Entropy **2013**, *xx*, 1-x; doi:10.3390/——

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

MODELLING AND SIMULATION OF SEASONAL RAINFALL USING THE PRINCIPLE OF MAXIMUM ENTROPY

Jonathan Borwein [1], **Phil Howlett** [2,*] **and Julia Piantadosi** [2]

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

[2] Scheduling and Control Group (SCG), Centre for Industrial and Applied Mathematics (CIAM) and Barbara Hardy Institute (BHI), University of South Australia, Mawson Lakes 5095, Australia.

* Author to whom correspondence should be addressed; Email: p.howlett@unisa.edu.au, Telephone: +61 8 8302 3195, Fax: +61 8 8302 5785.

Received: xx / Accepted: xx / Published: xx

Abstract: We use the principle of maximum entropy to propose a parsimonious model for generation of simulated rainfall during the wettest three-month season at a typical location on the east coast of Australia. The model uses a checkerboard copula of maximum entropy to model the joint probability distribution for total seasonal rainfall and a set of two-parameter gamma distributions to model each of the marginal monthly rainfall totals. The model allows us to match the grade correlation coefficients for the checkerboard copula to the observed Spearman rank correlation coefficients and hence provides a model that correctly describes the mean and variance for each of the monthly totals and also for the overall seasonal total. Thus we avoid the need for *a posteriori* adjustment of simulated monthly totals in order to correctly simulate the observed seasonal statistics.

Detailed results are presented for modelling and simulation of seasonal rainfall at the town of Kempsey on the mid-north coast of New South Wales. Empirical evidence from extensive simulations is used to validate this application of the model. A similar analysis for Sydney is also described.

Keywords: maximum entropy; checkerboard copula; rainfall modelling and simulation.

Classification: MSC 15B51; 60E05; 65C05; 52A41

---

1. Introduction

We propose a model for seasonal rainfall using the principle of maximum entropy. To demonstrate our model we used official records from the Australian Bureau of Meteorology for station 059017 (Wide Street) at Kempsey in New South Wales during the period 1889 to 2011. The town of Kempsey is a typical location on the east coast of Australia with a humid sub-tropical or temperate climate and no pronounced dry season. The Köppen classification is *Cfa*. Although the annual rainfall is relatively high and is generally regarded as reliable there is ample historical evidence of extended periods with well below average rainfall.

There is consensus amongst climate scientists that summer and autumn rainfall in eastern Australia is influenced on a recurring basis by the quasi-periodic seasonal climatic events El Niño and La Niña. During El Niño rainfall is inhibited and during La Niña it is enhanced. It is therefore not especially surprising to find positive correlation for monthly rainfall at Kempsey during the period February-March-April—the wettest time of the year. Our aim is to construct a parsimonious model for a vector-valued random variable $\boldsymbol{X} = (X_1, X_2, X_3) \in \mathbb{R}^3$ that can be used to simulate typical monthly rainfall time series for February-March-April at Kempsey. We will show that the key observed seasonal statistics lie well within the commonly accepted empirical confidence intervals established by repeated simulations with our proposed model. Our results also show that even seemingly significant trends in the observed data could be due to chance alone.

2. A brief literature review

A comprehensive review of the vast array of relevant literature is neither feasible nor helpful. Instead we shall be content to refer only to articles of fundamental historical significance or to those that are directly relevant to the methods used in this paper.

The topic of entropy has a long and distinguished research history dating back to the fundamental principles of thermodynamics proposed by Rudolf Clausius in 1855. The principle of maximum entropy, enunciated much later by the physicist E. T. Jaynes [7,8] in 1957, is a recurring theme in our discussion. We apply this principle to both discrete and continuous entropy. The modern notion of discrete entropy [12] was introduced by John von Neumann in his 1927 treatise on quantum mechanics in which he defined the entropy of a statistical operator $\rho = \{p_n, \psi_n\}_n$ where $p_n > 0$ and $\sum_n p_n = 1$ and where $\{\psi_n\}$ is a complete orthonormal system of basis vectors as the weighted ensemble average $S(\rho) = -k \langle \rho \log_e \rho \rangle_n = -k \operatorname{Tr}(\rho \log_e \rho) = -k \sum_n p_n \log_e p_n$. See [21, pp 348–353] for more details.

This measure was adopted in 1948 by C. E. Shannon [16] as a measure of information in the theory of communication systems. Shannon also introduced the analogous notion of continuous or differential entropy $S(f) = -\int_\Omega f(x) \log_e f(x) dx$ where $f(x) \geq 0$ and $\int_\Omega f(x) dx = 1$ for continuous probability distributions. The entropy of a system is commonly described as a measure of the inherent disorder within the system. Entropy is maximized when the system is in the highest possible state of disorder. For a system with a finite number of possible states the entropy is maximized when all probabilities are equal.

By contrast the topic of rainfall modelling has a much more recent research history. The early work, such as the paper by Stern and Coe [19], follows a classical style that is typical of the physical sciences.

However the focus has shifted in recent times to a more pragmatic approach that is less concerned with a logical axiomatic basis and more concerned with a positive utilitarian outcome. The most relevant recent paper in relation to our work is the comprehensive 2005 report to the Australian Cooperative Research Centre for Catchment Hydrology by Srikanthan [18]. We shall discuss this paper in some detail. Although the report describes a successful scheme for generation of daily rainfall data at multiple sites a substantive difficulty emerges in the accumulation of simulated daily rainfall totals. This difficulty lies at the very heart of the problem that we will address. At each site Srikanthan uses a simple two by two Markov chain to generate realistic sequences of wet and dry days. On the designated wet days a two-parameter gamma distribution is invoked to generate rainfall depths. However Srikanthan makes the following critical observation.

> The generated daily rainfall amounts when aggregated into monthly and annual totals will not in general preserve the monthly and annual characteristics.

Hence it is necessary to implement a nested correction process. In the first place the generated monthly total $\widetilde{X_i}$ for the current month i obtained by summing the daily totals is modified to produce a corrected monthly total X_i according to the iterative formula

$$\frac{X_i - \mu_i}{\sigma_i} = \rho_i \frac{X_{i-1} - \mu_{i-1}}{\sigma_{i-1}} + (1 - \rho_i^2)^{1/2} \frac{\widetilde{X_i} - \nu_i}{\tau_i}$$

where, in our simplified notation, μ_i is the observed mean and σ_i is the observed standard deviation for month i, ρ_i is the observed correlation between the normalized totals for the current month i and the previous month $i-1$, and ν_i is the theoretical mean and τ_i is the theoretical standard deviation for the sum of the generated daily totals in month i. The iterative process relies on knowledge of the previous corrected total but Srikanthan does not say how the initial correction is obtained. The generated daily totals are now multiplied by a factor $X_i/\widetilde{X_i}$ to ensure that the corrected monthly total is obtained by summing the corrected daily totals. In the second place a similar correction is necessary when converting the monthly totals to annual totals with a further corresponding correction required for the daily totals.

While this is an eminently sensible correction process—effectively a top-down approach based on the primary importance of the annual distribution—it also recognises a fundamental problem with the original model in which correlations in the daily rainfall, although undoubtedly very small, are apparently ignored. It is known that the introduction of a single correlation parameter allows one to adjust the standard deviation for the sum of the daily totals in a month so that it matches the observed monthly standard deviation. See for instance the correlative coherence analysis proposed by Getz [4]—which, incidentally, makes direct use of the Shannon entropy—and the model proposed by Hasan and Dunn [5] which uses a Tweedie distribution. It is perverse, in retrospect, to select a gamma distribution that will generate realistic daily rainfall depths if one intends, subsequently, to modify the generated data. This inevitably means that the modified daily rainfall depth distributions will be biassed relative to the observed distributions.

While we acknowledge the utility and undoubted practicality of the Srikanthan model we are concerned about what appears to be an *ad hoc* theoretical basis. There is no clear statement about the relationship between the sample measurements and the *a priori* criteria that will be used to judge goodness-of-fit. In this regard there is one further point we wish to make. In order to embed spatial

correlation into the rainfall generation process Srikanthan uses a multivariate normal distribution to generate appropriately correlated sequences of random numbers at the various sites. This is equivalent to using a multivariate normal copula to construct a joint distribution that preserves the marginal single site distributions. Srikanthan points out that there is no known theoretical method to calculate the corresponding grade correlation coefficients from the correlation matrix and hence concedes that it is necessary to use a process of iterative adjustment to find the correct correlation matrix. This is one more reason why calculation of a checkerboard copula of maximum entropy—as we propose in this paper—may be a more computationally efficient way to generate the required correlations. Since there are many other copulas that could be chosen for this task it would also seem to be good practice to provide some rationale for the choice.

There is a large number of other papers that we could legitimately cite but we mention only a few. For a more comprehensive review we refer to Srikanthan and McMahon [17] and to an earlier review by Wilks and Wilby [20]. The over-dispersion phenomenon that bedevils the Srikanthan model was studied by Katz and Parlange [9] who suggested that higher order Markov models can reduce apparent discrepancies in the number of generated wet days and the number of observed wet days. Rosenberg *et al* [15] constructed a joint density using Laguerre series to incorporate correlation between successive months and hence correct the seasonal variance but the optimal parametric structure of this model is unclear. Hasan and Dunn [5] have recently used a Tweedie distribution to model monthly rainfall. The model combines a Poisson process to generate wet and days and a collection of correlated gamma distributions to model daily rainfall depth. There is insufficient freedom in this model to match individual daily correlations but it is possible to adjust the correlation parameters so as to avoid the over-dispersion problem.

3. First experiment: comparing the observed seasonal rainfalls with simulated observations for a stationary time series at various timescales.

In practice it may be possible to observe only a finite number of terms $\{x_i\}_{i=1}^{N}$ in a doubly-infinite time series $\{x_i\}_{i=-\infty}^{\infty}$. In such cases we may define the moving average $\{x_i(k)\}_{i=1}^{N_k}$ where $N_k = N-k+1$ at scale k by

$$x_i(k) = \frac{1}{k}\sum_{\ell=i}^{i+k-1} x_\ell$$

for each $i = 1, \ldots, N_k$ and $k = 1, \ldots, N$. For each fixed time frame $s \leq N$ we can also define the mean $\{\mu_s(k)\}_{k=1}^{N_s}$ and standard deviation $\{\sigma_s(k)\}_{k=1}^{N_s}$ for time frame s at scale k by

$$\mu_s(k) = \frac{1}{s}\sum_{i=1}^{s} x_i(k) \quad \text{and} \quad \sigma_s(k)^2 = \frac{1}{s}\sum_{i=1}^{s}(x_i(k) - \mu_s(k))^2.$$

If the time series $\{x_i\}_{i=-\infty}^{\infty}$ is *stationary in the wide sense* then

1. $E[\mu_s(k)] = \mu$ for all $k = 1, \ldots, N_s$; and
2. $E[\sigma_s(k)^2] = R_s(k)$ for all $k = 1, \ldots, N_s$

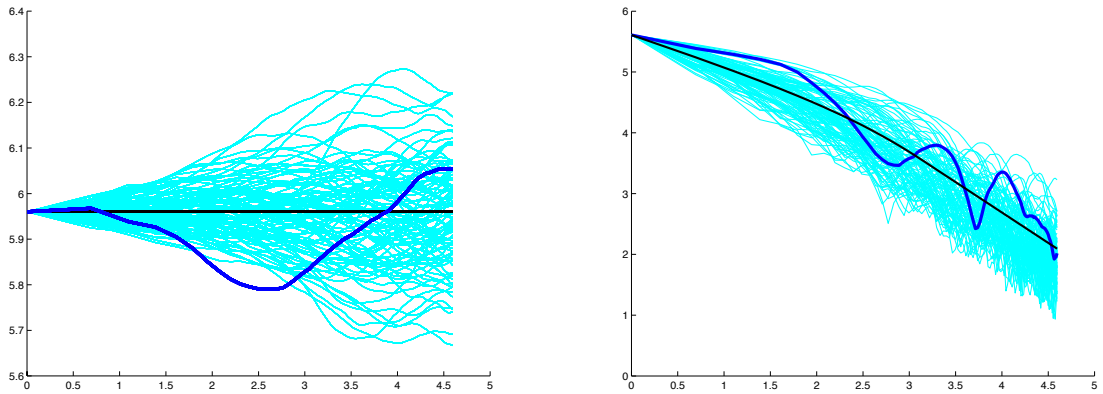
for each fixed time frame $s \geq 1$. We cannot test these properties directly for a partially observed time series $\{x_i\}_{i=1}^N$. Indeed Koutsoyannis [10] argues that it is difficult to tell, from a single realization, whether an observed time series is stationary. Nevertheless, he suggests that it is useful to compare the standard deviation at scale k for the partially observed time series to the adjusted standard deviation at scale k for simulated observations of a known stationary time series over the same period.

Consequently, we compared the observed time series $\{x_i(k)\}_{i=1}^N$ of seasonal rainfall totals at Kempsey, where $N = 123$ years, with a large number of independently generated simulated observed time series $\{u_i\}_{i=1}^N$ of uniformly distributed uncorrelated pseudo-random numbers. For a timeframe of $s = 25$ years we compared the graphs of the log mean $\{\log_e \mu_s(k)\}_{k=1}^{N_s}$ and log standard deviation $\{\log \sigma_s(k)\}_{k=1}^{N_s}$ at scale k against $\log_e k$ for the observed rainfall to the corresponding graphs of the adjusted log mean $\{\log_e c_s\, \nu_s(k)\}_{k=1} N_s$ where $c_s = \mu_s(1)/\nu_s(1)$, and adjusted log standard deviation $\{\log_e d_s\, \sigma_s(k)\}_{k=1}^{N_s}$ where $d_s = \sigma_s(1)/\tau_s(1)$ for the pseudo-random numbers. The pseudo-random numbers were generated in MATLAB.

If the partially observed time series of seasonal rainfalls is a wide sense stationary time series of independent random numbers with mean μ and standard deviation σ then for time frame s the expected values $E[\mu_s(k)]$ and $E[\sigma_s(k)^2]$ at scale k are given by

$$E[\mu_s(k)] = \mu, \quad E[\sigma_s(k)^2] = \begin{cases} \dfrac{1}{s-1}\left[\dfrac{s}{k} - 1 + \left(k - \dfrac{1}{k}\right)\dfrac{1}{3s}\right]\sigma^2 & \text{for} \quad 1 \leq k \leq s \\ \\ \dfrac{1}{3k^2}(s+1)\sigma^2 & \text{for} \quad s \leq k \leq N_s - 1. \end{cases} \tag{1}$$

Figure 1. Left: Plots of $\log_e \mu_s(k)$ against $\log_e k$ (blue graph) showing corresponding graphs for 100 simulated series of pseudo-random numbers (light blue graphs) and $\log_e E[\mu_s(k)]$ for a stationary time series (black graph). Right: Plots of $\log_e \sigma_s(k)$ against $\log_e k$ (blue graph) showing corresponding graphs for 100 simulated series of pseudo-random numbers (light blue graphs) and $\frac{1}{2}\log_e E[\sigma_s(k)^2]$ for a stationary time series (black graph).



The graphs in Figure 1 show that the mean and standard deviation at scale k for the series of seasonal rainfall totals $\{x_i\}_{i=1}^N$ exhibit similar behaviour to that of the corresponding adjusted mean and standard

deviation for the simulated observations $\{u_i\}_{i=1}^N$ of a stationary time series of uncorrelated pseudo-random numbers. We conclude that it is reasonable to model the observed seasonal rainfall as a stationary time series and hence deduce that the distribution of observed values can be modelled by a real-valued random variable.

4. First theoretical principle: the gamma distribution is the maximum entropy model for a random variable defined only by a finite number of strictly positive observations

In his influential 1957 papers the physicist E. T. Jaynes [7,8] wrote down the following general principle—now known as the principle of maximum entropy.

> In making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiassed assignment we can make; to use any other would amount to arbitrary assumption of information which, by hypothesis, we do not have.

We will use this principle to argue that the gamma distribution is the best model to represent the distribution of a random variable X provided the means of the partially observed totals $\{x_i\}_{i=1}^N$ and of the natural logarithm of the partially observed totals $\{\log_e x_i\}_{i=1}^N$ are both well-defined and finite. This is true if we assume that the observed totals $\{x_n\}_{n=1}^N$ are strictly positive.

It is useful to outline the mathematical argument. We wish to find a probability density $f : (0, \infty) \to [0, \infty)$ such that the differential entropy

$$h(f) = (-1) \int_0^\infty f(x) \log_e f(x) dx \tag{2}$$

is maximized subject to the additional constraints imposed by the observed means

$$E[X] = \overline{x} = \frac{1}{N} \sum_{n=1}^N x_n \quad \text{and} \quad E[\log_e X] = \overline{\log_e x} = \frac{1}{N} \sum_{n=1}^N \log_e x_n. \tag{3}$$

We can formulate this problem as a convex optimization with linear constraints. From the theory of Fenchel duality and the Fenchel-Young inequality ([3], pp. 171-178) we have

$$\begin{aligned} p &= \inf_{f \in L^1[(0,\infty)]} \left\{ -h(f) - 1 \mid E[1] = 1, E[X] = \overline{x}, E[\log_e X] = \overline{\log_e x} \right\} \\ &\geq \sup_{(\alpha,\beta,\kappa) \in \mathbb{R}^3} \left\{ \log_e \kappa - \overline{x}/\beta + (\alpha - 1)\overline{\log_e x} - \kappa \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx \right\} \\ &= \sup_{(\alpha,\beta,\kappa) \in \mathbb{R}^3} \left\{ \log_e \kappa - \overline{x}/\beta + (\alpha - 1)\overline{\log_e x} - \kappa\, \Gamma(\alpha) \beta^\alpha \right\} \\ &= \sup_{(\alpha,\beta,\kappa) \in \mathbb{R}^3} \varphi(\alpha, \beta, \kappa) \\ &= -\log_e[\Gamma(\alpha)\beta] + (\alpha - 1)\psi(\alpha) - (\alpha + 1) = d \end{aligned} \tag{4}$$

where the parameters α, β and $\kappa = \kappa(\alpha, \beta)$ are determined by the equations

$$\log_e \beta + \psi(\alpha) = \overline{\log_e x}, \quad \alpha\beta = \overline{x}, \quad \kappa = \frac{1}{\Gamma(\alpha)\beta^\alpha} \tag{5}$$

and where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. The supremum and the conditions (5) are found simply by solving the equations $\partial\varphi/\partial\alpha = 0$, $\partial\varphi/\partial\beta = 0$ and $\partial\varphi/\partial\kappa = 0$. The function

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

which arises naturally in (4) when solving the dual optimization problem to find d, is the probability density on $(0,\infty)$ for the gamma distribution with parameters α and β. If X is a random variable with this distribution we write $X \sim \Gamma(\alpha,\beta)$. We will also use the notation

$$F_{\alpha,\beta}(x) = \int_0^x f_{\alpha,\beta}(\xi)d\xi$$

for $x > 0$ to denote the corresponding cumulative probability distribution. In the case where α and β are determined by (5) then the additional constraints (3) are also satisfied. Since it is easy to show that $-h(f_{\alpha,\beta}) - 1 = d$ it follows that $p = d$ and that $f_{\alpha,\beta}$ is the unique solution to our original convex optimization problem.

We conclude that the gamma distribution with parameters α and β determined by solving the equations (5) is the least ordered or least prescriptive probability distribution on $(0,\infty)$ satisfying the given constraints. It is pleasing that the equations (5) are also the maximum likelihood equations used to estimate α and β if one has decided *a priori* to fit a gamma distribution to the observed values $\{x_n\}_{n=1}^N$.

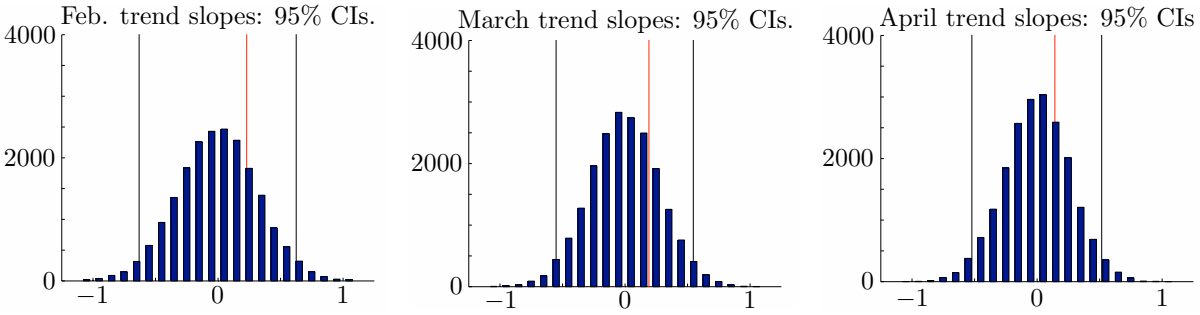
5. Second experiment: a comparison of the trend-slopes for the observed time series of monthly rainfall totals and the trend-slopes for simulated time series of monthly rainfall totals generated by the appropriate maximum likelihood gamma distributions.

The observed monthly rainfall totals for February, March and April at Kempsey for the period 1889 to 2011 are all strictly positive. Our tests so far suggest that we may reasonably propose the following null hypothesis—that the observed monthly rainfall totals at Kempsey in February, March and April can be modelled as outcomes of the stationary random variables $X_i \sim \Gamma(\alpha_i,\beta_i)$ for each $i = 1,2,3$ where

$$\boldsymbol{\alpha} = (1.5502, 2.0134, 1.2735) \quad \text{and} \quad \boldsymbol{\beta} = (100.4753, 77.2556, 91.1034).$$

We tested the null hypothesis by a linear regression on the observed time series of monthly rainfall totals and on each of 20000 simulated time series of monthly rainfall totals over the same period of 123 years generated by the proposed gamma distributions. Let $\{r_i(t)\}$ denote the rainfall in month i and year t. We used MATLAB to find (p_i, q_i) such that $\sum_{t=1}^{123} |r_i(t) - (p_i t + q_i)|^2$ is minimized. The slope p_i of this line is the trend-slope. In each case we found that the trend-slope of the observed rainfall data sets lay well within the empirical 95% confidence intervals for the trend-slopes of the simulated rainfall data sets. The trend-slopes for the observed data sets and the corresponding 95% confidence intervals for the trend-slopes of the simulated data sets were $p_1 = 0.22 \in [-0.63, 0.63]$ in February, $p_2 = 0.19 \in [-0.55, 0.55]$ in March and $p_3 = 0.14 \in [-0.52, 0.52]$ in April. We conclude that there is insufficient evidence to reject the hypothesis that the observed monthly rainfall totals $\{x_{i,j}\}_{j=1,\ldots,N}$ are the outcomes of a stationary random variable $X_i \sim \Gamma(\alpha_i,\beta_i)$ for each $i = 1,2,3$. This means the apparent observed trends could reasonably be regarded as due to chance alone. The results of our simulations are shown in Figure 2.

Figure 2. Trend-slope histograms for 20000 simulated rainfall data sets generated by maximum likelihood gamma distributions for February (left), March (centre) and April (right). The vertical red lines show the observed trend-slopes lying inside the empirical 95% confidence intervals for the simulated trend-slopes.



Since $r_i(t) \approx p_i t + q_i$ and since $p_i > 0$ for each $i = 1, 2, 3$ one could possibly argue that all observed trend-slopes are positive and that a null hypothesis of a positive trend slope in each case is more appropriate. If so a more complex time dependent model for both monthly and seasonal rainfall would be required. Although most climate scientists expect rainfall events in eastern Australia to become more extreme we believe there is currently no firm agreement about such rainfall trends.

6. Third experiment: using Q-Q plots to test the *goodness-of-fit* for the simulated time series of monthly rainfalls generated by the maximum likelihood gamma distributions to the observed time series of monthly rainfalls

We demonstrate the *goodness-of-fit* for the observed monthly rainfall data to the designated gamma distributions using Q-Q plots.

Firstly we used the designated gamma distribution to generate 1000 simulated data sets. Then we plotted the simulated quantiles against the theoretical quantiles. The results are shown in Figure 3. These plots show the full range of random variation that one should expect from observations of the designated gamma random variable from 1000 samples each of size $N = 123$. By discarding the bottom 25 and top 25 values for each quantile from the simulated data sets we found empirical 95% confidence intervals.

Secondly, we plotted the observed quantiles against the theoretical quantiles for the gamma distribution. The results are shown in Figure 4. We used grey bars on these plots to show the empirical 95% confidence intervals for the simulated quantiles. In all but one of twenty four cases the observed values lie within the desired intervals. Thus, on the basis of the Q-Q plots, there are no recognised statistical grounds to reject the hypothesis that the monthly rainfall totals can be modelled by the designated gamma distributions.

Remark 6.1. We have been taken to task by some in the engineering community for not comparing our proposed model to other models currently in popular use. Nevertheless, our model is based on a well-established scientific methodology. We use the principle of maximum entropy to argue that the

maximum likelihood gamma distribution is the most appropriate model for rainfall accumulations where the observed totals are strictly positive.

Thus, we use maximum likelihood to calculate the relevant parameter values and then test our model rigorously and impartially against the observed values according to accepted statistical wisdom. The conclusion is clear—there are no reasonable statistical grounds for rejecting the model. The argument that other models may provide a better fit to the observed data is essentially scientific nonsense. Indeed this criticism embraces a fundamental misconception that an observed sample is always a true representation of the entire population. Moreover, the suggested iterative correction methods used by Srikanthan and others are subject to concerns about overfitting [1].

A legitimate criticism of our model would need to argue either that the principle of maximum entropy is inappropriate or else that we should use more extensive data measurements.

Figure 3. Q-Q plots of quantiles for 1000 simulated data sets with each set covering a period of 123 years versus corresponding theoretical quantiles for $X_i \sim \Gamma(\alpha_i, \beta_i)$ at Kempsey in February (left), March (centre) and April (right).

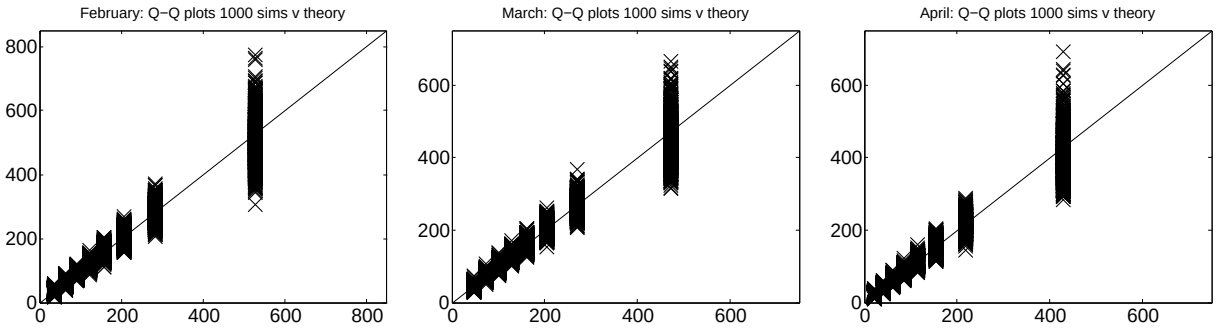
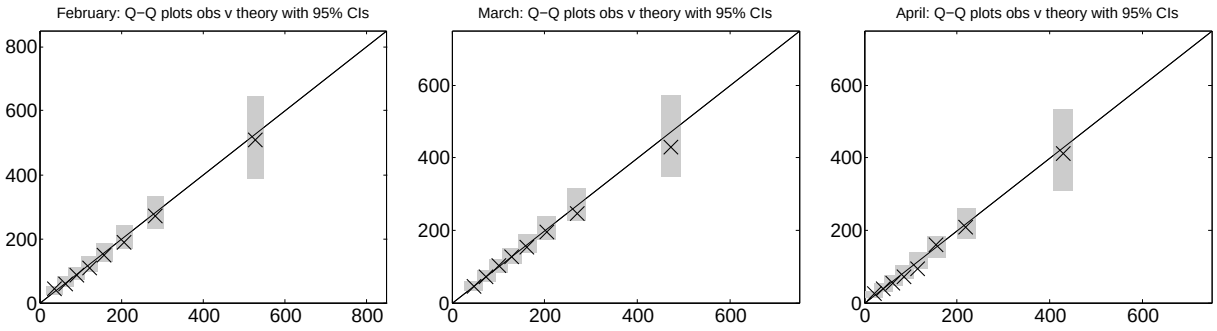


Figure 4. Q-Q plots for observed quantiles versus theoretical quantiles for $X_i \sim \Gamma(\alpha_i, \beta_i)$ at Kempsey in February (left), March (centre) and April (right). The vertical grey bars show empirical 95% confidence intervals for simulated quantiles generated by $X_i \sim \Gamma(\alpha_i, \beta_i)$.



7. Second theoretical principle: finding a checkerboard copula of maximum entropy to construct a joint density for seasonal rainfall that matches the observed marginal rank correlation coefficients and preserves the desired marginal monthly distributions

The next step in the modelling process is to construct a joint probability distribution for the entire three-month time period. We will do this in what we believe is the most natural way—by using the principle of maximum entropy. Past studies of rainfall accumulations over several months [9,15] have concluded that the variance of the simulated time series of seasonal rainfall totals generated by models with independent marginal distributions is often not consistent with the observed variance. Since the observed data at Kempsey shows positive correlation for February–March–April we expect the observed variance in seasonal rainfall to be higher than one would find with independent marginal monthly distributions.

Our aim will be to construct a joint distribution that not only preserves the desired monthly rainfall characteristics but also replicates the observed variance in the seasonal rainfall totals. We will do this using a checkerboard copula of maximum entropy. Once again we will proceed on the basis that our model should satisfy the fundamental physical requirements—that the marginal distributions are preserved and that the joint distribution is constrained by the observed rank correlation coefficients—and that we should test the model against the observed data using accepted statistical principles. In the first instance we need to show that the model satisfies the required constraints. In the second instance we wish to show that the variance of the simulated seasonal rainfall totals is consistent with the observed variance.

7.1. Copulas with prescribed grade correlation coefficients

An m-dimensional *copula*, where $m \geq 2$, is a cumulative probability distribution $C(\boldsymbol{u}) \in [0, \infty)$ defined on the m-dimensional unit hypercube $\boldsymbol{u} = (u_1, u_2, \ldots, u_m) \in [0,1]^m$ for a vector-valued random variable $\boldsymbol{U} = (U_1, U_2, \ldots, U_m)$ with uniform marginal distributions for $U_1, U_2, \ldots, U_m$. We refer to [11] for a full discussion. The correlation coefficients for the joint distribution are defined by

$$\rho_{r,s} = \frac{E[(U_r - 1/2)(U_s - 1/2)]}{\sqrt{E[(U_r - 1/2)^2]\, E[(U_s - 1/2)^2]}} = 12E[U_r U_s] - 3 \tag{6}$$

for each $1 \leq r < s \leq m$. In order to model the joint probability distribution for a vector-valued random variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_m) \in (0, \infty)^m$ with known marginals $u_i = F_i(x_i)$ we simply construct uniformly distributed random variables $U_i = F_i(X_i) \in (0,1)$ for each $i = 1, 2, \ldots, m$ and use the m-dimensional copula $C(\boldsymbol{u}) = C(\boldsymbol{F}(\boldsymbol{x})) = C(F_1(x_1), F_2(x_2), \ldots, F_m(x_m))$. We say that the *grade correlation coefficients* for $\boldsymbol{X}$ are simply the correlation coefficients for $\boldsymbol{U}$ defined above. That is

$$\rho_{r,s} = \frac{E[(F_r(X_r) - 1/2)(F_s(X_s) - 1/2)]}{\sqrt{E[(F_r(X_r) - 1/2)^2]\, E[(F_s(X_s) - 1/2)^2]}} = 12E[F_r(X_r)F_s(X_s)] - 3 \tag{7}$$

for each $1 \leq r < s \leq m$. In this paper we distinguish between the Spearman rank correlation coefficients $\widehat{\rho}_{r,s}$ obtained from the observed data $\{x_{i,j}\}_{j=1}^N$, or equivalently from the transformed observed data $\{u_{i,j}\}_{j=1}^N$, and the grade correlation coefficients $\rho_{r,s}$ defined by (7).

For our proposed application we make the following observation. Once it has been decided that the monthly rainfall X_i can be modelled by a gamma distribution $X_i \sim \Gamma(\alpha_i, \beta_i)$ with $F_i(x) = F_{\alpha_i,\beta_i}(x)$ then we can define transformed random variables $U_i = F_i(X_i)$ for each $i = 1, 2, 3$ which will be uniformly distributed on $(0, 1)$. The original observed data set $\{x_{i,j}\}_{j=1,2,\ldots,N}$ for the random variable $X_i \sim \Gamma(\alpha_i, \beta_i)$ can be transformed into a corresponding data set $\{u_{i,j} = F_i(x_{i,j})\}_{j=1,2,\ldots,N}$ for the uniformly distributed random variable $U_i \sim U([0, 1])$ for each $i = 1, 2, 3$. This transformation is also important insofar as it removes seasonal factors from the observed data.

7.2. Problem formulation and solution for the checkerboard copula of maximum entropy with prescribed grade correlation coefficients

We outline the basic ideas. An m-dimensional multivariate *checkerboard* copula is a probability distribution on the unit hypercube $[0, 1]^m$ defined by subdividing the hypercube into n^m congruent small hypercubes with constant density on each one. If the density on $I_{\boldsymbol{i}}$ where $\boldsymbol{i} = (i_1, i_2, \ldots, i_m)$ is defined by $n^{m-1}h_{\boldsymbol{i}} \geq 0$ then the marginal distributions will be uniform if

$$\sum_{\boldsymbol{i} \in S(r,i)} h_{\boldsymbol{i}} = 1 \quad \text{for each } r \text{ and each } i,$$

where $S(r, i) = \{\boldsymbol{i} \mid i_r = i\}$ for $r = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$. In such cases we say that $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ where $\ell = n^m$ is multiply-stochastic. We wish to construct a joint density in this form with the desired grade correlation coefficients. For sufficiently large n there are many ways that this can be done.

> The *principle of maximum entropy* means that the best such distribution is the most disordered or least prescriptive solution—the multiply-stochastic hypermatrix $\boldsymbol{h} \in \mathbb{R}^\ell$ which has the most equal subdivision of probabilities but still allows the required correlations. In mathematical terminology this is the hypermatrix that satisfies the grade correlation constraints and has the highest possible entropy.

We have the following formal statement of the problem.

Problem 7.1 (The primal problem). *Find the hypermatrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ where $\boldsymbol{i} = (i_1, \ldots, i_m)$ and $\ell = n^m$ to maximize the entropy*

$$J(\boldsymbol{h}) = (-1)\left[\frac{1}{n} \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1)\log_e n\right] \tag{8}$$

subject to the multi-stochastic constraints

$$\sum_{\boldsymbol{i} \in S(r,i)} h_{\boldsymbol{i}} = 1 \tag{9}$$

for all $r = 1, 2, ..., m$ and $i = 1, 2, ..., n$ and

$$h_{\boldsymbol{i}} \geq 0 \tag{10}$$

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

$$12 \left[\frac{1}{n^3} \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} (i_r - 1/2)(i_s - 1/2) \right] - 3 = \widehat{\rho}_{r,s} \tag{11}$$

for $1 \leq r < s \leq m$ *where the observed rank correlation coefficient* $\widehat{\rho}_{r,s}$ *is known for all* $1 \leq r < s \leq m$.

The problem can be neatly and rigorously solved using the theory of *Fenchel duality*. The solution and a full description of the construction process for the trivariate checkerboard copula of maximum entropy[1] can be found elsewhere [13,14]. The m-dimensional copula of maximum entropy is determined by $m(m-1)/2$ real parameters—the grade correlation coefficients—defined in equation (7).

7.3. The checkerboard copula of maximum entropy for seasonal rainfall in February-March-April at Kempsey

We set $m = 3$ and $n = 4$. The triply-stochastic hypermatrix $\boldsymbol{h} \in \mathbb{R}^{4\times4\times4}$ describing the trivariate checkerboard copula of maximum entropy for February–March–April rainfall at Kempsey is shown below to four decimal place accuracy. We set $\rho_{12} = \widehat{\rho}_{12} = 0.202$, $\rho_{13} = \widehat{\rho}_{13} = 0.112$ and $\rho_{23} = \widehat{\rho}_{23} = 0.152$ and calculate

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1262 & 0.0975 & 0.0733 & 0.0536 \\ 0.0870 & 0.0756 & 0.0639 & 0.0525 \\ 0.0567 & 0.0554 & 0.0527 & 0.0487 \\ 0.0350 & 0.0384 & 0.0411 & 0.0427 \end{bmatrix}, \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0920 & 0.0765 & 0.0618 & 0.0486 \\ 0.0750 & 0.0701 & 0.0637 & 0.0563 \\ 0.0578 & 0.0608 & 0.0621 & 0.0618 \\ 0.0422 & 0.0499 & 0.0573 & 0.0641 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0641 & 0.0573 & 0.0499 & 0.0422 \\ 0.0618 & 0.0621 & 0.0608 & 0.0578 \\ 0.0563 & 0.0637 & 0.0701 & 0.0750 \\ 0.0486 & 0.0618 & 0.0765 & 0.0920 \end{bmatrix}, \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0427 & 0.0411 & 0.0384 & 0.0350 \\ 0.0487 & 0.0527 & 0.0554 & 0.0567 \\ 0.0525 & 0.0639 & 0.0756 & 0.0870 \\ 0.0536 & 0.0733 & 0.0975 & 0.1262 \end{bmatrix},$$

where $\boldsymbol{h}_i = [h_{ijk}]$. The entropy is given by $J(\boldsymbol{h}) \approx -0.040714$. For the given rank correlation coefficients a simple MATLAB program computed the checkerboard copula in 0.78 s on a Macintosh OS laptop computer.

8. Fifth experiment: simulations for seasonal rainfall at Kempsey

We used the copula of maximum entropy to generate numerous simulated data sets each spanning a period of $N = 123$ years. The simulation finds the monthly rainfalls in each year, draws a histogram of the total seasonal rainfall and plots the corresponding time series. The simulations show that sample statistics are quite variable. Selected results in Figures 5 and 6 from 16 successively generated

[1] MATLAB algorithms to construct the copula in 3-dimensions are lodged at the *CARMA* website, `www.carma.newcastle.edu.au/hydro`, and are also available from the corresponding author Emeritus Professor Phil Howlett.

simulated data sets illustrate the typical range of variation in samples of this size. Consequently it is not unreasonable to believe that the statistics for an observed sample taken over a period of 123 years may exhibit similar variation. In view of these observations we suggest that it is prudent to regard minor changes in statistical parameters from the observed values during the 30-year standard period 1961 to 1990 as random fluctuations. To obtain more reliable observed samples it would be necessary to take observations over a much larger time span than 123 years.

Figure 5. Selected histograms for total rainfall from 15 successive random simulations of seasonal rainfall at Kempsey for February–March–April with each simulation spanning $N = 123$ years using the copula of maximum entropy. The histograms—for Sim #4 (left), Sim #6 (centre) and Sim #11 (right)—show typical sample variation for $N = 123$ years. The sample mean $\overline{x}$ and variance s^2 are shown on the plots.

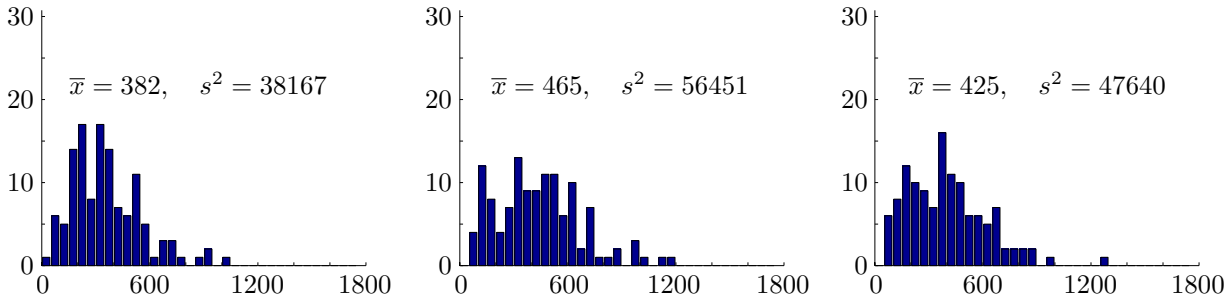
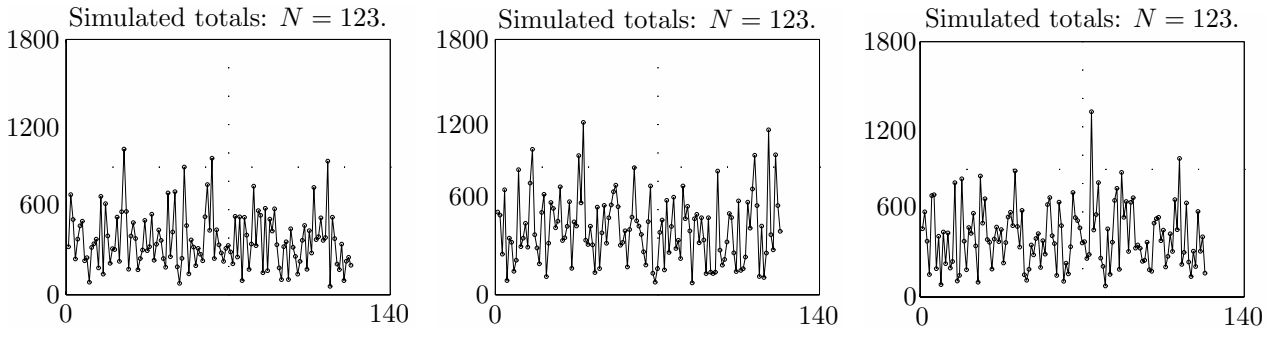


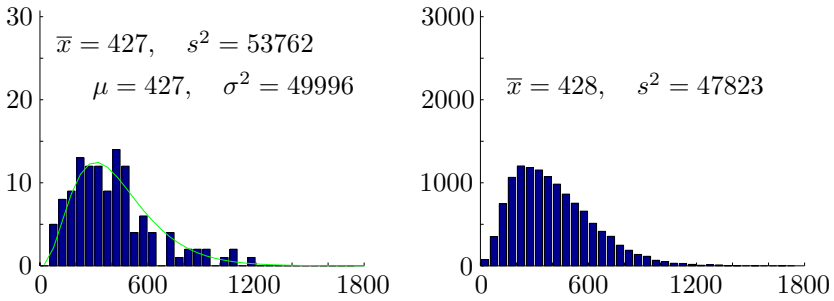
Figure 6. Selected time series for total rainfall from 15 successive random simulations of seasonal rainfall at Kempsey for February–March–April with each simulation spanning $N = 123$ years using the copula of maximum entropy. The time series—Sim #4 (left), Sim #6 (centre) and Sim #11 (right)—show typical sample variation for $N = 123$ years.



In Figure 7 we compare the observed frequencies for total rainfall with the probability density of the maximum likelihood gamma distribution $X \sim \Gamma(3.6524, 116.9983)$ for total rainfall[2] and also show generated frequencies from a typical simulation using the copula of maximum entropy spanning a period of 12300 years. Summary statistics for both models are shown in Figure 7.

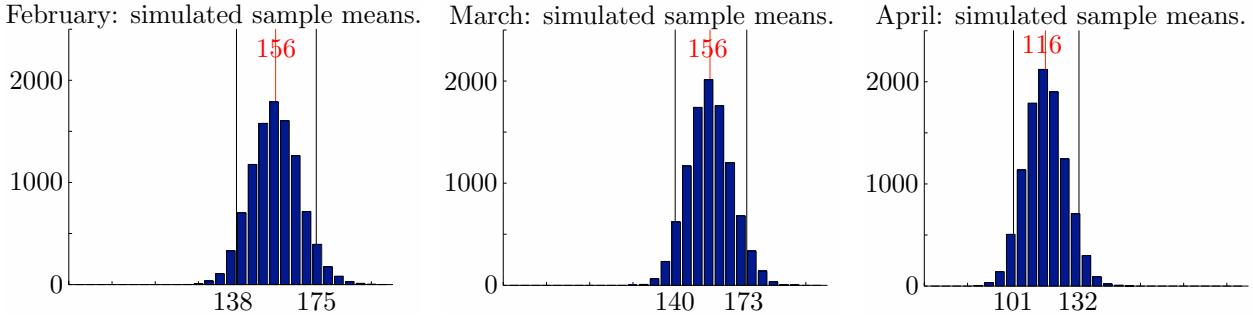
[2] This model generates simulated seasonal rainfall totals directly and does not generate individual monthly rainfall totals.

Figure 7. Histogram for observed total rainfall for February–March–April during the period 1889 to 2011 with mean $\overline{x}$ and variance s^2 and designated gamma distribution $X \sim \Gamma(3.6524, 116.9983)$ with mean μ and variance σ^2 (left) and typical histogram from 10 successive simulations for total rainfall at Kempsey for February–March–April with each simulation spanning a period of $N = 12300$ years using the copula of maximum entropy. The histogram—Sim # 9 of 10 successive simulations—is a true representation of the entire model population. The sample mean $\overline{x}$ and variance s^2 are shown on the graph.



In Figures 8, 9 and 10 we show histograms of sample means for monthly rainfall, sample means and variances for seasonal rainfall and corresponding values for the rank correlation coefficients from 10000 independently generated simulated data sets. Each data set covers a period of $N = 123$ years. In each case the empirical 95% confidence intervals established by the simulations are also shown. The mean values over all samples and the corresponding empirical 95% confidence intervals were $\overline{x_1} = 156 \in [138, 175]$, $\overline{x_2} = 156 \in [140, 173]$ and $\overline{x_3} = 116 \in [101, 132]$ for the monthly rainfalls, $\overline{x} = 427 \in [396, 459]$ and $s^2 = 47467 \in [36403, 60266]$ for the seasonal rainfall and seasonal variance, and $\widehat{\rho}_{12} = 0.201 \in [0.055, 0.340]$, $\widehat{\rho}_{13} = 0.111 \in [-0.037, 0.258]$ and $\widehat{\rho}_{23} = 0.152 \in [0.004, 0.297]$ for the rank correlation coefficients. Importantly we see that the observed overall variance $53762 \in [36403, 60266]$ lies well within the empirical 95% confidence intervals for the proposed model. We conclude that there are no reasonable statistical grounds to reject the proposed model.

Figure 8. Histograms for simulated monthly sample means at Kempsey from 10000 samples with each sample spanning a period of $N = 123$ years showing empirical 95% confidence intervals and overall means for February (left), March (centre) and April (right).



It is one thing to argue that the observed statistics lie within the empirical 95% confidence intervals generated by the proposed model. It is another thing altogether to turn this statement around and imagine

Figure 9. Histograms for simulated seasonal sample means (left) and sample variances (right) at Kempsey from 10000 samples with each sample spanning 123 years showing empirical 95% confidence intervals and overall mean and variance.

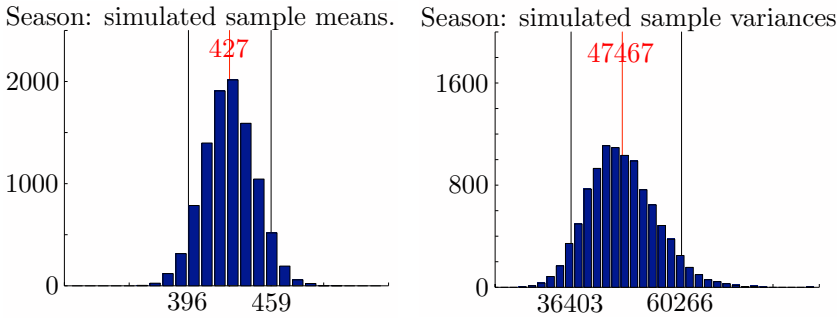
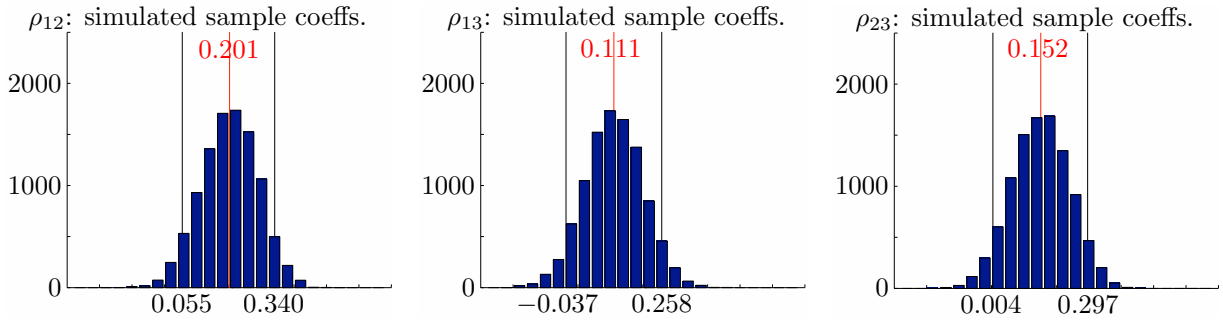


Figure 10. Histograms for simulated sample rank correlation coefficients at Kempsey from 10000 samples with each sample spanning 123 years showing empirical 95% confidence intervals and overall mean values for ρ_{12} (left), ρ_{13} (centre) and ρ_{23} (right).



that the observed statistics were actually generated by the model. If so one might ponder on what *should* be expected from the next set of generated statistics. The answer is that one should not be unduly surprised by a change of $\pm 7\%$ in the average seasonal rainfall. In purely statistical terms such changes lie within the empirical 95% confidence intervals and as such could be regarded as fluctuations due to chance alone.

9. Sixth experiment: an alternative joint distribution—the checkerboard normal copula

The normal distribution is popular and is generally regarded as easy to apply. Hence we decided to compare the copula of maximum entropy to a copula defined by a multivariate normal distribution.

9.1. Constructing a normal checkerboard copula with prescribed grade correlation coefficients

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 $\Sigma = E[\boldsymbol{Z}\boldsymbol{Z}^T] = [\cos\theta_{r,s}] \in [-1,1]^{m\times m}$ is the correlation matrix. The marginal distributions for Z_r are standard unit normal distributions 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 function $c : [0,1]^m \to [0,\infty)$ 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 an m-dimensional *normal copula* $C : [0,1]^m \to [0,1]$ given by

$$C(\boldsymbol{u}) = \int_{\times_{i=1}^{n}[0,u_i]} c(\boldsymbol{v}) d\boldsymbol{v}.$$

In practice the normal copula is approximated by a *checkerboard normal copula*. The idea is simple. We assume that the positive definite symmetric matrix Σ is known. The unit hypercube $I = [0,1]^m$ is divided into $\ell = n^m$ congruent hypercubes $I_{\boldsymbol{i}}$ where $\boldsymbol{i} = (i_1, \ldots, i_n)$ and we construct the corresponding multiply-stochastic hypermatrix $\boldsymbol{k} = [k_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ where $\ell = n^m$ by defining

$$k_{\boldsymbol{i}} = n \int_{I_{\boldsymbol{i}}} c(\boldsymbol{u}) d\boldsymbol{u}. \tag{12}$$

Write $\Sigma = P\Lambda P^T \Leftrightarrow \Lambda = P^T \Sigma P$ where P is orthogonal and Λ is diagonal. Now consider the successive transformations $\boldsymbol{w} = \Phi^{-1}(\boldsymbol{v}) \Leftrightarrow w_r = \Phi^{-1}(v_r)$ for each $r = 1, \ldots, m$, followed by the length distortion $\boldsymbol{y} = \Lambda^{1/2}\boldsymbol{w}$, the orthogonal transformation $\boldsymbol{z} = P\boldsymbol{y}$ and finally the transformation $\boldsymbol{u} = \Phi(\boldsymbol{z}) \Leftrightarrow u_r = \Phi(z_r)$ for each $r = 1, \ldots, m$. It has been shown [14] that the collective transformation $\boldsymbol{u} = \Phi[\Lambda^{-1/2} P^T \Phi^{-1}(\boldsymbol{v})]$ maps $\boldsymbol{v} \in [0,1]^m$ into $\boldsymbol{u} \in [0,1]^m$. If we define the region $J_{\boldsymbol{i}} = \Phi[\Lambda^{-1/2} P^T \Phi^{-1}(I_{\boldsymbol{i}})] \subset [0,1]^m$ then

$$\int_{I_{\boldsymbol{i}}} c(\boldsymbol{u}) d\boldsymbol{u} = \int_{J_{\boldsymbol{i}}} d\boldsymbol{v} = V(J_{\boldsymbol{i}})$$

where $V(J_{\boldsymbol{i}})$ denotes the volume of $J_{\boldsymbol{i}}$. Thus, it is necessary to calculate the volume of $J_{\boldsymbol{i}}$ for each index $\boldsymbol{i}$ in order to find the hypermatrix $\boldsymbol{k}$. Unfortunately calculation of these volumes is not straightforward. See [14] for more details. In a further complication we also note that the grade correlation coefficients for the checkerboard normal copula must be calculated directly using the formula

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

This means that the positive definite symmetric matrix Σ must be iteratively adjusted in order to satisfy the equations $\rho_{r,s} = \widehat{\rho}_{r,s}$.

9.2. The normal checkerboard copula for Kempsey

We set $m = 3$ and $n = 4$. The triply-stochastic hypermatrix $\boldsymbol{k} \in \mathbb{R}^{4\times4\times4}$ describing the trivariate checkerboard copula of maximum entropy for February–March–April rainfall at Kempsey is shown below to four decimal place accuracy. We define Σ by setting $\theta_{12} = 1.3179$, $\theta_{13} = 1.4309$, $\theta_{23} = 1.3813$ to give $\rho_{1,2} = \widehat{\rho}_{1,2} = 0.202$, $\rho_{1,3} = \widehat{\rho}_{1,3} = 0.112$ and $\rho_{2,3} = \widehat{\rho}_{2,3} = 0.152$ and calculate

$$\boldsymbol{k}_1 \approx \begin{bmatrix} 0.1324 & 0.0961 & 0.0759 & 0.0529 \\ 0.0829 & 0.0719 & 0.0637 & 0.0517 \\ 0.0585 & 0.0566 & 0.0540 & 0.0487 \\ 0.0341 & 0.0381 & 0.0402 & 0.0420 \end{bmatrix}, \boldsymbol{k}_2 \approx \begin{bmatrix} 0.0891 & 0.0724 & 0.0616 & 0.0475 \\ 0.0720 & 0.0687 & 0.0648 & 0.0575 \\ 0.0593 & 0.0629 & 0.0638 & 0.0627 \\ 0.0424 & 0.0516 & 0.0579 & 0.0662 \end{bmatrix},$$

$$\boldsymbol{k}_3 \approx \begin{bmatrix} 0.0662 & 0.0579 & 0.0516 & 0.0424 \\ 0.0627 & 0.0638 & 0.0629 & 0.0593 \\ 0.0575 & 0.0648 & 0.0687 & 0.0720 \\ 0.0475 & 0.0616 & 0.0724 & 0.0891 \end{bmatrix}, \boldsymbol{k}_4 \approx \begin{bmatrix} 0.0420 & 0.0402 & 0.0381 & 0.0341 \\ 0.0487 & 0.0540 & 0.0566 & 0.0585 \\ 0.0517 & 0.0637 & 0.0719 & 0.0829 \\ 0.0529 & 0.0759 & 0.0961 & 0.1324 \end{bmatrix},$$

where $\boldsymbol{k}_i = [k_{ijk}]$. The entropy $J(\boldsymbol{k}) = -0.041158$ of the checkerboard normal copula is slightly less than the maximum entropy but it is nevertheless true that $\boldsymbol{k} \approx \boldsymbol{h}$.

Simulation results are very similar to those obtained using the checkerboard copula of maximum entropy. It is necessary to search for the values of $\theta_{r,s}$. For given values $\theta_{r,s}$ a MATLAB program to find the necessary volumes by counting the transformed distribution of 64^3 equally spaced points in $(0, 1)^3$ took 31 s to run and calculated the corresponding $\boldsymbol{k}$ to only 2 decimal place accuracy. This program was used to iteratively adjust $\theta_{r,s}$. For final adjustment the program used 256^3 points and took 1970 s to run. This shows that calculation of the checkerboard normal copula is considerably more difficult than calculation of the checkerboard copula of maximum entropy.

10. Commentary: the modelling process and model evaluation

In this section we wish to comment on the modelling process. Thus, we compare our proposed model with two elementary models in order to highlight the key points. We will not compare our model directly with any of the more sophisticated models—such as the model proposed by Srikanthan—because it is not our intention to reach a conclusion about which model is best. Our intention is to examine the difficulties that arise when modelling rainfall accumulations and to examine the physical and theoretical basis for various assumptions. Our objective is to build a model based on the measurement of certain key statistics and to decide—on the basis of standard statistical performance criteria—whether the model should be rejected or accepted.

The question of which population model is best depends on the measurements that have been made and on the selection criteria. In our case we argue that if we have a single sample of independently generated monthly rainfall totals and if only the sample mean of the observed values and the sample mean of the logarithm of the observed values are known, then the principle of maximum entropy tells us that the maximum likelihood gamma distribution is the best model for random generation of monthly totals. For the joint distribution of seasonal rainfall we show that if a piecewise constant joint probability

density is used on a subdivision of the unit cube such that the marginal distributions are uniform and the grade correlation coefficients are equal to the observed rank correlation coefficients then the principle of maximum entropy shows that the checkerboard copula of maximum entropy is the best solution. Thus in a purely theoretical sense, on that axiomatic basis, there is no point in trying to compare our model to another model based on different measurements and different selection criteria. A much more sensible debate is to argue about the appropriate measurements and the best selection criteria to be used.

We understand why hydrologists have focussed on model outputs. The problems of catchment hydrology are real and they require solution. Simulation of realistic rainfall and run-off regimes over an entire river basin is vital for planning of sustainable agricultural practice and for implementation and operation of effective flood mitigation infrastructure. There is ample recent evidence in Australia of failures to understand and appreciate both sustainability of agriculture and management of flood mitigation infrastructure. The failure of water supplies for irrigation in the Murray-Darling basin during an extended period of below average rainfall in eastern Australia from 2003–2008 and the flooding of Brisbane in January 2011 are prime examples.

The main modelling motivations in catchment management are to understand long-term behaviour and to cope with the management of extreme events. Thus, the model construction should reflect the most appropriate data collection processes and the best statistical design in relation to the desired objectives. These issues are addressed in the study by Srikanthan [18] but the approach is somewhat informal and is general rather than particular. Despite the acknowledged practical imperatives of catchment management—which seem to have captured the attention of engineers in recent times—we argue that it is both necessary and beneficial to continue investigating more basic modelling questions for which the answers are less tangible and the benefits less clear.

In the first place there is the apparent contradiction in the process of modelling rainfall accumulations at a single site over different timescales. For each fixed timescale there is general agreement that a mixed gamma distribution with cumulative distribution $F(x) = P[0 \leq X < x]$ for $x \geq 0$ given by

$$F(x) = p_0 + (1 - p_0) \int_0^x f_{\alpha,\beta}(x)dx$$

provides a satisfactory description. However, there is no simple model for a joint distribution of daily rainfall in a particular month which incorporates an appropriate marginal gamma distribution for rainfall on each separate day and an appropriate gamma distribution for the total monthly rainfall.

The same dilemma arises in the relationship between monthly and seasonal rainfall. This is the problem that we have tried to address. To test the utilitarian value of our model we must test the properties of the seasonal rainfall totals. By contrast, the Srikanthan model [18] at an individual site is essentially a *disaggregation* model—since the model is ultimately adjusted at monthly and daily level to match the annual characteristics. Hence it was tested, with generally positive results, at a monthly and daily level. The model will be acceptable—in our view—if it can be shown that the bias introduced by modifying the monthly and daily rainfall totals is not significant. Incidentally it can be seen that if the generated rainfall for a particular day at some given site is $\widetilde{X} \sim \Gamma(\alpha, \beta)$ then the modified rainfall in the Srikanthan model will take the form $X = k\widetilde{X} \sim \Gamma(\alpha, k\beta)$. Thus both the daily mean and daily standard deviation will be multiplied by the correction factor k.

There are other important modelling issues. We use a model in which monthly rainfall distribution does not depend on the year. We have tested samples generated by our model and have concluded using stablished statistical procedures that the apparent trend-slopes in the observed data are well within the 95% confidence intervals for the randomly generated trend-slopes in the simulated data from our time independent model. This means that the trends in the observed data could have been generated by chance alone. This does not mean that the observed data is time independent and it does not mean that a time dependent model would be unsuitable. Indeed we noted earlier in the paper that a linear regression against time showed a positive trend-slope for the observations in each month. Thus one could argue that an unbiassed model should incorporate those observed trends and that one should expect random fluctuations in these base trend-slopes during simulation. The only clear reason to prefer a time independent model is that it will be simpler.

The final point we wish to address is the issue of parameter estimation. Perhaps the most contentious estimation is that of the rank correlation coefficients. In each case subsequent simulations with the proposed model showed the observed values and the empirical 95% confidence intervals as $\widehat{\rho}_{12} = 0.201 \in [0.055, 0.340]$, $\widehat{\rho}_{13} = 0.111 \in [-0.037, 0.258]$ and $\widehat{\rho}_{23} = 0.152 \in [0.004, 0.297]$. This suggests that one could legitimately argue that a model with $\rho_{13} = 0$ would be simpler and would possibly also fit the observed data. Nor is it entirely clear, on the basis of one sample, that all correlations are truly positive. Indeed the strongest argument for assuming that there is positive monthly correlation is probably that a model with independent marginals will seriously underestimate the seasonal variance. Even this argument is not entirely certain since the calculated variance of 38236 for the model with independent marginals lies within the empirical 95% confidence interval $[36403, 60266]$ for the simulated variance obtained from our model.

In Table 1 summary statistics for (a) observed total rainfall in February–March–April at Kempsey are compared to summary population statistics for models using (b) the maximum likelihood gamma distribution (c) the checkerboard copula of maximum entropy with marginal gamma distributions[3] and (d) the joint distribution with independent marginal gamma distributions.

Table 1. Model comparison for total rainfall in February–March–April at Kempsey.

Distribution	mean	variance
(a) observed	427	53762
(b) maximum likelihood gamma	427	49996
(c) checkerboard copula of maximum entropy	427	47448
(d) independent	427	38236

11. Seventh experiment: simulations for seasonal rainfall at Sydney

We also successfully simulated seasonal rainfall in March-April-May at Sydney. We used official records from the Australian Bureau of Meteorology for Station 066062 (Observatory Hill) at Sydney

[3] Details of the theoretical calculation procedures for (c) can be found in [13,14].

in NSW during the period 1859 to 2008. The monthly rainfall totals were again modelled by gamma distributions $X_i \sim \Gamma(\alpha_i, \beta_i)$ with the respective parameters defined by

$$\boldsymbol{\alpha} = (1.7413, 1.3329, 1.2579), \quad \boldsymbol{\beta} = (74.5972, 94.6996, 95.9645).$$

The copula of maximum entropy is shown below to four decimal place accuracy. We set $\rho_{12} = \widehat{\rho}_{12} = 0.112$, $\rho_{13} = \widehat{\rho}_{13} = 0.043$ and $\rho_{23} = \widehat{\rho}_{23} = 0.183$ and calculate

$$\boldsymbol{h}_1 \approx \begin{bmatrix} 0.1070 & 0.0847 & 0.0649 & 0.0482 \\ 0.0766 & 0.0710 & 0.0638 & 0.0555 \\ 0.0526 & 0.0571 & 0.0600 & 0.0612 \\ 0.0346 & 0.0440 & 0.0542 & 0.0647 \end{bmatrix}, \boldsymbol{h}_2 \approx \begin{bmatrix} 0.0916 & 0.0739 & 0.0577 & 0.0437 \\ 0.0719 & 0.0680 & 0.0622 & 0.0551 \\ 0.0542 & 0.0599 & 0.0643 & 0.0667 \\ 0.0391 & 0.0507 & 0.0637 & 0.0774 \end{bmatrix},$$

$$\boldsymbol{h}_3 \approx \begin{bmatrix} 0.0774 & 0.0637 & 0.0507 & 0.0391 \\ 0.0668 & 0.0643 & 0.0599 & 0.0542 \\ 0.0551 & 0.0622 & 0.0680 & 0.0719 \\ 0.0437 & 0.0577 & 0.0739 & 0.0916 \end{bmatrix}, \boldsymbol{h}_4 \approx \begin{bmatrix} 0.0647 & 0.0542 & 0.0440 & 0.0346 \\ 0.0612 & 0.0600 & 0.0571 & 0.0526 \\ 0.0555 & 0.0638 & 0.0710 & 0.0766 \\ 0.0482 & 0.0649 & 0.0847 & 0.1070 \end{bmatrix},$$

where $\boldsymbol{h}_i = [h_{ijk}]$. The entropy is $J(\boldsymbol{h}) \approx -0.026749$. For purposes of comparison we also modelled the total seasonal rainfall by a gamma distribution $X \sim \Gamma(\alpha, \beta)$ where we used maximum likelihood to set $\alpha = 3.5157$ and $\beta = 107.1866$. Summary statistics for seasonal rainfall at Sydney are shown in Table 2. Simulations for Sydney showed similar behaviour to the corresponding simulations for Kempsey.

Table 2. Model comparison for total rainfall in March-April-May at Sydney.

Distribution	mean	variance
(a) observed	377	37363
(b) maximum likelihood gamma	377	40392
(c) copula of maximum entropy	377	39009
(d) independent	377	33228

12. Conclusions

The problem of seasonal rainfall modelling has no obvious solution. Our model is derived on a solid theoretical basis and the standard tests of simulated data generated by the model against observed data showed there was insufficient evidence to reject the model on statistical grounds. There are many other models that have been proposed recently for the purpose of modelling catchment hydrology. In most cases researchers report on the successful use of these models in simulation of catchment rainfall. Although such models are tested extensively to ensure that the simulated data is a good approximation to the observed data it seems there is often no clear axiomatic model structure, no explicit *a priori* measurement strategy for parameter estimation and no coherent testing procedure. It is more like *measure everything you can, use as many parameters as you like and test every statistical parameter that comes to mind*. See [1] for further discussion of the backtest overfitting problem.

References

1. D.H. Bailey, J.M. Borwein, M. Lopez de Prado, and Qiji Zhu, Pseudo mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. Submitted *Notices of the AMS*, September 2013. See http://www.carma.newcastle.edu.au/jon/backtest.pdf.
2. Borwein, J.M., Lewis, A.S., Convex Analysis and Nonlinear Optimization, Theory and Examples, Second Edition. CMS Books in Mathematics, **3**, Springer-Verlag New York, Inc. (2006).
3. Borwein, J .M., Vanderwerff, J .D.: Convex Functions: Constructions, Characterizations and Counterexamples, **109**, Encyclopedia of Mathematics and its Applications. Cambridge University Press (2010).
4. W.M. Getz (2003), Correlative coherence analysis: Variation from intrinsic and extrinsic sources in competing populations. *Theoretical Population Biology*, **64**(1), 89-99.
5. Md Masud Hasan and Peter K. Dunn (2011), Two Tweedie distributions that are near optimal for modelling monthly rainfall in Australia, *International J Climatology*, **31**(9), 1389-1397, DOI: 10.1002/joc.2162.
6. Hurst, H.E., (1951). Long Term Storage Capacities of Reservoirs. *Transactions of the American Society of Civil Engineering*, **116**, 776-808.
7. Jaynes, E. T. (1957a), Information theory and statistical mechanics. *Physical Review*, **106**(4), 620-630.
8. Jaynes, E. T. (1957b), Information theory and statistical mechanics ll. *Physical Review*, **108**(2), 171-190.
9. R. W. Katz and M. B. Parlange (1998), Overdispersion phenomenon in stochastic modelling of precipitation, *J. Climate*, **11**, 591-601.
10. Demetris Koutsoyiannis (2011), Hurst-Kolmogorov dynamics and uncertainty, *J. American Water Resources Association*, **47** (3), 481-495.
11. Roger B. Nelsen, *An Introduction to Copulas*, Lecture Notes in Statistics, **139**, Springer-Verlag, New York, (1999).
12. John von Neumann, *Mathematical Foundations of Quantum Mechanics*, Princeton University Press (1996).
13. J. Piantadosi, P.G. Howlett and J.M. Borwein (2012), Copulas of maximum entropy, *Optimization Letters*, **6**(1), 99-125, DOI: 10.1007/s11590-010-0254-2.
14. J. Piantadosi, P.G. Howlett, J.M. Borwein, J. Henstridge (2012), Maximum entropy methods for generating simulated rainfall, Invited submission to *Numerical Algebra, Control and Optimization*, Special Issue for Charles Pearce's 70th birthday, **2**(2), 233-256. doi: 10.3934/naco.2012.2.233.
15. K. Rosenberg, J. Boland and P.G. Howlett (2004), Simulation of monthly rainfall totals, *ANZIAM J.*, **46**, (E), E85-E104.
16. C. E. Shannon (1948), A Mathematical Theory of Communication, *The Bell System Technical Journal*, **27**, 379-423 and 623-656, July and October, 1948.
17. R. Srikanthan and T. A. McMahon (2001), Stochastic generation of annual, monthly and daily climate data: A review, *Hydr. and Earth Sys. Sci.*, **5**, 633-670.

18. Ratnasingham Srikanthan (2005), *Stochastic generation of daily rainfall at a number of sites*, Cooperative Research Centre for Catchment Hydrology, Technical Report 05/7.
19. R. D. Stern and R. Coe (1984), A model fitting analysis of daily rainfall, *J. Roy. Statist. Soc. A*, **147**, 1-34.
20. D. S. Wilks and R. L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Prog. Phys. Geog.*, **23**, 329-357.
21. Eberhard Ziedler, (1997), *Applied Functional Analysis, Applications to Mathematical Physics*, Applied Mathematical Sciences **108**, Springer.