Piantadosi *et al.* [7, 9] 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. Since entropy is a measure of disorder the solution proposed by Piantadosi *et al.* for $c_{\boldsymbol{h}}$ can be interpreted as the *most disordered* or *least prescriptive* choice of step function for the selected value of n that satisfies the required grade correlation constraints. The corresponding checkerboard copula $C = C_{\boldsymbol{h}}$ is the most disordered such copula.

1.3 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

$$\Sigma = E[\boldsymbol{Z}\boldsymbol{Z}^T] = [\cos \theta_{r,s}] \in [-1, 1]^{m \times m} \tag{1.3}$$

is the correlation matrix. The Hilbert space interpretation is that $\theta_{r,s}$ represents the angle between the unit vectors representing the random variables Z_r and Z_s. Consequently there are geometric restrictions on the permissible angles [9]. The marginal distributions for Z_r are standard unit normal distributions [13] 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 the m–dimensional *normal copula* $C : [0, 1]^m \to [0, 1]$ defined by

$$C(\boldsymbol{u}) = \int_{[0,u_1] \times \cdots \times [0,u_m]} c(\boldsymbol{v}) d\boldsymbol{v}.$$

We note from [13] that for any $1 \leq r < s \leq m$ the marginal distribution of $(Z_r, Z_s)^T$ is a bi–variate normal distribution with *correlation matrix* $\Sigma_{r,s}$ given by

$$\Sigma_{r,s} = \begin{bmatrix} 1 & \cos\theta_{r,s} \\ \cos\theta_{r,s} & 1 \end{bmatrix} \tag{1.4}$$

and hence the *grade correlation coefficients* for the copula C can be calculated from the formula

$$\rho_{r,s} = \frac{6}{\pi \sin\theta_{r,s}} \int_{\mathbb{R}^2} \Phi(z_r)\Phi(z_s) \exp\left[-\frac{1}{2\sin^2\theta_{r,s}} \left(z_r^2 - 2\cos\theta_{r,s} z_r z_s + z_s^2 \right) \right] dz_r dz_s - 3. \tag{1.5}$$

From [15] the entropy is given by

$$J(C) = \frac{1}{2} \log_e \det \Sigma. \tag{1.6}$$

Since $\det\Sigma$ represents the volume of an m–dimensional parallelepiped defined by unit vectors representing the random variables $Z_1, \ldots, Z_m$ it follows that $0 < \det\Sigma \leq 1$ and hence $J(C) \leq 0$. We will adjust the parameters $\theta_{r,s}$ in order to match the observed grade correlation coefficients. Note also [15] in one dimension for two distributions with the same mean and variance that entropy is maximized by the normal distribution. In our case, with a multi–variate normal distribution and additional constraints on the correlation, it nevertheless seems intuitively reasonable to expect that the normal copula may be close to the copula of maximum entropy.

2 Constructing a copula of maximum entropy

We now outline the method proposed by Piantadosi *et al.* [9] to find a copula of maximum entropy. 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 it follows that the step function $c_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, \infty)$ defines a corresponding copula $C_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, 1]$ by the formula

$$C_{\boldsymbol{h}}(\boldsymbol{u}) = \int_{\bigtimes_{i=1}^n [0, u_i]} c_{\boldsymbol{h}}(\boldsymbol{v}) d\boldsymbol{v}$$

for all $\boldsymbol{u} \in [0, 1]^m$. The formulae (1.1) and (1.2) can be established by direct integration. It is also possible to show that

$$-1 + \frac{1}{n^2} \leq \rho_{r,s} \leq 1 - \frac{1}{n^2}. \tag{2.7}$$

See [9] for more details. The checkerboard *copula of maximum entropy* is the checkerboard copula $C_{\boldsymbol{h}}$ defined by the hyper–matrix $\boldsymbol{h}$ that solves the following problem.

Problem 2.1 (The primal problem) *Find the hyper–matrix* $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ *to maximize the entropy*

$$J(\boldsymbol{h}) = (-1)\left[\frac{1}{n}\sum_{\boldsymbol{i}\,\in\,\{1,\ldots,n\}^m} h_{\boldsymbol{i}}\log_e h_{\boldsymbol{i}} + (m-1)\log_e n\right] \tag{2.8}$$

subject to the constraints

$$\sum_{j\neq r,\, i_j\in\{1,\ldots,n\}} h_{\boldsymbol{i}} = 1 \tag{2.9}$$

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

$$h_{\boldsymbol{i}} \geq 0 \tag{2.10}$$

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

$$12\left[\frac{1}{n^3}\cdot\sum_{\boldsymbol{i}\,\in\,\{1,\ldots,n\}^m} h_{\boldsymbol{i}}(i_r-1/2)(i_s-1/2)\right] - \;3 = \rho_{r,s} \tag{2.11}$$

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

Piantadosi *et al.*[9] 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. To do this it is best to begin by writing the primal problem in standard form. Define $g : \mathbb{R}^{\ell} \mapsto [0,\infty)\cup\{+\infty\}$ by setting

$$g(\boldsymbol{h}) = \begin{cases} (-1)J(\boldsymbol{h}) & \text{if } h_{\boldsymbol{j}} \geq 0 \text{ for all } \boldsymbol{j} \in \{1,2,\ldots,m\}^n \\ +\infty & \text{otherwise} \end{cases}$$

where we have used the convention that $h\log_e h = 0$ when $h = 0$ and where we allow functions to take values in an extended set of real numbers. With appropriate definitions we can write the constraints (2.9) and (2.11) in the form $A\boldsymbol{h} = b$ where $A \in \mathbb{R}^{k\times\ell}$ and $b \in \mathbb{R}^k$ and where k is the collective rank of the coefficient matrix defining the two sets of linear constraints. We can omit constraint (2.10) as it is enforced by the entropy. The primal problem can be restated in standard mathematical form.

Problem 2.2 (Mathematical statement of the primal problem) *Find*

$$\inf_{\boldsymbol{h}\,\in\,\mathbb{R}^{\ell}}\left\{\; g(\boldsymbol{h}) \mid A\boldsymbol{h} = b \;\right\}. \tag{2.12}$$

The *Fenchel conjugate* function $g^* : \mathbb{R}^{\ell} \mapsto \mathbb{R}\cup\{-\infty\}$ is defined by

$$g^*(\boldsymbol{k}) = \sup_{\boldsymbol{h}\,\in\,\mathbb{R}^{\ell}}\Big\{\langle\boldsymbol{k},\boldsymbol{h}\rangle - g(\boldsymbol{h})\Big\} \tag{2.13}$$

from which it follows by elementary calculus that

$$g^*(\boldsymbol{k}) = \frac{1}{n}\sum_{\boldsymbol{i}\,\in\,\{1,\ldots,n\}^m} \exp\left[nk_{\boldsymbol{i}}\right] - (m-1)\log_e n.$$

If we denote the the adjoint matrix by $A^* \in \mathbb{R}^{\ell\times k}$ then we can write down a standard mathematical statement of the dual problem.

Problem 2.3 (Mathematical statement of the dual problem) *Find*

$$\sup_{\boldsymbol{\varphi}\in\mathbb{R}^k}\left\{\langle \boldsymbol{b},\boldsymbol{\varphi}\rangle - g^*(A^*\boldsymbol{\varphi})\right\}. \tag{2.14}$$

If we let

$$H(\boldsymbol{\varphi}) = \sum_{j=1}^{k} b_j\varphi_j - \frac{1}{n}\sum_{i=1}^{\ell}\exp\left[n\cdot\sum_{j=1}^{k} a_{ij}^*\varphi_j\right] + (m-1)\log_e n$$

then we can use elementary calculus once again to show that if the maximum of $H(\boldsymbol{\varphi})$ occurs when $\boldsymbol{\varphi}=\overline{\boldsymbol{\varphi}}$ then

$$\sum_{i=1}^{\ell} a_{ir}^*\exp\left[n\cdot\sum_{j=1}^{k} a_{ij}^*\overline{\varphi}_j\right] = b_r \tag{2.15}$$

for all $r=1,2,\ldots,k$.
Piantadosi *et al.* [9] showed that the dual problem is much easier to solve than the primal problem and that the solution to the primal problem can be recovered from the solution to the dual problem. Indeed, we can use a Newton iteration to solve the key equations (2.15). The key equations take the form

$$\boldsymbol{q}(\boldsymbol{\varphi}) = \mathbf{0}$$

where

$$q_r(\boldsymbol{\varphi}) = \sum_{i=1}^{\ell} a_{ir}^*\exp\left[n\cdot\sum_{j=1}^{k} a_{ij}^*\varphi_j\right] - b_r$$

for each $r=1,2,\ldots,k$. Now the Newton iteration is given by

$$\boldsymbol{\varphi}^{(j+1)} = \boldsymbol{\varphi}^{(j)} - J^{-1}[\boldsymbol{\varphi}^{(j)}]\boldsymbol{q}[\boldsymbol{\varphi}^{(j)}]$$

where we use the MATLAB inverse of the Jacobian matrix $J\in\mathbb{R}^{k\times k}$. In general there is a closed form for the primal solution $\overline{\boldsymbol{h}}$. Let $\overline{\boldsymbol{k}} = A^*\overline{\boldsymbol{\varphi}}$ and suppose $\overline{k}_{\boldsymbol{j}} > 0$ for all $\boldsymbol{j}\in\{1,2,\ldots,m\}^n$. Then the unique solution to the primal problem (2.2) is given by

$$\overline{\boldsymbol{h}} = \nabla g^*(A^*\overline{\boldsymbol{\varphi}}). \tag{2.16}$$

The underlying analysis is described in the book by Borwein and Lewis [1]. See also the recent survey paper by Borwein [2].

3 Constructing a multi–variate normal copula

We note from [15] that the entropy of the multi–variate normal distribution φ is given by

$$J(\varphi) = \frac{m}{2}\log_e 2\pi + \frac{m}{2} + \frac{1}{2}\log_e\det\Sigma, \tag{3.17}$$

where Σ is the correlation matrix (1.3). It follows that the entropy for the multi–variate normal copula C with density

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

is given by

$$\begin{aligned}
J(C) &= -\int_{[0,1]^m} c(\boldsymbol{u}) \log_e c(\boldsymbol{u})\ d\boldsymbol{u} \\
&= -\int_{\mathbb{R}^m} \frac{\varphi(\boldsymbol{z})}{\Phi'(z_1)\cdots\Phi'(z_m)} \left[\log_e \varphi(\boldsymbol{z}) - \sum_{r=1}^{m} \log_e \Phi'(z_r)\right] \Phi'(z_1)\cdots\Phi'(z_m)\ d\boldsymbol{z} \\
&= -\int_{\mathbb{R}^m} \varphi(\boldsymbol{z}) \log_e \varphi(\boldsymbol{z})\ d\boldsymbol{z} + \sum_{r=1}^{m} \int_{\mathbb{R}^m} \varphi(\boldsymbol{z}) \log_e \Phi'(z_r)\ d\boldsymbol{z} \\
&= J(\varphi) - \sum_{r=1}^{m} \int_{\mathbb{R}^m} \varphi(\boldsymbol{z}) \left(\frac{1}{2}\log_e 2\pi + \frac{z_i^2}{2}\right) d\boldsymbol{z} \\
&= J(\varphi) - \frac{m}{2}\log_e 2\pi - \frac{m}{2} \\
&= \frac{1}{2}\log_e \det \Sigma.
\end{aligned}$$

To match the observed grade correlation coefficients we must find $\theta = \theta_{r,s}$ by solving 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 $1 \le r < s \le m$.

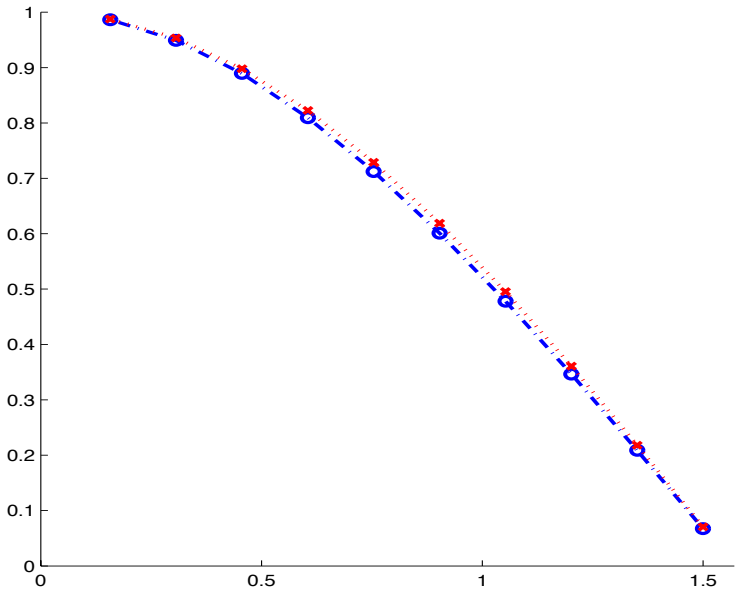


Figure 1: Comparison of $f(\theta)$ in blue circles and $\cos\theta$ in red crosses.

The graph in Figure 1 shows that $f(\theta) \approx \cos\theta$ decreases on $[0, \pi/2]$. Since $f(\theta)$ is an odd function about $\theta = \pi/2$ we know $f(\theta)$ decreases throughout $[0, \pi]$ and hence the equation $f(\theta) = \rho_{r,s}$ can be solved using a simple midpoint iteration. Evaluation of $f(\theta)$ requires a suitable numerical integration package. We have used the MATLAB package *dblquad*. When θ is small the integration becomes unstable but in this region the value of $f(\theta)$ is very close to $\cos\theta$.
For the purpose of simulation and to enable a direct comparison with the method of the previous section, it is convenient to approximate the multi–variate normal copula by a checkerboard copula of the same size. This copula is defined by a hyper–matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ determined from the multi–variate normal copula by the formula

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

for each $\boldsymbol{i} \in \{1, \ldots, n\}^m$. If the partition $-\infty = b(1) < b(2) < \cdots < b(n) < b(n+1) = +\infty$ and the corresponding intervals $J_{\boldsymbol{i}} = \bigtimes_{r=1}^{m} (b(i_r), b(i_r + 1))$ are defined by solving the equations $\Phi(b(k)) = a(k)$ for each $k = 1, \ldots, n+1$ then

$$h_{\boldsymbol{i}} = n \int_{J_{\boldsymbol{i}}} \varphi(\boldsymbol{z}) d\boldsymbol{z}$$

for each $\boldsymbol{i} \in \{1, \ldots, n\}^m$. This should mean, for instance, that standard MATLAB functions can be used for the numerical calculations[1]. 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 manner explained earlier in Section 1.2. For convenience we will refer to this copula as a *normal checkerboard copula*. When using this approximation we also use the formula (1.1) to calculate $\rho_{r,s}$ for each $1 \le r < s \le m$. Thus, we choose $\theta_{r,s}$ so that the calculated values of the grade correlation coefficients $\rho_{r,s}$ match the observed values. The properties of the normal distribution mean that these calculations can be done separately for each $1 \le r < s \le m$ using the relevant marginal bi–variate normal copula.

4 Monthly rainfall data for Sydney

We used official monthly rainfall records for the 150 year period 1859–2008 at station number 0662062, Observatory Hill, Sydney, NSW, Australia. These records are available on the Australian Bureau of Meteorology website http://www.bom.gov.au/climate/data/. 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)

[1]Evaluation of the relevant integrals in MATLAB turned out to be more difficult than we had first imagined. See later notes about the numerical calculations.

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.

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	

4.1 Modelling individual monthly rainfall totals

There are no observed zero rainfall totals and the distributions for individual months can be modelled effectively using a gamma distribution [5, 8, 10, 11, 12, 16]. 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

$$\boldsymbol{\alpha} = (1.817,\ 1.359,\ 1.741,\ 1.333,\ 1.258,\ 1.338,\ 1.202,\ 1.051,\ 1.412,\ 1.468,\ 1.461,\ 1.777)$$

and

$$\boldsymbol{\beta} = (56.40,\ 86.75,\ 74.60,\ 94.70,\ 95.97,\ 97.64,\ 81.56,\ 78.12,\ 49.33,\ 52.31,\ 57.29,\ 43.92).$$

4.2 Simulating individual monthly rainfall totals

Simulated data for the individual monthly totals can be generated in the following way. If $F(x) = P[0 < X \leq x]$ is the fitted cumulative probability distribution for the monthly rainfall total $X \in (0, \infty)$ then the random variable $U = F(X)$ is uniformly distributed on the interval $[0, 1]$. If we generate uniformly distributed pseudo–random numbers $\{u_r\} \in (0, 1)$ then we can generate corresponding pseudo–random monthly rainfall totals $\{x_r\} \in (0, \infty)$ with the desired distribution by setting $x_r = F^{-1}(u_r)$.

4.3 Rainfall in the Spring Quarter

For our particular case study we consider the Spring Quarter rainfall in Sydney. We begin by modelling the total rainfall in the months of September, October and November using gamma distributions with the parameter values

$$\boldsymbol{\alpha} = (1.4115, 1.4682, 1.4608) \quad \text{and} \quad \boldsymbol{\beta} = (49.3327, 52.3126, 57.2866).$$

Figures 2, 3 and 4 show, respectively, histograms of the observed frequency versus the fitted gamma probability density on the left and a histogram of the observed frequency versus a histogram of the pseudo–randomly generated rainfall on the right for each of the months September, October and November.

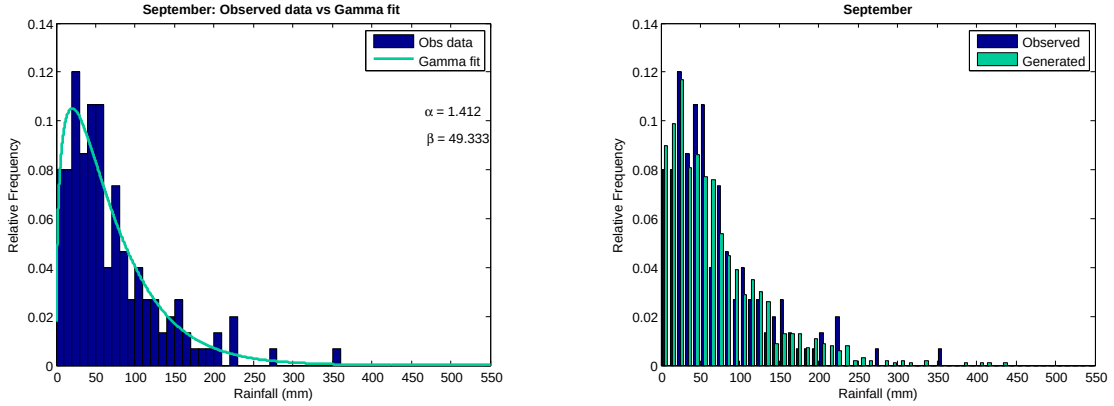


Figure 2: September: Observed rainfall with fitted distribution (left) and generated data (right).

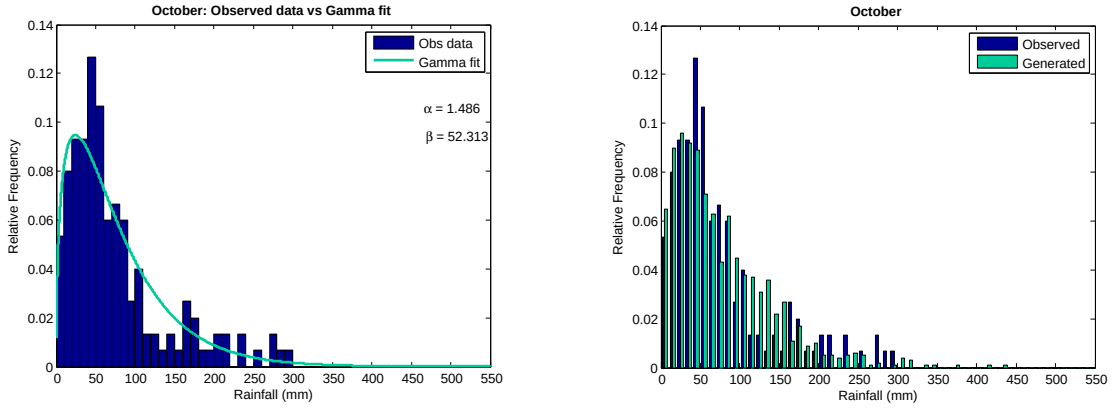


Figure 3: October: Observed rainfall with fitted distribution (left) and generated data (right).

A comparison of the means and variances for the observed, fitted and generated data is given in Table 3. The Kolmogorov–Smirnov goodness–of–fit test was used to assess the fit between the observed and fitted rainfall totals and between the observed and generated totals. The P–values were greater than 0.05 and so we conclude that the null hypothesis, that the samples came from the same distribution as the observations, should not be rejected at the 5% significance level.

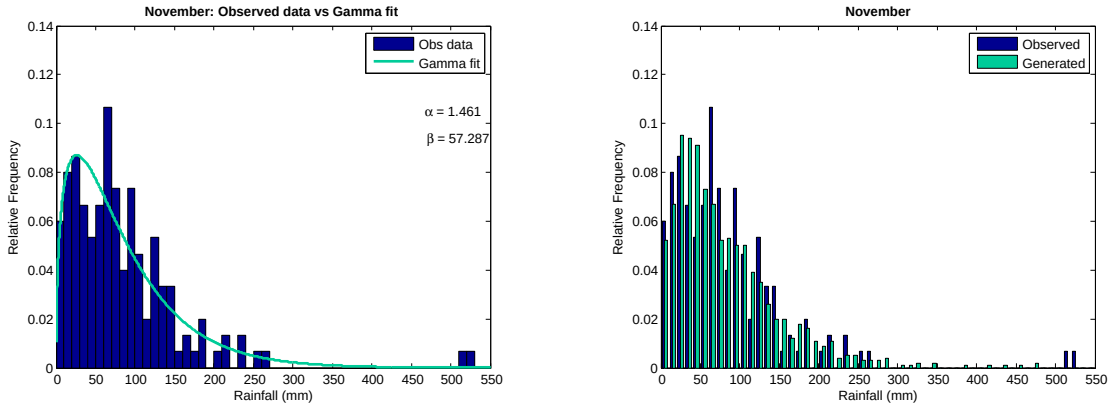


Figure 4: November: Observed rainfall with fitted distribution (left) and generated data (right).

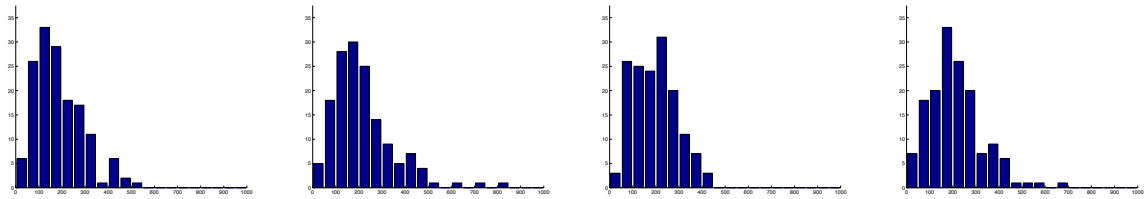
Table 3: Key statistics for observed, fitted and generated data

Data	Mean			Variance		
	September	October	November	September	October	November
Observed	69.633	76.805	83.684	3596.596	4411.235	5699.169
Fitted	69.633	76.805	83.684	3435.189	4017.888	4793.987
Generated	69.795	76.394	84.065	3418.217	4038.737	4760.255

The observed grade correlation coefficients are $\rho_{12} \approx 0.0305$ for September and October, $\rho_{13} \approx 0.0707$ for September and November and $\rho_{23} \approx 0.2169$ for October and November. The generally positive correlation means that we should expect a higher variance in the overall total for the Spring Quarter than would be the case if the monthly rainfall totals were independent. This expectation was confirmed by simulating rainfall in the Spring Quarter using a model where the monthly totals were treated as independent random variables. The MATLAB histograms in Figure 5 for the total rainfall were selected from 10 successive simulations, each one spanning a period of 150 years, using a model in which the monthly totals were treated as independent random variables. The bins were defined by the MATLAB instruction $0 : 50 : 1000$. The sample mean and variance for each simulation are shown under the histograms. Very large and very small monthly totals are relatively infrequent and if the monthly totals are treated as independent random variables then the probability of large totals in all three months or small totals in all three months is extremely small. This probability will increase if the monthly rainfalls are positively correlated. The variance will also increase. Hence, in our case, we expect that the variance predicted by the independent model will be too small. This is indeed the case for all except the sample INDS2 which is apparently a statistical outlier.

These observations strongly suggest we should seek a model using a correlated joint distribution. Suppose the random variable $\boldsymbol{X} = (X_1, \ldots, X_m)^T$ is distributed according to the joint probability density $g : (0, \infty)^m \to (0, \infty)$ defined by

$$g(x_1, \ldots, x_m) = n^{m-1} h_{\boldsymbol{i}} f_1(x_1) \cdots f_m(x_m) \quad \text{when} \quad (F_1(x_1), \ldots, F_m(x_m)) \in I_{\boldsymbol{i}}$$



INDS1: $(214, 10788)$ INDS2: $(239, 17242)$ INDS4: $(221, 9072)$ INDS8: $(236, 12660)$

Figure 5: Spring rainfall in selected simulations using independent random variables

where $\boldsymbol{i} = (i_1, \ldots, i_m)$ and $I_{\boldsymbol{i}}$ is the usual uniform subdivision of the unit hyper–cube $[0,1]^m$ and $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ where $\ell = n^m$ is a multiply–stochastic hyper–matrix. Let $S = \sum_{r=1}^m X_r$ and $\mu = \sum_{r=1}^m \mu_r$ where $\mu_r = E[X_r]$ for each $r = 1, \ldots, m$ and define the interval $K_{\boldsymbol{i}}$ as the inverse image of $I_{\boldsymbol{i}}$ under the mapping $\boldsymbol{F} = (F_1, \ldots, F_m) : (0,\infty)^m \to (0,1)^m$. We have

$$\begin{aligned} E[(S-\mu)^2] &= \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} n^{m-1} h_{\boldsymbol{i}} \int_{K_{\boldsymbol{i}}} (S-\mu)^2 f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m \\ &= \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} n^{m-1} h_{\boldsymbol{i}} \int_{K_{\boldsymbol{i}}} \left[\sum_{r=1}^m (x_r - \mu_r)^2 \right. \\ &\qquad \left. + 2 \sum_{1 \le r < s \le m} (x_r - \mu_r)(x_s - \mu_s) \right] f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m. \end{aligned}$$

If we write $K_{\boldsymbol{i}} = (c_1(i_1), c_1(i_1+1)) \times \cdots \times (c_m(i_m), c_m(i_m+1))$ for each $\boldsymbol{i} = (i_1, \ldots, i_m)$ then we can show by direct integration that

$$\int_{K_{\boldsymbol{i}}} \sum_{r=1}^m (x_r - \mu_r)^2 f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m = \frac{\sum_{r=1}^m \sigma_r(i_r)^2}{n^{m-1}}$$

where

$$\sigma_r(k)^2 = \int_{c_r(k)}^{c_r(k+1)} (x_r - \mu_r)^2 f_r(x_r) dx_r$$

for each $r = 1, \ldots, m$ and each $k = 1, 2, \ldots, n$. We can also show that

$$\int_{K_{\boldsymbol{i}}} \sum_{1 \le r < s \le m} (x_r - \mu_r)(x_s - \mu_s) f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m = \frac{\sum_{1 \le r < s \le m} m_r(i_r) m_s(i_s)}{n^{m-2}}$$

where

$$m_r(k) = \int_{c_r(k)}^{c_r(k+1)} (x_r - \mu_r) f_r(x_r) dx_r$$

for each $r = 1, \ldots, m$ and each $k = 1, 2, \ldots, n$. By noting that $\boldsymbol{h}$ is multiply–stochastic and summing over the relevant terms it follows that

$$E[(S-\mu)^2] = \sum_{r=1}^m \sigma_r^2 \; + \; 2n \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} \left(\sum_{1 \le r < s \le m} m_r(i_r) m_s(i_s) \right) \tag{4.18}$$

where

$$\sigma_r^2 = \int_0^\infty (x_r - \mu_r)^2 f_r(x_r) dx_r$$

for each $r = 1, \ldots, m$. In practice it may be easier to check the validity of the model by simply computing the variance for a sufficiently large pseudo–random sample. We discuss generation of pseudo–random samples in the next subsection.

4.4 Simulating rainfall in the Spring Quarter using a 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 $\ell = n^3$ and $\boldsymbol{i} = (i, j, k) \in \{1, \ldots, n\}^3$ on a uniform partition $\{I_{\boldsymbol{i}}\}$ of the unit cube $(0, 1)^3$. Simulated data for monthly rainfall triples may be generated as follows. Define an order for the indices $\boldsymbol{i} = (i, j, k)$ by saying that $(i, j, k) \prec (i_0, j_0, k_0)$ if $i < i_0$ or if $i = i_0$ and $j < j_0$ or if $i = i_0$ and $j = j_0$ and $k < k_0$. For each pseudo–random number $r \in (0, 1)$ select the interval $I_{i_0 j_0 k_0} = (a(i_0), a(i_0 + 1)) \times (a(j_0), a(j_0 + 1)) \times (a(k_0), a(k_0 + 1))$ if

$$\sum_{(i,j,k) \prec (i_0,j_0,k_0)} h_{ijk} \;<\; nr \;<\; \left[\sum_{(i,j,k) \prec (i_0,j_0,k_0)} h_{ijk} \right] + h_{i_0 j_0 k_0}.$$

Once the interval $I_{i_0 j_0 k_0}$ has been selected the precise position of the pseudo–random point $(u_r, v_r, w_r) \in I_{i_0 j_0 k_0}$ is fixed by generating three more (independent) random numbers $(q_r, s_r, t_r) \in (0, 1)^3$ and setting

$$(u_r, v_r, w_r) = \left(\frac{(i_0 - 1) + q_r}{n}, \frac{(j_0 - 1) + s_r}{n}, \frac{(k_0 - 1) + t_r}{n} \right)$$

and the corresponding rainfall triple is defined by

$$(x_r, y_r, z_r) = \big(F_x^{-1}(u_r), F_y^{-1}(v_r), F_z^{-1}(w_r) \big)$$

where F_x, F_y and F_z are the given marginal distributions.

4.4.1 The fitted tri–variate checkerboard copula of maximum entropy

We set $\rho_{12} = 0.0305$, $\rho_{13} = 0.0707$ and $\rho_{23} = 0.2169$ and calculate

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1040 & 0.0751 & 0.0517 & 0.0339 \\ 0.0800 & 0.0701 & 0.0584 & 0.0463 \\ 0.0589 & 0.0625 & 0.0630 & 0.0606 \\ 0.0415 & 0.0532 & 0.0650 & 0.0757 \end{bmatrix}, \qquad \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0940 & 0.0720 & 0.0525 & 0.0364 \\ 0.0733 & 0.0680 & 0.0600 & 0.0504 \\ 0.0547 & 0.0614 & 0.0656 & 0.0668 \\ 0.0390 & 0.0530 & 0.0686 & 0.0845 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0845 & 0.0686 & 0.0530 & 0.0390 \\ 0.0668 & 0.0656 & 0.0614 & 0.0547 \\ 0.0504 & 0.0600 & 0.0680 & 0.0733 \\ 0.0364 & 0.0525 & 0.0720 & 0.0940 \end{bmatrix}, \qquad \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0757 & 0.0650 & 0.0532 & 0.0415 \\ 0.0606 & 0.0630 & 0.0625 & 0.0589 \\ 0.0463 & 0.0584 & 0.0701 & 0.0800 \\ 0.0339 & 0.0517 & 0.0751 & 0.1040 \end{bmatrix},$$

where $\boldsymbol{h}_i = [h_{ijk}]$. The entropy is given by $J(\boldsymbol{h}) \approx -0.030252$.

4.4.2 The fitted tri–variate normal checkerboard copula

We set $\theta_{12} = 1.5328$, $\theta_{13} = 1.4826$ and $\theta_{23} = 1.2989$ and calculate $\rho_{12} \approx 0.0305$, $\rho_{13} \approx 0.0707$, $\rho_{23} \approx 0.2169$ and also

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1072 & 0.0718 & 0.0531 & 0.0331 \\ 0.0777 & 0.0688 & 0.0604 & 0.0472 \\ 0.0605 & 0.0638 & 0.0633 & 0.0584 \\ 0.0408 & 0.0540 & 0.0635 & 0.0764 \end{bmatrix}, \quad \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0950 & 0.0690 & 0.0538 & 0.0360 \\ 0.0701 & 0.0671 & 0.0620 & 0.0520 \\ 0.0554 & 0.0629 & 0.0656 & 0.0652 \\ 0.0380 & 0.0540 & 0.0669 & 0.0871 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0871 & 0.0669 & 0.0540 & 0.0380 \\ 0.0652 & 0.0656 & 0.0629 & 0.0554 \\ 0.0520 & 0.0620 & 0.0671 & 0.0701 \\ 0.0360 & 0.0538 & 0.0690 & 0.0950 \end{bmatrix}, \quad \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0764 & 0.0635 & 0.0540 & 0.0408 \\ 0.0584 & 0.0633 & 0.0638 & 0.0605 \\ 0.0472 & 0.0604 & 0.0688 & 0.0777 \\ 0.0331 & 0.0531 & 0.0718 & 0.1072 \end{bmatrix},$$

where $\boldsymbol{h}_i = [h_{ijk}]$. The entropy is given by $J(\boldsymbol{h}) \approx -0.030624$.

4.5 Numerical calculations

We used MATLAB for the numerical calculations. The subintervals K_{ijk} are defined by

r	$c_r(1)$	$c_r(2)$	$c_r(3)$	$c_r(4)$	$c_r(5)$
1	0	26.962	54.054	95.635	∞
2	0	30.586	60.243	105.292	∞
3	0	33.207	65.553	114.750	∞

and we calculate

r	$m_r(1)$	$m_r(2)$	$m_r(3)$	$m_r(4)$	sum
1	−13.730	−7.431	0.779	20.381	0.000
2	−14.970	−8.004	0.931	22.042	0.000
3	−16.355	−8.747	1.005	24.077	0.000

for the corresponding moments about the mean and

r	$\sigma_r(1)^2$	$\sigma_r(2)^2$	$\sigma_r(3)^2$	$\sigma_r(4)^2$	$\sigma_r{}^2$
1	767.330	236.020	37.637	2394.201	3435.189
2	913.258	274.367	44.796	2785.467	4017.888
3	1087.259	327.624	53.328	3325.776	4793.987

for the corresponding variances. Note that the row sums of the moments give the total first moment about the mean for each of the three variables and the row sums of the variances give the total variance for each of the three variables. These moments are used in conjunction with the relevant hyper–matrices to calculate the theoretical variances for each of the two copulas using the formula (4.18).

4.6 Simulations using the fitted checkerboard copulas

Suppose that the joint probability is defined by a checkerboard copula on a uniform subdivision I_{ijk} of the unit cube $(0,1)^3$ by a triply–stochastic hyper–matrix $\boldsymbol{h} = [h_{ijk}]$. The probability in simulation that a rainfall triple (x_1, x_2, x_3) will be selected from the interval $K_{ijk} = \boldsymbol{F}^{-1}(I_{ijk})$ is given by $p_{ijk} = 4h_{ijk}$. Convergence of the simulations in probability is extremely slow in terms of real time. Each realization represents one year of real time and 10^6 realizations were required to obtain 3 decimal place accuracy for the hyper–matrices. The simulated values were

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1039 & 0.0752 & 0.0514 & 0.0331 \\ 0.0807 & 0.0710 & 0.0588 & 0.0466 \\ 0.0589 & 0.0623 & 0.0628 & 0.0607 \\ 0.0422 & 0.0535 & 0.0651 & 0.0760 \end{bmatrix}, \quad \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0938 & 0.0718 & 0.0517 & 0.0356 \\ 0.0732 & 0.0680 & 0.0603 & 0.0501 \\ 0.0547 & 0.0611 & 0.0659 & 0.0667 \\ 0.0392 & 0.0530 & 0.0682 & 0.0855 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0841 & 0.0690 & 0.0525 & 0.0386 \\ 0.0678 & 0.0664 & 0.0615 & 0.0544 \\ 0.0502 & 0.0594 & 0.0681 & 0.0729 \\ 0.0369 & 0.0525 & 0.0721 & 0.0943 \end{bmatrix}, \quad \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0755 & 0.0645 & 0.0532 & 0.0415 \\ 0.0610 & 0.0628 & 0.0628 & 0.0592 \\ 0.0453 & 0.0583 & 0.0697 & 0.0792 \\ 0.0342 & 0.0516 & 0.0744 & 0.1050 \end{bmatrix},$$

for the copula of maximum entropy and

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1070 & 0.0719 & 0.0533 & 0.0329 \\ 0.0768 & 0.0695 & 0.0596 & 0.0463 \\ 0.0605 & 0.0642 & 0.0630 & 0.0580 \\ 0.0408 & 0.0536 & 0.0629 & 0.0774 \end{bmatrix}, \quad \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0953 & 0.0685 & 0.0539 & 0.0356 \\ 0.0705 & 0.0671 & 0.0629 & 0.0531 \\ 0.0554 & 0.0625 & 0.0646 & 0.0655 \\ 0.0383 & 0.0546 & 0.0668 & 0.0862 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0870 & 0.0678 & 0.0545 & 0.0382 \\ 0.0648 & 0.0658 & 0.0619 & 0.0563 \\ 0.0527 & 0.0621 & 0.0667 & 0.0712 \\ 0.0353 & 0.0535 & 0.0686 & 0.0948 \end{bmatrix}, \quad \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0768 & 0.0642 & 0.0542 & 0.0411 \\ 0.0575 & 0.0625 & 0.0630 & 0.0603 \\ 0.0473 & 0.0614 & 0.0689 & 0.0775 \\ 0.0335 & 0.0534 & 0.0717 & 0.1070 \end{bmatrix},$$

for the normal copula. In this context one can see that simulation runs of 150 realisations are really very small samples and as such we may expect them to produce quite variable results. This logic can be turned around to speculate that, even without the effects of climate change, rainfall in the Spring Quarter over the next period of 150 years could be quite different from the observed rainfall thus far. One could also argue that such variability undermines our implicit assumption that the observed data provides a representative basis for a model of the entire population.

We began our investigation of the simulations by looking at detailed rainfall patterns in two particular simulations. Each simulation covers a period of 150 years. Monthly rainfalls for selected years from the two simulations are shown in Table 4 and Table 5.

The simulation in Table 4 using the copula of maximum entropy showed the wettest Spring Quarter in year 72 with a total of 662. The driest quarter was in year 33 with a total of 22. The simulation suggests that high quarterly totals may be associated with above average rainfall in all three months, such as in years 78 and 135, or with one or two extreme totals as depicted in years 42 and 69. Very dry quarters were infrequent but not unusual. There was no instance of sustained severe drought but there were instances of successive predominantly below average quarterly totals, such as those in years 116–121.

The simulation in Table 5 using the normal copula showed the wettest Spring Quarter in year 87 with a total of 732 and the driest in year 68 with a total of 33. Once again it is apparent

Table 4: Selected years from a typical simulation using the maximum entropy copula

Year	September	October	November	Total
13	0	139	24	163
33	6	8	8	22
42	36	7	246	289
45	8	25	16	49
69	285	203	1	489
70	91	93	184	368
71	21	38	8	67
72	303	19	340	662
73	148	107	128	383
78	192	215	118	525
87	7	26	4	37
116	4	5	42	51
117	69	59	36	164
118	65	30	95	190
119	83	105	81	269
120	24	2	12	38
121	28	68	45	141
127	288	10	29	327
134	8	37	17	62
135	284	101	250	635

that high quarterly totals may be associated with above average rainfall in all three months or with one or two extreme monthly rainfalls. Very dry quarters may precede or follow very wet quarters. See for instance years 68 and 69 and years 7 and 8. Successive below average totals were generated in years 142–145.

The simulated totals compare favourably with the observed totals. The wettest observed quarter was in 1961 where the total rainfall was 644 and the driest was 1968 when the total was 30. Very wet quarters in 1917, 1950, 1959 and 1976 resulted from above average rainfall in all three months while there were numerous instances, most notably in 1877, 1916, 1943, 1954, 1981, 1987 and 1995 where extreme rainfall was recorded in two of the three months. Successive quarters with consistently below average totals were recorded in 1904–1908, 1936–1941 and 1944–1948.

We tested the intrinsic variability in Spring Quarter rainfall by considering repeated simulations covering a period of 150 years. The first sequence was generated using the copula of maximum entropy and the second using the normal copula. The histograms for selected simulations are displayed in Figure 6. For the copula of maximum entropy the simulation MES8 is essentially an archetypal simulation with a mean value of 231 and a variance of 14399; MES2 has the highest mean of 242 and the highest variance of 20594; MES1 has the lowest mean of 218; and MES10 has the lowest variance of 11977. For the normal copula NS3 has a mean of 230 which is close to the expected value but the variance of 12052 is smaller than expected; NS5 has the highest mean of 244; NS8 has the highest variance of 16468; and NS2 has the lowest variance of 10786. In Table 6 we have shown summary statistics for each sequence of 150–year simulations. The

Table 5: Selected years from a typical simulation using the normal copula

Year	September	October	November	Total
6	53	37	17	107
7	169	152	119	440
8	1	21	55	77
9	4	7	67	78
13	51	9	220	280
14	64	104	311	479
33	209	33	323	565
68	2	25	6	33
69	43	131	304	478
72	18	0	47	65
87	162	64	506	732
91	54	21	17	92
92	148	243	15	406
118	32	16	15	63
119	63	22	19	104
120	3	13	31	47
142	73	87	10	170
143	7	40	78	125
144	22	132	13	167
145	52	53	65	170

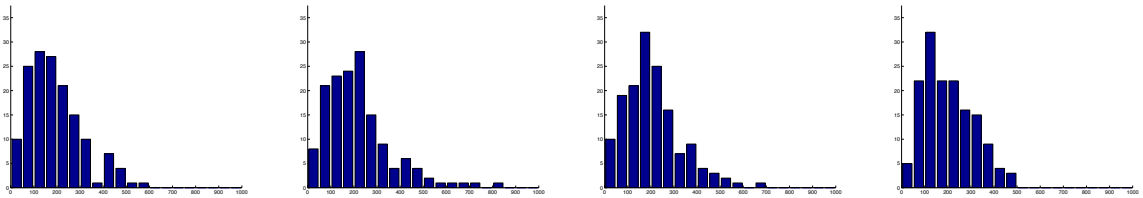
summary statistics vary significantly with the mean lying in the range $(211, 242)$ and the variance lying in the range $(10785, 20594)$. The error vector $\boldsymbol{e}$ is defined as the difference between the theoretical probabilities defined for the intervals K_{ijk} by the relevant triply–stochastic hyper–matrix and the corresponding relative frequencies generated by the pseudo–random simulation. The probability error $e = \|\boldsymbol{e}\|$ displayed in Table 6 is the Euclidean norm of this error vector. We can analyse the error e more precisely in the following way. If there are N realizations and if we renumber the intervals K_{ijk} where $(i, j, k) \in \{1, 2, 3, 4\}^3$ in the form $K_1, K_2, \ldots, K_\ell$ where $\ell = 4^3$ then for each $r = 1, \ldots, \ell$ we have

$$E[e_r^2] = E\left[\left(p_r - \frac{N_r}{N}\right)^2\right] = \sum_{N_1+\cdots+N_\ell=N} \left(p_r - \frac{N_r}{N}\right)^2 \binom{N}{N_1 \cdots N_\ell} p_1^{N_1} \cdots p_\ell^{N_\ell} = \frac{p_r(1-p_r)}{N}$$

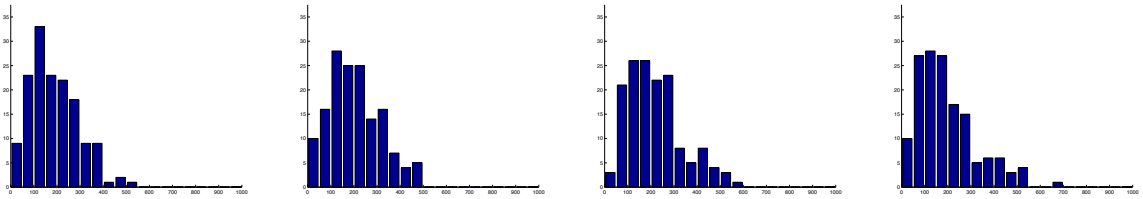
since $p_1 + \cdots + p_\ell = 1$ and hence the expected square error is

$$E[e^2] = \sum_{r=1}^{\ell} E[e_r^2] = \frac{1}{N} \sum_{r=1}^{\ell} p_r(1-p_r) \leq \frac{1}{N}.$$

It follows that $\sqrt{E[e^2]} \leq 1/\sqrt{N}$. Note that $1/\sqrt{150} \approx 0.081650$ and $1/\sqrt{15000} \approx 0.008165$. We tested stochastic convergence by considering simulations covering a period of 15000 years. In Figure 7 we have displayed histograms from selected archetypal simulations, each covering



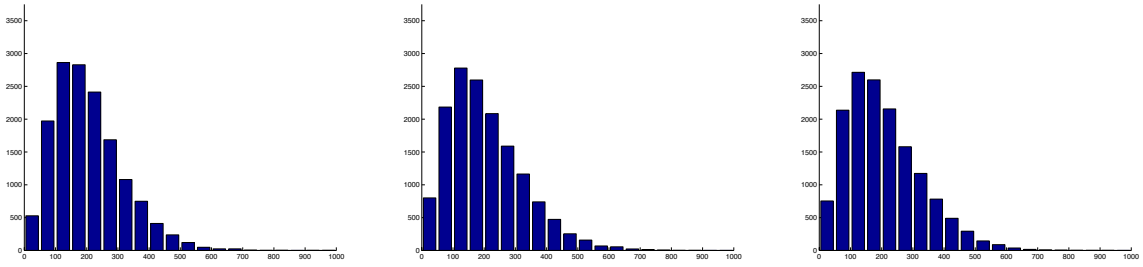
MES1: $(218, 13740)$ MES2: $(242, 20594)$ MES8: $(231, 14399)$ MES10: $(227, 11977)$



NS2: $(211, 10786)$ NS3: $(230, 12052)$ NS5: $(244, 14327)$ NS8: $(219, 16468)$

Figure 6: Total rainfall for selected simulations using a checkerboard copula

a period of $15,000$ years, for the independent model and the two correlated models. Although the histograms are visually stable for simulations with this number of realizations the sample statistics still show some variation. Thus we have chosen to select simulations with sample mean and variance that are close to the theoretical values. Although the histogram for the independent simulation is slightly taller and narrower the the visual differences are minimal. Nevertheless our numerical calculations show that the variance for the independent simulation is significantly smaller. The histograms and associated summary statistics are shown in Figure 7.



IND15000A: $(230, 12240)$ ME15000A: $(229, 14321)$ N15000A: $(231, 14340)$

Figure 7: Comparison of three archetypal simulations–independent (left); maximum entropy copula (centre); normal copula (right).

For the copula of maximum entropy and the normal copula we ran also repeated simulations covering a period of $15,000$ years. The summary statistics are shown in Table 7. In moving from simulations with 150 realizations shown in Table 6 to simulations with $15,000$ realizations shown in Table 7 the probability errors are reduced by one order of magnitude.

Table 6: Statistics and probability errors for typical simulations (150 years)

run	me mean	me var	e	n mean	n var	e
1	218.051	13740.47	0.081178	228.886	13751.66	0.077355
2	242.358	20593.60	0.083936	211.358	10785.57	0.080275
3	235.513	14223.73	0.080700	229.887	12051.58	0.079029
4	224.125	12190.76	0.079674	232.618	13684.20	0.068920
5	227.902	15024.04	0.086871	244.082	14326.78	0.092515
6	231.622	15137.70	0.089484	221.677	13538.77	0.080973
7	227.888	12985.29	0.074645	238.924	14163.22	0.087583
8	230.708	14399.34	0.096069	219.864	14468.45	0.081098
9	231.820	15105.23	0.081699	218.375	13739.63	0.079944
10	226.928	11977.45	0.092291	241.420	15304.71	0.085598

Table 7: Statistics and probability errors for typical simulations (15000 years)

run	me mean	me var	e	n mean	n var	e
1	231.451	14539.69	0.007605	229.735	14430.86	0.008110
2	230.754	14560.60	0.007935	229.370	14166.38	0.010244
3	230.372	14340.78	0.008827	228.909	14249.83	0.009168
4	228.611	14244.27	0.007842	230.913	14271.26	0.007875
5	229.524	14139.78	0.006840	230.635	14313.10	0.008829
6	229.238	14384.39	0.008706	230.539	14430.73	0.007616
7	230.030	14479.19	0.007641	230.329	14519.17	0.008243
8	229.788	14089.51	0.007622	229.806	14365.04	0.007749
9	229.322	14379.86	0.008251	231.063	14388.86	0.009063
10	229.997	14035.33	0.007016	230.757	14406.03	0.007634

In Table 8 the summary statistics for the observed sums for the Spring Quarter rainfall are compared to the summary statistics for the generated sums from all three models. The simulation statistics were obtained as averages over a period of 3×10^6 years.
We emphasize that the variance of the synthetic rainfall totals generated by the independent model is significantly less than that for the observed sums. The variance of the total generated using either the copula of maximum entropy or the normal copula is much closer to the observed value. If we propose a conventional null hypothesis that the two simulated Spring Quarter totals come from the same population as the observed totals then the Kolmogorov–Smirnov goodness–of–fit test shows that the hypothesis should not be rejected at the 5% significance level.

5 Conclusions

Our investigation shows that the variance of the generated quarterly rainfall totals can be significantly changed by using a copula that allows us to incorporate the observed correlation.

Table 8: Comparison of three models for Spring Quarter rainfall

	mean	variance
observed	230.123	14391.34
independent (theory)	230.123	12247.06
independent (simulation)	230.155	12236.11
maximum entropy copula (theory)	230.123	14318.11
maximum entropy copula (simulation)	230.085	14319.45
normal copula (theory)	230.123	14348.46
normal copula (simulation)	230.058	14347.05

We 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 tri–variate distribution is the most disordered distribution that preserves the prescribed marginal distributions and matches the observed grade correlation.

There are several reasons why we wished to compare the checkerboard copula of maximum entropy with a normal checkerboard copula. In the first place the normal distribution contains a specific parameter for the correlation. Thus, it seems sensible to investigate a copula derived from the normal distribution. In the second place the normal distribution is a natural distribution to describe the addition of unrelated random events. One could argue, at a microscopic level, that rainfall is a process of this type. At a macroscopic level there are climatic processes that cause systematic variations and dependencies that we may wish to describe. There are also technical problems that must be overcome. Rainfall accumulations are non–negative and so some transformation of the raw data is necessary.

More pedantically there is a difference between the correlation of the marginal normal distributions, represented by the matrix $\Sigma = [\cos\theta_{r,s}]$, and the grade correlation coefficients. Our numerical calculation of the grade correlations shows that this difference is quite small. If one used the normal copula directly it would be very easy and not unreasonable to ignore this difference. On the other hand, if one uses the checkerboard normal copula, as we have done, then the correct values for $\Theta = [\theta_{r,s}]$ must be computed numerically from the associated triply–stochastic hyper–matrix $\boldsymbol{h}$. This computation is quite straightforward in MATLAB.

In comparing the two methods there is little to distinguish them. The normal checkerboard copula turns out to be close to the maximum entropy checkerboard copula of the same size in all of the examples we considered, irrespective of the number of subdivisions. In two dimensions we found 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.* [9] 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 our preliminary calculations for the case study considered in this paper with the standard MATLAB package *triplequad* did not given sufficiently accurate answers for the required probability integrals. More work is still required to determine why this was so. Thus we used an

alternative procedure to determine the probabilities for the normal checkerboard copula in our three dimensional example. This procedure, which we now describe, is essentially a counting procedure and it should be generally well suited to calculation in MATLAB.
Select a large number of equally spaced points $\boldsymbol{v} \in [0,1]^m$. Map these points into $\mathbb{R}^m$ using the transformation $\boldsymbol{w} = \Phi^{-1}(\boldsymbol{v}) \Leftrightarrow w_r = \Phi^{-1}(v_r)$. Choose the orthogonal matrix P such that $\Lambda = P^T \Sigma P \Leftrightarrow P \Lambda P^T = \Sigma$ where Λ is a positive diagonal matrix and rescale the points according to the transformation $\boldsymbol{y} = \Lambda^{1/2} \boldsymbol{w}$. Map the points $\boldsymbol{y} \in \mathbb{R}^m$ onto points $\boldsymbol{z} = P\boldsymbol{y} \in \mathbb{R}^m$. Then return the points to $[0,1]^m$ using the map $\boldsymbol{u} = \Phi(\boldsymbol{z}) \Leftrightarrow u_r = \Phi(z_r)$. Thus

$$\boldsymbol{u} = \Phi[P\,\Lambda^{1/2}\Phi^{-1}(\boldsymbol{v})] \Leftrightarrow \boldsymbol{v} = \Phi[\Lambda^{-1/2}P^T\Phi^{-1}(\boldsymbol{u})].$$

Now count the proportion $p_{\boldsymbol{i}}$ of points in each of the intervals $I_{\boldsymbol{i}}$ where $\boldsymbol{i} \in \{1,2,\ldots,n\}^m$. The multiply–stochastic hyper–matrix $\boldsymbol{h}$ is defined by $h_{\boldsymbol{i}} = np_{\boldsymbol{i}}$ for all $\boldsymbol{i} \in \{1,2,\ldots,n\}^m$. The rationale is that the points $\boldsymbol{v}$ are independently distributed but the points $\boldsymbol{u}$ are distributed according to the correlation specified by φ. In mathematical terms we have

$$\begin{aligned}
P[\boldsymbol{u} \in I_{\boldsymbol{i}}] &= \int_{I_{\boldsymbol{i}}} \frac{\varphi(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_m))}{\Phi'(\Phi^{-1}(u_1))\cdots\Phi'(\Phi^{-1}(u_m))}\,d\boldsymbol{u} \\
&= \int_{\Phi^{-1}(I_{\boldsymbol{i}})} \varphi(\boldsymbol{z})d\boldsymbol{z} \\
&= \int_{P^T\Phi^{-1}(I_{\boldsymbol{i}})} \varphi(P\boldsymbol{y})\,d\boldsymbol{y} \qquad (\text{since } \det P = 1) \\
&= \int_{\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})} \varphi(P\Lambda^{1/2}\boldsymbol{w})\det\Lambda^{1/2}\,d\boldsymbol{w} \\
&= \int_{\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})} \frac{1}{(2\pi)^{m/2}(\det\Sigma)^{1/2}}\exp[-\boldsymbol{w}^T\Lambda^{1/2}P^T\Sigma^{-1}P\Lambda^{1/2}\boldsymbol{w}/2]\,(\det\Lambda)^{1/2}\,d\boldsymbol{w} \\
&= \int_{\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})} \frac{1}{(2\pi)^{m/2}}\exp[-\boldsymbol{w}^T\boldsymbol{w}/2]\,d\boldsymbol{w} \qquad (\text{since } \det\Sigma = \det\Lambda) \\
&= \int_{\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})} \Phi'(w_1)\cdots\Phi'(w_m)\,d\boldsymbol{w} \\
&= \int_{\Phi[\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})]} d\boldsymbol{v} \\
&= V\left(\Phi[\Lambda^{-1/2}P^T\Phi^{-1}(I_{\boldsymbol{i}})]\right).
\end{aligned}$$

In order to calculate the required tri–variate normal copula with sufficient accuracy in the above case study it was necessary to choose 256^3 equally spaced points in the unit cube. We were able to use an elementary MATLAB program to make the calculations. For higher dimensional examples the numerical calculations for the normal copula may be more challenging.

References

[1] Jonathan M. Borwein and Adrian S. Lewis (2000), Convex Analysis and Nonlinear Optimization, Theory and Examples, CMS Books in Mathematics, Springer–Verlag New York, Inc. Expanded second edition (2005).

[2] Jonathan M. Borwein (2011), Maximum entropy and feasibility methods for convex and nonconvex inverse problems, *Optimization*, Invited survey paper, (to appear), pre–print http://carma.newcastle.edu.au/jon/inverse-paper.pdf.

[3] Fowler, H.J., Kilsby, C.G., O'Connell, P.E. & Burton, A. (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.

[4] Md Masud Hasan, Peter K. Dunn (2010), Two Tweedie distributions that are near optimal for modelling monthly rainfall in Australia, *International J Climatology*, DOI: 10.1002/joc.2162.

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

[6] Nelsen, R. B. (1999), An Introduction to Copulas. Springer Lecture Notes in Stat., New York.

[7] Julia Piantadosi, Phil Howlett and John Boland (2007), Matching the Grade Correlation Coefficient using a copula with maximum disorder, *Journal of Industrial and Management Optimization*, **3**(2), 305–312.

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

[9] Julia Piantadosi, Phil Howlett, Jonathan Borwein (2010), Copulas with Maximum Entropy, *Optimization Letters*, doi: 10.1007/s11590-010-0254-2.

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

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

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

[13] Wang, Ruye (2006), Conditional and marginal of multivariate Gaussian, http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node5.html.

[14] C. S. Withers and S. Nadarajah (2011), On the compound Poisson–Gamma distribution, *Kybernetika*, **47**(1), 15–37.

[15] http://en.wikipedia.org/wiki/Differential_entropy.

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