MAXIMUM ENTROPY METHODS FOR GENERATING SIMULATED RAINFALL

JULIA PIANTADOSI AND PHIL HOWLETT

Scheduling and Control Group
Centre for Industrial and Applied Mathematics (CIAM)
University of South Australia
Mawson Lakes, SA 5095, Australia

JONATHAN BORWEIN

Centre for Computer Assisted Research Mathematics and its Applications (CARMA)
University of Newcastle
Callaghan, NSW 2308, Australia

AND

JOHN HENSTRIDGE

Data Analysis Australia Pty Ltd
School of Mathematics and Statistics
University of Western Australia
Crawley WA, 6009, Australia

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 prescribed 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. Introduction. It has been usual to model both short-term and long-term rainfall accumulations at a specific location by a gamma distribution [3, 4, 12, 17]. Some authors [5, 15] 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 that of 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

2000 Mathematics Subject Classification. Primary: 15B51, 60E05, 65C05; Secondary: 52A41.
Key words and phrases. Maximum entropy methods, multiply stochastic hyper-matrices, multi-dimensional copulas, rainfall simulation.
to construct a joint probability distribution which preserves the prescribed marginal distributions and matches the observed grade correlation coefficients. We propose two alternative ways to do this. Both methods use multi-dimensional copulas.

1.1. Structure of the paper. In Section 2 we explain how a multi-dimensional copula can be used systematically to define a joint probability distribution with prescribed marginals. We recall the checkerboard copula of maximum entropy designed by Piantadosi et al. [7, 9] to match the observed grade correlation coefficients and propose a multi-variate normal copula that can be used for a similar purpose. In Section 3 we outline construction of the checkerboard copula of maximum entropy and state the key results obtained by Piantadosi et al. in [9]. The multi-variate normal copula and a corresponding checkerboard approximation are discussed in Section 4 and we show that the parameters can be chosen to match the observed grade correlations. In Section 5 we apply the checkerboard copulas to modelling and simulation of rainfall during the Spring Quarter (September, October, November) for Sydney, Australia. This is the main section. We note that monthly rainfall totals can be modelled as random variables using a Gamma distribution but argue that rainfall accumulations for the entire Spring Quarter are not modelled effectively if the monthly totals are regarded as independent random variables. In particular we show that the variance of the Spring Quarter rainfall generated by a model with independent monthly totals is much smaller than the observed variance. Subsequently we show that a joint probability distribution defined either by a checkerboard copula of maximum entropy or by a normal checkerboard copula provides a much better model. In each case the new model preserves the marginal monthly distributions and matches the observed grade correlation coefficients. Thus the mean of the Spring Quarter rainfall is preserved while the variance is increased to a value that is close to the observed value. This is our main result. Our findings are supported by theoretical arguments and by extensive simulations. The numerical calculations are also described. In Section 6 we state our conclusions and discuss an unexpected difficulty that arose in connection with numerical calculation of multi-dimensional normal probabilities on small hyper-cubes.

2. Modelling joint probability distributions with prescribed marginals. In this section we use a multi-dimensional copula to construct a joint probability density with prescribed marginals and we introduce two special checkerboard copulas that allow us to match the observed grade correlations—the copula of maximum entropy proposed by Piantadosi et al. [7, 9] and a copula defined by a multi-variate normal distribution. We refer to Nelsen [6] for general information about copulas.

2.1. 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

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

where $x = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$ is a joint probability distribution for the vector-valued random variable $X = (X_1, \ldots, X_m)^T$ 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
\[ g(x) = c(F_1(x_1), \ldots, F_m(x_m))f_1(x_1) \cdots f_m(x_m) \]
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 \( U = (U_1, \ldots, U_m)^T \). The \textit{grade correlation coefficients} for \( X \) are defined by
\[ \rho_{r,s} = \frac{E[(F_r(X_r) - 1/2)(F_s(X_s) - 1/2)]}{\sqrt{E[(F_r(X_r) - 1/2)^2] \cdot E[(F_s(X_s) - 1/2)^2]}} \]
\[ = \frac{E[(U_r - 1/2)(U_s - 1/2)]}{\sqrt{E[(U_r - 1/2)^2] \cdot E[(U_s - 1/2)^2]}} \]
\[ = 12E[U_rU_s] - 3 \]
for each \( 1 \leq r < s \leq m \). Thus, the grade correlation coefficients for \( X \) are simply the correlations for \( U \). The \textit{entropy} for the copula \( C \) with density \( c \) is defined by
\[ J(C) = (-1) \int_{[0,1]^m} c(u) \log_e c(u) \, du \]
where \( u = (u_1, \ldots, u_m)^T \in [0,1]^m \). The entropy \( J(C) \) of the copula measures the inherent disorder of the distribution. The most disordered copula is the one with \( c(u) = 1 \) for all \( u \in [0,1]^m \) for which \( J(C) = 0 \).

2.2. \textbf{Checkerboard copulas of maximum entropy}. An \( m \)-dimensional \textit{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 \( n \in \mathbb{N} \) 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 \( \ell = n^m \) congruent hyper-cubes. We can now construct an elementary checkerboard copula \( C = C_h \) with density \( c = c_h \) defined by the elements of an \( m \)-dimensional hyper-matrix \( h \in \mathbb{R}^\ell \) in such a way that the density takes a constant non-negative value on each hyper-cube of the subdivision. If \( h = [h_i] \in \mathbb{R}^\ell \) where \( i = (i_1, \ldots, i_m) \in \{1, \ldots, n\}^m \) is scaled to be multiply stochastic then it follows from the discussion in Section 3 that the grade correlation coefficients for \( C_h \) are given by
\[ \rho_{r,s} = 12 \left[ \frac{1}{n^3} \sum_{i \in \{1,\ldots,n\}^m} h_i(i_r - 1/2)(i_s - 1/2) \right] - 3 \quad (1) \]
and the \textit{entropy} of \( h \) is given by
\[ J(h) = (-1) \left[ \frac{1}{n} \sum_{i \in \{1,\ldots,n\}^m} h_i \log_e h_i + (m - 1) \log_e n \right] . \quad (2) \]
If \( n \in \mathbb{N} \) is sufficiently large then Piantadosi et al. [7, 9] showed that \( h \) can be chosen in such a way that the observed grade correlations are imposed and the entropy of the hyper-matrix is maximized. Since entropy is a measure of disorder the solution proposed by Piantadosi et al. for \( c_h \) can be interpreted as the \textit{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 \( C = C_h \) is the most disordered such copula.

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

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

where \( z = (z_1, \ldots, z_m)^T \in \mathbb{R}^m \) and where

\[
\Sigma = E[ZZ^T] = [\cos \theta_{r,s}] \in [-1, 1]^{m \times m}
\]

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 \([14]\) given by

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

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

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

is the density for the \( m \)-dimensional normal copula \( C : [0, 1]^m \to [0, 1] \) defined by

\[
C(u) = \int_{\prod_{i=1}^{m} [0, u_i]} c(v) dv.
\]

We note from \([14]\) 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

\[
\Sigma_{r,s} = \begin{bmatrix}
1 & \cos \theta_{r,s} \\
\cos \theta_{r,s} & 1
\end{bmatrix}
\]

and hence the grade correlation coefficients for the copula \( C \) can be calculated from the formula

\[
\rho_{r,s} = \frac{6}{\pi \sin \theta_{r,s}} \int_{\mathbb{R}^2} \Phi(z_r)\Phi(z_s) \exp \left[ -\frac{(z_r^2 - 2\cos \theta_{r,s} z_r z_s + z_s^2)}{2\sin^2 \theta_{r,s}} \right] dz_r dz_s - 3 \quad (5)
\]

for each \( r < s \). From \([16]\) the entropy is given by

\[
J(C) = \frac{1}{2} \log e \det \Sigma.
\]

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 \([16]\) in one dimension for 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. For purposes of comparison with the copula of maximum entropy we will approximate the normal copula by a checkerboard copula.

3. Constructing a checkerboard 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 \( h \) be a non-negative \( m \)-dimensional hyper-matrix given by \( h = [h_i] \in \mathbb{R}^\ell \) where \( \ell = n^m \) and \( i \in \{1, \ldots, n\}^m \) with \( h_i \in [0, 1] \). Define the marginal sums \( \sigma_r : \{1, \ldots, n\} \mapsto \mathbb{R} \) by the formulae

\[
\sigma_r(i_r) = \sum_{j \neq r, i_j \in \{1, \ldots, n\}} h_i
\]

for each \( r = 1, 2, \ldots, m \). If \( \sigma_r(i_r) = 1 \) for all \( i_r \in \{1, \ldots, n\} \) and all \( r = 1, 2, \ldots, m \) then we say that \( h \) 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 \( c_h : [0, 1]^m \mapsto [0, \infty) \) almost everywhere by the formula

\[
c_h(u) = n^{m-1} \cdot h_i \quad \text{if} \quad u \in I_i = \times_{r=1}^m (a(i_r), a(i_r + 1))
\]

for each \( i = (i_1, \ldots, i_m) \in \{1, 2, \ldots, n\}^m \). Now it follows that the step function \( c_h : [0, 1]^m \mapsto [0, \infty) \) defines a corresponding copula \( C_h : [0, 1]^m \mapsto [0, 1] \) by the formula

\[
C_h(u) = \int_{\times_{r=1}^m [0, u_i]} c_h(v) dv
\]

for all \( u \in [0, 1]^m \). The formulae (1) and (2) can be established by direct integration. It is also possible to show that

\[
-1 + \frac{1}{n^2} \leq \rho_{r,s} \leq 1 - \frac{1}{n^2}.
\]

(7)

See [9] for more details. The checkerboard copula of maximum entropy is the checkerboard copula \( C_h \) defined by the hyper-matrix \( h \) that solves the following problem.

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

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

subject to the constraints

\[
\sum_{j \neq r, i_j \in \{1, \ldots, n\}} h_i = 1
\]

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

\[
h_i \geq 0
\]

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

\[
12 \left[ \frac{1}{n^3} \sum_{i \in \{1, \ldots, n\}^m} h_i(i_n - 1/2)(i_n - 1/2) \right] - 3 = \rho_{r,s}
\]

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
\[
g(h) = \begin{cases} \frac{-1}{n} J(h) & \text{if } h_j \geq 0 \text{ for all } j \in \{1, 2, \ldots, m\}^n \\ +\infty & \text{otherwise} \end{cases}
\]
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 (9) and (11) in the form \( A h = b \) 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 (10) as it is enforced by the entropy. The primal problem can be restated in standard mathematical form.

**Problem (Mathematical statement of the primal problem).** Find
\[
\inf_{h \in \mathbb{R}^\ell} \left\{ g(h) \mid A h = b \right\}.
\]
(12)

The Fenchel conjugate function \( g^* : \mathbb{R}^\ell \mapsto \mathbb{R} \cup \{-\infty\} \) is defined by
\[
g^*(k) = \sup_{h \in \mathbb{R}^\ell} \left\{ (k, h) - g(h) \right\}
\]
(13)
from which it follows by elementary calculus that
\[
g^*(k) = \frac{1}{n} \sum_{i \in \{1, \ldots, n\}^m} \exp[nk_i] - (m - 1) \log_e n.
\]
If we denote the adjoint matrix by \( A^\ast \in \mathbb{R}^{\ell \times k} \) then we can write down a standard mathematical statement of the dual problem.

**Problem (Mathematical statement of the dual problem).** Find
\[
\sup_{\varphi \in \mathbb{R}^k} \left\{ (b, \varphi) - g^*(A^\ast \varphi) \right\}.
\]
(14)

If we let
\[
H(\varphi) = \sum_{j=1}^{k} b_j \varphi_j - \frac{1}{n} \sum_{i=1}^{\ell} \exp \left[ n \cdot \sum_{j=1}^{k} a_{ij}^* \varphi_j \right] + (m - 1) \log_e n
\]
then we can use elementary calculus once again to show that if the maximum of \( H(\varphi) \) occurs when \( \varphi = \varphi^\ast \) then
\[
\sum_{i=1}^{\ell} a_{ir}^* \exp \left[ n \cdot \sum_{j=1}^{k} a_{ij}^* \varphi_j \right] = b_r
\]
(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 (15). These equations take the form
\[
q(\varphi) = 0
\]
where

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

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

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

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 \( h \). Let \( k = A^* \varphi \) and suppose \( k_j > 0 \) for all \( j \in \{1, 2, \ldots, m\} \). Then the unique solution to the primal problem (3) is given by

\[ h = \nabla g^*(A^* \varphi) \] (16)

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

4. Constructing a multi-variate normal copula. We note from [16] that the entropy of the multi-variate normal distribution \( \varphi \) is given by

\[ J(\varphi) = \frac{m}{2} \log_2 e^{2\pi} + \frac{m}{2} + \frac{1}{2} \log_2 \det \Sigma, \] (17)

where \( \Sigma \) is the correlation matrix (3). It follows that the entropy for the multi-variate normal copula \( C \) with density

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

is given by

\[
J(C) = (-1) \int_{[0,1]^m} c(u) \log_2 c(u) \, du \\
= \int_{\mathbb{R}^m} \varphi(z) \log_2 \varphi(z) \, dz + \sum_{r=1}^m \int_{\mathbb{R}^m} \varphi(z) \log_2 \Phi'(z_r) \, dz \\
= J(\varphi) - \sum_{r=1}^m \int_{\mathbb{R}^m} \varphi(z) \left( \frac{1}{2} \log_2 e^{2\pi} + \frac{z_r^2}{2} \right) \, dz \\
= J(\varphi) - m \frac{1}{2} \log_2 e^{2\pi} - m \frac{1}{2} \\
= \frac{1}{2} \log_2 \det \Sigma.
\]

To match the observed grade correlation coefficients we must find \( \theta = \theta_{r,s} \) by solving the equation \( f(\theta) = \rho_{r,s} \) where

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

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

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 \texttt{dblquad}. When \( \theta \) 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. This copula is defined by a hyper-matrix $h = [h_i] \in \mathbb{R}^d$ determined from the multi-variate normal copula by the formula

$$h_i = \int_{I_i} c(u) du$$

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

$$h_i = \int_{J_i} \varphi(z) dz$$

for each $i \in \{1, \ldots, n\}^m$. This should mean, for instance, that standard MATLAB functions can be used for the numerical calculations. Finally the step function $c_h : [0, 1]^m \mapsto \mathbb{R}$ and the corresponding copula $C_h : [0, 1]^m \mapsto [0, 1]$ are defined in the manner explained earlier in Section 2.2. For convenience we will refer to this copula as a normal checkerboard copula. When using this approximation we also use the formula (1) to calculate $\rho_{r,s}$ for each $1 \leq r < s \leq m$. 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 $1 \leq r < s \leq m$ using the relevant marginal bi-variate normal copula.

---

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.
5. 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

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$</td>
<td>103</td>
<td>118</td>
<td>130</td>
<td>126</td>
<td>103</td>
<td>131</td>
<td>98</td>
<td>82</td>
<td>70</td>
<td>77</td>
<td>84</td>
<td>78</td>
</tr>
<tr>
<td>$s$</td>
<td>76</td>
<td>110</td>
<td>103</td>
<td>112</td>
<td>111</td>
<td>116</td>
<td>82</td>
<td>84</td>
<td>60</td>
<td>66</td>
<td>76</td>
<td>63</td>
</tr>
</tbody>
</table>

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.

Table 2. Grade correlation coefficients for all monthly pairs

<table>
<thead>
<tr>
<th></th>
<th>Ja</th>
<th>Fe</th>
<th>Mr</th>
<th>Ap</th>
<th>Ma</th>
<th>Jn</th>
<th>Jl</th>
<th>Au</th>
<th>Se</th>
<th>Oc</th>
<th>No</th>
<th>De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
<td>.18</td>
<td>.06</td>
<td>-19</td>
<td>.01</td>
<td>-.02</td>
<td>-.02</td>
<td>.13</td>
<td>.09</td>
<td>-.16</td>
<td>.05</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Fe</td>
<td>.18</td>
<td>.03</td>
<td>.08</td>
<td>-.09</td>
<td>.05</td>
<td>-.01</td>
<td>.10</td>
<td>.09</td>
<td>-.05</td>
<td>.08</td>
<td>.07</td>
<td>.06</td>
</tr>
<tr>
<td>Mr</td>
<td>-.06</td>
<td>-.03</td>
<td>.11</td>
<td>.04</td>
<td>.19</td>
<td>-.14</td>
<td>-.15</td>
<td>-.12</td>
<td>.15</td>
<td>-.05</td>
<td>-.01</td>
<td>-.01</td>
</tr>
<tr>
<td>Ap</td>
<td>-.19</td>
<td>-.08</td>
<td>.11</td>
<td>.18</td>
<td>.05</td>
<td>.13</td>
<td>.12</td>
<td>-.08</td>
<td>.11</td>
<td>.09</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>Ma</td>
<td>-.01</td>
<td>-.09</td>
<td>.04</td>
<td>.18</td>
<td>.05</td>
<td>-.02</td>
<td>-.05</td>
<td>-.08</td>
<td>.07</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>Jn</td>
<td>-.02</td>
<td>.05</td>
<td>.19</td>
<td>.05</td>
<td>.05</td>
<td>-.04</td>
<td>-.07</td>
<td>-.17</td>
<td>.02</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>Jl</td>
<td>-.02</td>
<td>-.01</td>
<td>.14</td>
<td>.13</td>
<td>-.02</td>
<td>-.04</td>
<td>.11</td>
<td>.12</td>
<td>.08</td>
<td>-.08</td>
<td>-.02</td>
<td>-.02</td>
</tr>
<tr>
<td>Au</td>
<td>.13</td>
<td>.10</td>
<td>-.15</td>
<td>.12</td>
<td>-.05</td>
<td>-.07</td>
<td>.11</td>
<td>.13</td>
<td>.13</td>
<td>.12</td>
<td>.09</td>
<td>.09</td>
</tr>
<tr>
<td>Se</td>
<td>.09</td>
<td>.09</td>
<td>.12</td>
<td>-.08</td>
<td>-.08</td>
<td>-.17</td>
<td>.12</td>
<td>.13</td>
<td>.04</td>
<td>.07</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Oc</td>
<td>-.16</td>
<td>-.05</td>
<td>.15</td>
<td>.11</td>
<td>-.07</td>
<td>.02</td>
<td>.08</td>
<td>.13</td>
<td>.04</td>
<td>.22</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>No</td>
<td>.05</td>
<td>.08</td>
<td>.05</td>
<td>.09</td>
<td>.05</td>
<td>-.08</td>
<td>.12</td>
<td>.07</td>
<td>.22</td>
<td>.08</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>De</td>
<td>-.04</td>
<td>-.07</td>
<td>-.01</td>
<td>-.03</td>
<td>-.06</td>
<td>-.05</td>
<td>-.02</td>
<td>-.09</td>
<td>-.01</td>
<td>-.03</td>
<td>.08</td>
<td>.08</td>
</tr>
</tbody>
</table>

5.1. 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, 11, 12, 13, 17]. The gamma distribution is defined on $(0, \infty)$ by the formula

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

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

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

and

$\beta = (56.40, 86.75, 74.60, 94.70, 95.97, 97.64, 81.56, 78.12, 49.33, 52.31, 57.29, 43.92)$. 
5.2. Simulating monthly rainfall. 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 \( x_r = F^{-1}(u_r) \).

5.3. 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

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

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

![Histogram of observed and generated rainfall for September](image)

**Figure 2.** September: Observed rainfall with fitted distribution (above) and with generated data (below).
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.

**Table 3.** Key statistics for observed, fitted and generated data

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>69.633</td>
<td>76.805</td>
</tr>
<tr>
<td>Fitted</td>
<td>69.633</td>
<td>76.805</td>
</tr>
<tr>
<td>Generated</td>
<td>69.795</td>
<td>76.394</td>
</tr>
</tbody>
</table>
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 $X = (X_1, \ldots, X_m)^T$ is distributed according to the joint probability density $g : (0, \infty)^m \to (0, \infty)$ defined by

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

where $i = (i_1, \ldots, i_m)$ and $I_i$ is the usual uniform subdivision of the unit hyper-cube $[0, 1]^m$ and $h = [h_i] \in \mathbb{R}^\ell$ 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 $K_i$ as the inverse image of $I_i$ under the mapping $F = (F_1, \ldots, F_m) : (0, \infty)^m \to (0, 1)^m$. We have

$$E[(S - \mu)^2] = \sum_{i \in \{1, \ldots, n\}^m} n^{m-1} h_i \int_{K_i} (S - \mu)^2 f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m$$

$$= \sum_{i \in \{1, \ldots, n\}^m} n^{m-1} h_i \int_{K_i} \left[ \sum_{r=1}^m (x_r - \mu_r)^2 + 2 \sum_{1 \leq r < s \leq m} (x_r - \mu_r)(x_s - \mu_s) \right] f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m.$$

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

$$\int_{K_i} \sum_{r=1}^m (x_r - \mu_r)^2 f_1(x_1) \cdots f_m(x_m) dx_1 \cdots dx_m = \frac{\sum_{r=1}^m \sigma_r(i_r)^2}{n^{m-1}}$$
Define an order for the indices where

\[ \ell \]

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

\[
\int_{c_r(k)}^{c_r(k+1)} (x_r - \mu_r)^2 f_r(x_r) dx_r
\]

for each \( r = 1, \ldots, m \) and each \( k = 1, 2, \ldots, n \). By noting that \( h \) is multiply-stochastic and summing over the relevant terms it follows that

\[
E[(S - \mu)^2] = \sum_{r=1}^{m} \sigma_r^2 + 2n \sum_{i \in \{1, \ldots, n\}^3} h_i \left( \sum_{1 \leq r < s \leq m} m_r(i_r)m_s(i_s) \right)
\]

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 Section 5.4.

### 5.4. Simulating Spring rainfall using a checkerboard copula

Suppose we have obtained a checkerboard copula \( C_h \) defined by a matrix \( h = [h_i] \in \mathbb{R}^{\ell} \) where \( \ell = n^3 \) and \( i = (i, j, k) \in \{1, \ldots, n\}^3 \) on a uniform partition \( \{I_i\} \) of the unit cube \( (0, 1)^3 \). Simulated data for monthly rainfall triples may be generated as follows. Define an order for the indices \( i = (i, j, k) \) 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

\[
(x_r, y_r, z_r) = \left( F_x^{-1}(u_r), F_y^{-1}(v_r), F_z^{-1}(w_r) \right)
\]

where \( F_x, F_y \) and \( F_z \) are the given marginal distributions.
5.4.1. 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

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

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

5.4.2. 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

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

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

5.5. Numerical calculations. We used MATLAB for the numerical calculations. The subintervals \(K_{ijk}\) are defined by

<table>
<thead>
<tr>
<th>(r)</th>
<th>(c_r(1))</th>
<th>(c_r(2))</th>
<th>(c_r(3))</th>
<th>(c_r(4))</th>
<th>(c_r(5))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>26.962</td>
<td>54.054</td>
<td>95.635</td>
<td>(\infty)</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>30.586</td>
<td>60.243</td>
<td>105.292</td>
<td>(\infty)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>33.207</td>
<td>65.553</td>
<td>114.750</td>
<td>(\infty)</td>
</tr>
</tbody>
</table>

and we calculate

<table>
<thead>
<tr>
<th>(r)</th>
<th>(m_r(1))</th>
<th>(m_r(2))</th>
<th>(m_r(3))</th>
<th>(m_r(4))</th>
<th>(\text{sum})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-13.730</td>
<td>-7.431</td>
<td>0.779</td>
<td>20.381</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>-14.970</td>
<td>-8.004</td>
<td>0.931</td>
<td>22.042</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>-16.355</td>
<td>-8.747</td>
<td>1.005</td>
<td>24.077</td>
<td>0.000</td>
</tr>
</tbody>
</table>

for the corresponding moments about the mean and

<table>
<thead>
<tr>
<th>(r)</th>
<th>(\sigma_r(1)^2)</th>
<th>(\sigma_r(2)^2)</th>
<th>(\sigma_r(3)^2)</th>
<th>(\sigma_r(4)^2)</th>
<th>(\sigma_r^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>767.330</td>
<td>236.020</td>
<td>37.637</td>
<td>2394.201</td>
<td>3435.189</td>
</tr>
<tr>
<td>2</td>
<td>913.258</td>
<td>274.367</td>
<td>44.796</td>
<td>2785.467</td>
<td>4017.888</td>
</tr>
<tr>
<td>3</td>
<td>1087.259</td>
<td>327.624</td>
<td>53.328</td>
<td>3325.776</td>
<td>4793.987</td>
</tr>
</tbody>
</table>
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 (18).

5.6. 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 \(h = [h_{ijk}]\). The probability
in simulation that a rainfall triple \((x_1, x_2, x_3)\) will be selected from the interval
\(K_{ijk} = F^{-1}(I_{ijk})\) 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

\[
\begin{align*}
    h_1 &\approx \begin{bmatrix}
            0.1039 & 0.0752 & 0.0514 & 0.0331 \\
            0.0870 & 0.0710 & 0.0588 & 0.0466 \\
            0.0589 & 0.0623 & 0.0628 & 0.0607 \\
            0.0422 & 0.0535 & 0.0651 & 0.0760 \\
\end{bmatrix},
    h_2 &\approx \begin{bmatrix}
            0.0938 & 0.0718 & 0.0517 & 0.0356 \\
            0.0732 & 0.0680 & 0.0603 & 0.0501 \\
            0.0547 & 0.0611 & 0.0659 & 0.0667 \\
            0.0392 & 0.0530 & 0.0682 & 0.0855 \\
\end{bmatrix},
    h_3 &\approx \begin{bmatrix}
            0.0841 & 0.0690 & 0.0525 & 0.0386 \\
            0.0678 & 0.0664 & 0.0615 & 0.0544 \\
            0.0502 & 0.0594 & 0.0681 & 0.0729 \\
            0.0369 & 0.0525 & 0.0721 & 0.0943 \\
\end{bmatrix},
    h_4 &\approx \begin{bmatrix}
            0.0755 & 0.0645 & 0.0532 & 0.0415 \\
            0.0610 & 0.0628 & 0.0628 & 0.0592 \\
            0.0453 & 0.0583 & 0.0697 & 0.0792 \\
            0.0342 & 0.0516 & 0.0744 & 0.1050 \\
\end{bmatrix},
\end{align*}
\]

for the copula of maximum entropy and

\[
\begin{align*}
    h_1 &\approx \begin{bmatrix}
            0.1070 & 0.0719 & 0.0533 & 0.0329 \\
            0.0768 & 0.0695 & 0.0596 & 0.0463 \\
            0.0605 & 0.0642 & 0.0630 & 0.0580 \\
            0.0408 & 0.0536 & 0.0629 & 0.0774 \\
\end{bmatrix},
    h_2 &\approx \begin{bmatrix}
            0.0953 & 0.0685 & 0.0539 & 0.0356 \\
            0.0705 & 0.0671 & 0.0629 & 0.0531 \\
            0.0554 & 0.0625 & 0.0646 & 0.0655 \\
            0.0383 & 0.0546 & 0.0668 & 0.0862 \\
\end{bmatrix},
    h_3 &\approx \begin{bmatrix}
            0.0870 & 0.0678 & 0.0545 & 0.0382 \\
            0.0648 & 0.0658 & 0.0619 & 0.0563 \\
            0.0527 & 0.0621 & 0.0667 & 0.0712 \\
            0.0353 & 0.0535 & 0.0686 & 0.0948 \\
\end{bmatrix},
    h_4 &\approx \begin{bmatrix}
            0.0768 & 0.0642 & 0.0542 & 0.0411 \\
            0.0575 & 0.0625 & 0.0630 & 0.0603 \\
            0.0473 & 0.0614 & 0.0689 & 0.0775 \\
            0.0335 & 0.0534 & 0.0717 & 0.1070 \\
\end{bmatrix},
\end{align*}
\]

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 pro-
cede 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
Table 4. Selected years from a typical simulation using the maximum entropy copula

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

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

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

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 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 \( e \) 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 \( \|e\| \) 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

\[
E[e_r^2] = E \left[ \left( p_r - \frac{N_r}{N} \right)^2 \right]
= \sum_{N_1 + \cdots + N_\ell = N} \left( p_r - \frac{N_r}{N} \right)^2 \left( \begin{array}{c} N \\ N_1 \cdots N_\ell \end{array} \right) p_1^{N_1} \cdots p_\ell^{N_\ell}
= p_r \left( 1 - p_r \right) \frac{1}{N}
\]
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 a period of 15000 years, for the independent model.
Table 6. Statistics and probability errors for typical simulations (150 years)

<table>
<thead>
<tr>
<th>run</th>
<th>me mean</th>
<th>me var</th>
<th>e</th>
<th>n mean</th>
<th>n var</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>218.051</td>
<td>13740.47</td>
<td>0.081178</td>
<td>228.886</td>
<td>13751.66</td>
<td>0.077355</td>
</tr>
<tr>
<td>2</td>
<td>242.358</td>
<td>20593.60</td>
<td>0.083936</td>
<td>229.887</td>
<td>12051.58</td>
<td>0.079029</td>
</tr>
<tr>
<td>3</td>
<td>235.513</td>
<td>14223.73</td>
<td>0.079674</td>
<td>232.618</td>
<td>13684.20</td>
<td>0.068920</td>
</tr>
<tr>
<td>4</td>
<td>224.125</td>
<td>14223.73</td>
<td>0.083936</td>
<td>229.887</td>
<td>12051.58</td>
<td>0.079029</td>
</tr>
<tr>
<td>5</td>
<td>227.902</td>
<td>15024.04</td>
<td>0.086871</td>
<td>244.082</td>
<td>14326.78</td>
<td>0.092515</td>
</tr>
<tr>
<td>6</td>
<td>231.622</td>
<td>15137.70</td>
<td>0.089484</td>
<td>221.677</td>
<td>13538.77</td>
<td>0.080973</td>
</tr>
<tr>
<td>7</td>
<td>227.888</td>
<td>12985.29</td>
<td>0.074645</td>
<td>238.924</td>
<td>14163.22</td>
<td>0.087583</td>
</tr>
<tr>
<td>8</td>
<td>230.708</td>
<td>14399.34</td>
<td>0.096069</td>
<td>219.864</td>
<td>16468.45</td>
<td>0.081098</td>
</tr>
<tr>
<td>9</td>
<td>231.820</td>
<td>15105.23</td>
<td>0.081699</td>
<td>218.375</td>
<td>13739.63</td>
<td>0.079944</td>
</tr>
<tr>
<td>10</td>
<td>226.928</td>
<td>11977.45</td>
<td>0.092291</td>
<td>241.420</td>
<td>15304.71</td>
<td>0.085598</td>
</tr>
</tbody>
</table>

and the two correlated models. Although the histograms are relatively 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 close to the theoretical values. The histogram for the independent simulation is slightly taller and narrower but 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.

For the copula of maximum entropy and the normal copula we ran repeated simulations covering a period of 15000 years. The summary statistics are shown in Table 7. In moving from simulations with 150 realizations shown in Table 6 to simulations with 15000 realizations shown in Table 7 the probability errors are reduced by one order of magnitude.

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

<table>
<thead>
<tr>
<th>run</th>
<th>me mean</th>
<th>me var</th>
<th>e</th>
<th>n mean</th>
<th>n var</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>231.451</td>
<td>14539.69</td>
<td>0.007605</td>
<td>229.735</td>
<td>14430.86</td>
<td>0.008110</td>
</tr>
<tr>
<td>2</td>
<td>230.754</td>
<td>14560.60</td>
<td>0.007935</td>
<td>229.370</td>
<td>14166.38</td>
<td>0.010244</td>
</tr>
<tr>
<td>3</td>
<td>230.372</td>
<td>14340.78</td>
<td>0.008827</td>
<td>228.909</td>
<td>14249.83</td>
<td>0.009168</td>
</tr>
<tr>
<td>4</td>
<td>228.611</td>
<td>12424.27</td>
<td>0.007842</td>
<td>230.913</td>
<td>14271.26</td>
<td>0.007875</td>
</tr>
<tr>
<td>5</td>
<td>229.524</td>
<td>14139.78</td>
<td>0.006840</td>
<td>230.635</td>
<td>14313.10</td>
<td>0.008299</td>
</tr>
<tr>
<td>6</td>
<td>229.238</td>
<td>14384.39</td>
<td>0.008706</td>
<td>230.539</td>
<td>14430.73</td>
<td>0.007616</td>
</tr>
<tr>
<td>7</td>
<td>230.030</td>
<td>14479.19</td>
<td>0.007641</td>
<td>230.329</td>
<td>14519.17</td>
<td>0.008243</td>
</tr>
<tr>
<td>8</td>
<td>229.788</td>
<td>14089.51</td>
<td>0.007622</td>
<td>229.806</td>
<td>14365.04</td>
<td>0.007749</td>
</tr>
<tr>
<td>9</td>
<td>229.322</td>
<td>14379.86</td>
<td>0.008251</td>
<td>231.063</td>
<td>14388.86</td>
<td>0.009063</td>
</tr>
<tr>
<td>10</td>
<td>229.997</td>
<td>14035.33</td>
<td>0.007016</td>
<td>230.757</td>
<td>14406.03</td>
<td>0.007634</td>
</tr>
</tbody>
</table>

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 $3 \times 10^6$ years.

**Table 8. Comparison of three models for Spring Quarter rainfall**

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td>230.123</td>
<td>14391.34</td>
</tr>
<tr>
<td>independent (theory)</td>
<td>230.123</td>
<td>12247.06</td>
</tr>
<tr>
<td>independent (simulation)</td>
<td>230.155</td>
<td>12236.11</td>
</tr>
<tr>
<td>maximum entropy copula (theory)</td>
<td>230.123</td>
<td>14318.11</td>
</tr>
<tr>
<td>maximum entropy copula (simulation)</td>
<td>230.085</td>
<td>14319.45</td>
</tr>
<tr>
<td>normal copula (theory)</td>
<td>230.123</td>
<td>14348.46</td>
</tr>
<tr>
<td>normal copula (simulation)</td>
<td>230.058</td>
<td>14347.05</td>
</tr>
</tbody>
</table>

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.

6. 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. Our original idea was to use a copula of maximum entropy. The rationale for this choice was a desire to avoid unwarranted assumptions about the unobserved statistics. The corresponding tri-variate distribution is the most disordered distribution that preserves the prescribed marginal distributions and matches the observed grade correlation.

There were several reasons why we decided 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 normal checkerboard 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 $h$. 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 the numerical calculations for each method are of similar complexity [10] 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 CIRCUIT did not give 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 $v \in [0, 1]^m$. Map these points into $\mathbb{R}^m$ using the transformation $w = \Phi^{-1}(v) \leftrightarrow w_r = \Phi^{-1}(v_r)$. Choose the orthogonal matrix $P$ such that $\Lambda = P^T \Sigma P \leftrightarrow \Lambda P^T = \Sigma$ where $\Lambda$ is a positive diagonal matrix and rescale the points according to the transformation $y = \Lambda^{1/2} w$. Map the points
\( y \in \mathbb{R}^m \) onto points \( z = Py \in \mathbb{R}^m \). Then return the points to \([0,1]^m\) using the map \( u = \Phi(z) \iff u_r = \Phi(z_r) \). Thus

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

Now for each \( i \in \{1, 2, \ldots, n\}^m \) count the points \( p_i \) in the intervals \( I_i \) and hence calculate the relative proportion for every interval. The multiply-stochastic hypermatrix \( h \) is defined by \( h_i = \eta p_i \) for all \( i \in \{1, 2, \ldots, n\}^m \). The rationale is that the points \( v \) are independently and evenly distributed but the points \( u \) are distributed according to the correlation specified by \( \varphi \). In mathematical terms we have

\[
P[u \in I_i] = \int_{I_i} \frac{\varphi(P^{-1}(u_1), \ldots, P^{-1}(u_m))}{\varphi(P^{-1}(u_1)) \cdots \varphi(P^{-1}(u_m))} \, du = \int_{\Phi^{-1}(I_i)} \varphi(z) \, dz = \int_{P^T \Phi^{-1}(I_i)} \varphi(Py) \, dy
\]

since \( \det P = 1 \) and hence

\[
P[u \in I_i] = \int_{\Lambda^{-1/2} P^T \Phi^{-1}(I_i)} \varphi(P \Lambda^{1/2} w) \det \Lambda^{1/2} \, dw = \int_{\Lambda^{-1/2} P^T \Phi^{-1}(I_i)} \frac{\exp[-w^T \Lambda^{1/2} P^T \Sigma^{-1} \Lambda^{1/2} w/2]}{(2\pi)^m/2 (\det \Sigma)^{1/2}} \det \Lambda^{1/2} \, dw = \int_{\Lambda^{-1/2} P^T \Phi^{-1}(I_i)} \frac{1}{(2\pi)^m/2} \exp[-w^T w/2] \, dw
\]

where we have used the fact that \( \det \Sigma = \det \Lambda \). Now it follows that

\[
P[u \in I_i] = \int_{\Lambda^{-1/2} P^T \Phi^{-1}(I_i)} \varphi(w_1) \cdots \varphi(w_m) \, dw = \int_{\Phi[\Lambda^{-1/2} P^T \Phi^{-1}(I_i)]} dv = V \left( \Phi[\Lambda^{-1/2} P^T \Phi^{-1}(I_i)] \right).
\]

In order to calculate the required tri-variate normal copula to our desired accuracy in the above case study we chose 256\(^3\) equally spaced points in the unit cube. We used an elementary MATLAB program to make the calculations. For higher dimensional examples the numerical calculations for the normal copula may be more challenging.

**Acknowledgements.** This research was supported in part by ARC Discovery Grant DP1096551 held by Phil Howlett and Charles Pearce. We would also like to acknowledge the use of publicly available data supplied by the Australian Bureau of Meteorology.

**REFERENCES**


Received September 2011; revised March 2012.

E-mail address: julia.piantadosi@unisa.edu.au
E-mail address: phil.howlett@unisa.edu.au
E-mail address: jonathan.borwein@newcastle.edu.au
E-mail address: john@daa.com.au