Hydrol. Earth Syst. Sci., 13, 2299–2314, 2009
www.hydrol-earth-syst-sci.net/13/2299/2009/

Hydrology and Earth System Sciences

Copula based multisite model for daily precipitation simulation

A. Bárdossy[1] **and G. G. S. Pegram**[2]

[1]Institute of Hydraulic Engineering, University of Stuttgart, Stuttgart, 70569, Germany
[2]Civil Engineering Program, University of KwaZulu-Natal, Durban, South Africa

Received: 25 May 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 19 June 2009
Revised: 29 October 2009 – Accepted: 7 November 2009 – Published: 3 December 2009

Abstract. From the point of view of multisite stochastic daily rainfall modelling, there are two new ideas introduced in this paper. The first is the use of asymmetrical copulas to model the spatial interdependence structure of the rainfall amounts together with the rainfall occurrences in one relationship. The second is in the evaluation of the (necessary but often ignored) congregating behaviour of the higher values of simulated rainfall; this evaluation is performed by calculating the entropy of the observations at all the near equilateral triangles that can be formed from the sequences at the gauge sites, as a function of their mutual separation distance. It turns out that the model captures the qualities desired and offers a fresh approach to a relatively mature problem in hydrometeorology.

1 Introduction

In 21st century hydrology, there is a growing need to match precipitation simulation to the fine spatial scale of distributed hydrological models. Random fields mimicking radar rainfields offer an attractive possibility, but the lowly rain-gauge is still the instrument of choice for conditioning such fields. Without a good model for multi-site daily rainfall at gauge sites, a random field model would likely be missing some connection to the reality of the detail, particularly those aspects dependent on location. It is the purpose of this paper to delve into the underlying interdependence between daily rain-gauge readings in a variable terrain and model it faithfully.

Careful study of the dependence structure of many hydrometeorological data sets, both spatial and temporal, reveals that the dependence is more complex than that modelled by conventional correlation of the multivariate normal. The tool used in this paper for modelling this complex dependence is the multivariate copula, relating observations at many sites in space and time. The advantages that can be ascribed to using copula densities to represent interdependence between variables include the following properties:

- The empirical copulas (probability density scatterplots in many dimensions) are independent of their corresponding marginal distributions, so that copulas display interdependence between variables in its purest or essential form.
- Empirical copulas are easily computed from data.
- Differences in types of association between variables are readily identified by copula shape.
- A suite of theoretical copula density functions has been developed to model these attributes.

We proceed by describing how copulas can be used for modelling spatial and temporal relationships between variables sampled at fixed locations. As a point of departure, the normal score (or quantile) transform (NQT) has been used in multi-site stochastic simulation in hydrology for the last 40 or more years, starting with the work of Thomas and Fiering (1962) and Matalas (1967). In their approach, to achieve the simulation, field variables are analysed and their marginal distributions fitted. These are then transformed to normal variates and these in turn are associated by their pairwise correlations through the multivariate normal distribution function. The unique advantage of the multivariate normal is that all higher order correlations are implicitly defined by the second order moments. Thus, in simulation, generation of the multi-site replicates of the field variables is readily achieved by generating properly associated multivariate normal variates which are back-transformed to synthetic field variables. These ideas are now common practice in hydrological applications.

Correspondence to: A. Bárdossy
(bardossy@iws.uni-stuttgart.de)

Published by Copernicus Publications on behalf of the European Geosciences Union.

A careful examination of sample copulas of hydrometeorological processes reveals, however, that there is frequently a significant departure of the dependence structure from that of the bivariate Gaussian which is simply defined by its correlation coefficient. The question posed and answered in this paper is: How does one exploit this knowledge in simulation of multi-site daily rainfall?

1.1 Simulation of multi-site Time Series of daily precipitation

Srikanthan and McMahon (2001) gave an extensive review of rainfall models, including multisite network modelling. There has since been continued interest in the subject and we name some of the activity (this is not an exhaustive list as we use a different modelling procedure in this paper). Srikanthan (2005) generated the daily rainfall values via a season-dependent multisite AR(1) model, based on the work of Wilks (1998), which was then post-conditioned to give the observed monthly and annual variations. Mehrotra et al. (2006) compared three multi-site stochastic weather generators, including (i) a parametric hidden Markov model (HMM), (ii) the Srikanthan (2005) multi-site stochastic precipitation generation model proposed by Wilks (1998) and (iii) a non-parametric K-nearest neighbour (KNN) model, concluding that Wilks-based model had the edge. Apipattanavis et al. (2007) used a semiparametric blend of Markov chains and Bootstrap resampling. Mehrotra and Sharma (2007) proposed a modification of the traditional Markov chain approach for modelling the occurrences, using an analytically derived factor that represents the influence of rainfall aggregated over long time periods in an attempt to incorporate low-frequency variability in simulation. Srikanthan and Pegram (2009) extended the post-conditioning in the earlier work of Srikanthan (2005) to include spatial as well as temporal dependence to the same end.

One of the notable features of rainfall estimated by radar is that the low frequency fluctuations of the random field are evident in zones of higher rainfall distinct from zones of lower rainfall. Data sampled from raingauge networks also exhibits this behaviour (the grouping of high separate from low values) but is not so obvious. This strong spatial dependence has its origins in the physics of rainfall accumulations and has traditionally been modelled by two-dimensional tools in the past (covariance and variogram functions, for example) and we suggest that the spatial interdependence is stronger than can be captured by pairwise modelling.

To model this spatial dependence we go part of the way to multidimensional copulas by introducing stronger dependence structures for high rainfall values than low ones. To test this congregating property by using bivariate correlations is not enough. To ascertain whether rainfall data and simulations can capture the higher dimensional spatial dependence structure, we have devised a spatial statistic based on the joint probabilities of rainfall behaviour, above certain thresholds, measured at the vertices of nearly equilateral triangles over a range of sizes in the gauge network. The measure of association depends on the entropy of the 3-variate 2-state probabilities of each triangular set, defined by a quantile threshold of choice. We chose this configuration because acute angled triangles, in the limit, degenerate to straight lines, destroying the spatial property we wish to explore. We choose to call this behaviour "congregating" rather than use the word "clustering" which has other connotations.

By congregation, we mean that on a particular day of rain on a region, the wet gauges will tend to group in a small number of sub-regions, with the interior of each of the congregations experiencing more rainfall than the edges fringed by dry gauges, as observed in the records. We suggest that the multisite copula-based model will dynamically model that feature better, because the wet-dry process and the amount process are jointly described by the dependencies captured by the multivariate copulas. This is the important innovation of the modelling procedure introduced herein.

Turning to the detail of jointly modelling occurrences and amounts of rainfall at more than one site, Herr and Krzysztofowicz (2005) developed a normal quantile based procedure for modelling pairs of rain gauge records, in particular they separately modelled the dry and wet probabilities and the marginal and joint wet distributions. In this paper, we use that approach as a point of departure to model many sites through the use of multivariate copulas which implicitly capture the pair-wise joint wet and dry processes as well as the mixed wet and dry possibilities. The approach goes beyond the combination of Markov chains for occurrences and the separate generation of correlated amounts for the wet sites used previously by most authors quoted above. These are typified by Srikanthan and Pegram (2009), who used a two-tiered model: a multi-site bivariate Markov chain to model the wet-dry occurrences and a multivariate Gamma AR(1) process to model the jointly wet events. Although that model satisfactorily mimics and reproduces the historical daily, monthly and annual statistics, there is no feature built in to that approach to mimic the congregation of high values, other than cross-correlation.

The novelty in the copula-based application is that both occurrences and amounts are generated from the same interdependence process, which was an idea that Serinaldi (2009) published while this paper was being submitted. In the procedure adopted in this paper, in simulating a rainfield on a particular day, a set of hidden correlated normal Y-vectors is generated via a multi-site autoregressive process, dependent on each other and on the previous day's values. These are transformed to uniformly distributed copula variables through the non-linear V-transform (to be introduced in the sequel) and in turn using these quantiles, the field values Z (occurrences and amounts) are obtained, with the congregation of similar quantiles of wetness at gauges being implicitly modelled.

The paper is structured as follows after this introduction: the problem is defined and copulas are explained; the copula-based multi-site time series model of daily rainfall is outlined theoretically and the theory behind the entropy-based congregation measure is developed; comparisons between (i) the observed and (ii) sets of simulated replicate daily data are made; the congregation behaviour of the modelled rainfall is compared with the observed through innovative entropy calculations; conclusions are drawn.

2 Problem definition

2.1 Description of the data to be modelled

The sites of the records studied in this paper appear in Fig. 1 in a region which has non-homogeneous meteorological characteristics, effectively a mesoscale sized area in Baden-Württemberg in South-West Germany. 32 stations with daily records from 1958 to 2001 were selected and are located around the Black Forest on the west and east side of the mountains, chosen particularly because of the heterogeneity of the statistics of the records.

The resolution of the data is 0.1 mm, so that any observation less than this threshold is taken to be a dry day. Examining such a record reveals intermittency of rainfall, in both space and time. From experience and observation, clouds produce highly variable precipitation in random congregations. These events range from intermittent, locally intense, and fragmentary convective ones, to long-lasting, well spread, persistent, stratiform ones. The statistics of the daily raingauge observations of the observed events reflect these characteristics, occurrences being isolated or congregated in space and time, while the amounts are relatively highly skewed.

The spatial dependence structure of daily rainfall is more complex than can be modelled by a simple correlation coefficient over the range of observed values. Where the observed values are low, they tend to be scattered and intermittent, exhibiting a poor spatial dependency. By contrast, where the rainfalls are high, they tend to be spatially and temporally more dependent. This interdependence which is related to the amount of rain needs an appropriate technique to describe it. The choice here is the multivariate copula; only a short introduction on copulas is given here; readers interested in more general details are referred to Joe (1997), Nelsen (1999), or Salvadori et al. (2007).

A copula is defined as a distribution function on the n-dimensional unit cube. All marginal distributions are uniform:

$$C : [0,1]^n \rightarrow [0,1] \tag{1}$$

$$C(u) = u_i \text{ if the vector } u = (1,\ldots,1,u_i,1,\ldots,1) \tag{2}$$

For every n dimensional hypercube within the unit hypercube, the corresponding probability has to be non-negative. Copulas and multivariate distributions are linked to each other by Sklar's theorem Sklar (1959). Sklar proved that each multivariate distribution $F(t_1,\ldots,t_n)$ can be represented with the help of a copula:

$$F(t_1,\ldots,t_n) = C\left(F_{t_1}(t_1),\ldots F_{t_n}(t_n)\right) \tag{3}$$

where $F_{t_i}(t_i)$ represents the i-th one-dimensional marginal distribution of the multivariate distribution. If the distribution is continuous then the copula C is unique. Copulas can be constructed from distribution functions, as described by Nelsen (1999):

$$C(u) = C(u_1,\ldots,u_n) = F\left(F_{t_1}^{-1}(t_1),\ldots F_{t_n}^{-1}(t_n)\right) \tag{4}$$

The advantage of using a copula is that it is invariant to strictly increasing monotonic transformations of the variables. Thus the frequent dilemma whether to transform data or not (for example taking the natural logarithms) does not occur in this case.

An interesting and important property of a copula is whether the corresponding dependence is the same for high or low values. A bivariate copula expresses a symmetrical dependence if:

$$C(u,v) = C(1-u,1-v) - 1 + u + v \tag{5}$$

which means that the copula density is symmetrical with respect to the secondary diagonal $u = 1-v$ of the unit square and the copula density c satisfies:

$$c(u,v) = c(1-u,1-v) \tag{6}$$

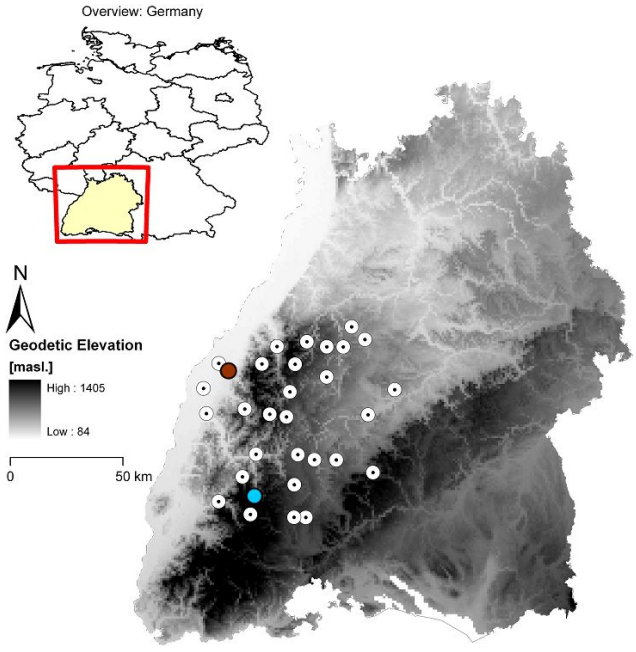

Fig. 1. The locations of the rain gauge stations indicated by circled dots, around the Black Forest within the German state of Baden-Württemberg used in this study; shading darkens with increasing altitude above sea level. Stations 1 and 23 are coloured brown and blue respectively and will be used in the discussion in Section 2.2. The full numbering appears in Fig. 8.

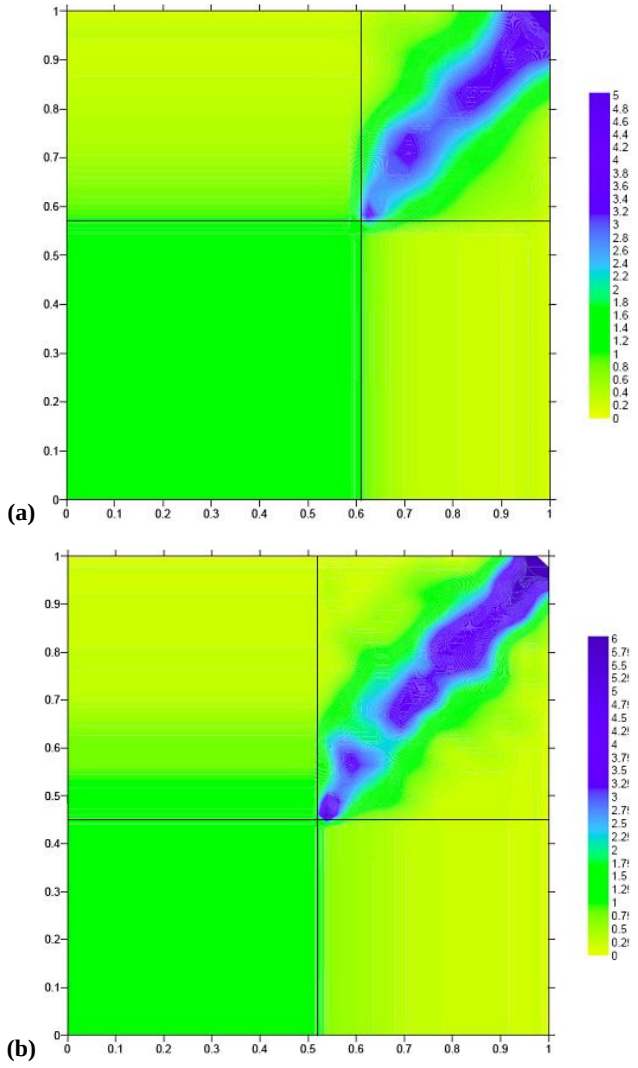

Fig. 2. Two sample copulas for station pairs 1 and 23, the upper panel **(a)** for Summer (June to August) and the bottom panel **(b)** for the relatively wetter Winter (December to February); see Fig. 10 for station locations and Table 1 for $P[0]$ values in January and July, which are close to the thresholds in these figures.

2.2 Examples of empirical copulas of daily rainfall

Copulas are most usefully visualised as bivariate densities rather than bivariate cumulative functions of joint probability on the space $[0,1]^2$. In Fig. 2 we show two examples of empirical copulas derived from scatter-plots of pairs of daily recording rain gauges. The empirical copula densities sampled from the pair of stations 1 and 23 (on opposite sides of the divide as will be seen in Fig. 1), appear in Fig. 2a and b, respectively for summer (June, July, and August) and for winter (December, January, and February). The horizontal and vertical lines indicate the probability limits for the dry/wet boundaries, both evidently drier in summer than in winter. These figures exhibit a constant density for the "both dry" condition in the plateau defined in the lower left corner. The upper left and lower right quadrants show the one-dimensional marginal densities of the wet gauge given that the other is dry, thus one can see all the conditional distributions in one figure.

2.3 Examples of models of theoretical copulas for multisite rainfall

Figure 3a shows an example of the bivariate normal copula with a relatively high constant correlation of 0.85, which is seen to exhibit symmetry about both diagonals. We note that any Normal Quantile (scores) transform from arbitrary joint distribution functions to Gaussian, allowing the modelling of the correlation structure of the bivariate Normal to dictate the interdependence, will maintain the bisymmetric nature of the bivariate normal copula. Thus low values will be as well correlated as high values, as is expected in the normal copula. The image in Fig. 3b shows the partitioned version of the normal copula in Fig. 3a.

Evidently there is a need to model the observed dependence structure more carefully than via the normal score transform, whose dependence structure is locked into the bivariate normal copula, independent of any monotonic quantile transform. This fundamental property (of the interdependence between the variables being completely unaffected by the transformations of the field variables' marginal distributions) is not easy to grasp until it is fully understood that the dependence structure displayed by the copula is dependent only on the rank of the observation, not its marginal distribution.

No monotonic transform of the normal distribution is going to yield a copula in character different to Fig. 3. Theoretical models of copulas are difficult to come by, however, Bárdossy (2006) introduced some which provide a basis for achieving the type of asymmetrical behaviour found in empirical copulas, of which Fig. 2 is an example.

Meta-elliptical copulas Fang et al. (2002) offer another alternative for the description of the dependence structure of precipitation. These copulas however are radially symmetrical and, except for the special case of the normal distribution, cannot be parameterized to represent independent marginals.

In our first choice to achieve asymmetry, a hidden standard normal density is folded about its origin by a modular transform (and variants) that ensure that highly negative and positive values occupy the same location on the copula density space. The new V-copula model chosen was introduced by Bárdossy and Li (2008) and is based on the modulus transform, where all negative values are mapped onto their corresponding positive values. An example of such a copula is shown in Fig. 4c (a partitioned version of Fig. 4a), which will be seen to reflect the desired characteristics of the sampled spatial daily rainfall copulas in Fig. 2. Figure 4b which has different parameters does not exhibit the "waist" appearing

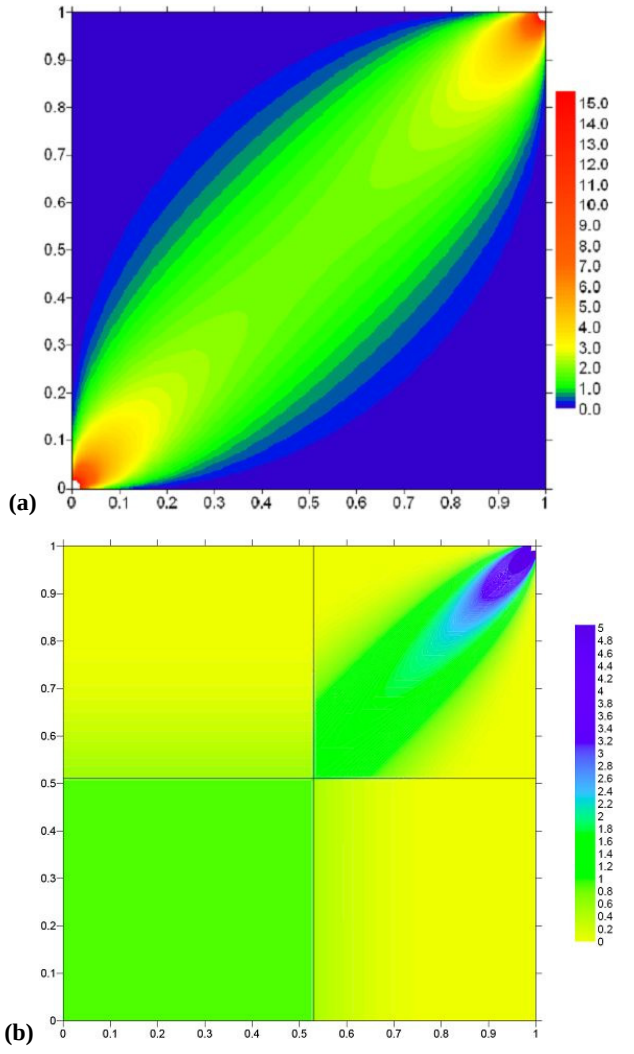

Fig. 3. (a) Bivariate normal copula density with ρ=0.85, the lower panel **(b)** is the partitioned version of the upper panel (a), where P[zero rainfall]=0.51 in each case.

in Fig. 4a and in the copulas of data of Fig. 2a and b. The parameters will be defined in the theoretical development in the next section.

In contrast to the work of Serinaldi (2008) who concentrated on the upper-tail dependence structure of 2-copulas, we are concerned about the joint interdependence of congregations of rainfall stations. Congregation behaviour is the feature of widespread flood-causing rainfall and it is this feature which we do not want to lose in the rainfall model. We are particularly interested in the joint modelling of multisite daily rainfall and would eventually like to model the spatial behaviour, not as a conglomerate of bivariate relationships as has been done with the multinormal, but as jointly multidimensional copulas as was done by Bárdossy and Li (2008). However, the copula model we chose for this rainfall network

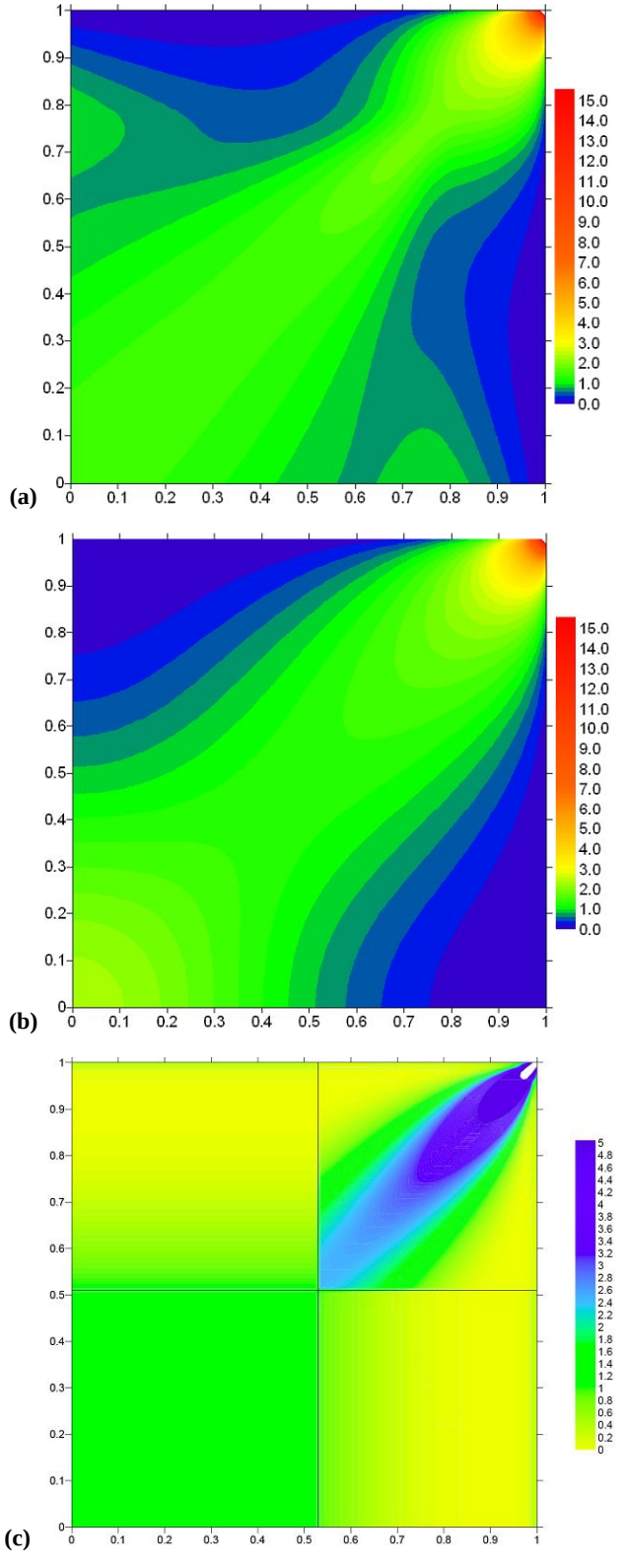

Fig. 4. Copula densities using the v- transform copula model with hidden normal density: **(a)** $k = 4.0$, $m = 0$, $\rho = 0.75$, **(b)** $k = 1.0$, $m = 1$, $\rho = 0.75$, and **(c)** the partitioned version of the v-transformed copula in Fig. 4a.

application is still a multisite bivariate model in principle, as it is predicated on a hidden covariance Gaussian model. To check if this copula-based model can capture spatial dependence better than the conventional covariance based model, we use the entropy of the observations of the triangular triples. This entropy procedure is fully developed in Sect. 3.4.

3 The model

3.1 Theory behind the model

The rainfall observed at site $i = 1, 2, \ldots, n$ on day t is labelled $z_i(t)$. To model these data, we work through three stages: a hidden multisite AR(1) time series model; a non-linear transform to an intermediate variable defining the copulas; the transform of the quantiles defined by the copulas to the simulated field variables.

The multi-site Auto-Regressive lag-1, or AR(1) model, suggested by Pegram and James (1972), based on the seminal paper of Matalas (1967), is used as the driver for the multi-site rainfall model, for both the dry and wet occurrences and the wet amounts, specified as follows:

$$\boldsymbol{y}(t) = \mathrm{diag}\{r_i(t)\}\,\boldsymbol{y}(t-1) + \mathrm{diag}\left\{\left[1-r_i^2(t)\right]^{1/2}\right\}\boldsymbol{a}(t)$$

where:

- $\boldsymbol{y}(t)$ = $\{y_i(t), i = 1, 2, \ldots, n\}$ is a vector of correlated Gaussian variates corresponding to each of the n sites on day t, suitably transformed from the $z_i(t)$ observations.
- $r_i(t)$ = are the serial correlations between the y_i-values and depend on the month in which t falls; they have to be inferred after the transformation by the copula of the $\boldsymbol{z}$-values to the $\boldsymbol{y}$s in the Gaussian domain
- $\boldsymbol{a}(t)$ = $B(t)\boldsymbol{e}(t)$ is a cross-correlated "noise" vector
- $B(t)$ is the "square root" matrix of the cross-correlation matrix $G(t)$ relating the $\boldsymbol{y}(t)$ through the $\boldsymbol{a}(t)$-values during the month corresponding to day t, in the sense that $B^T B = G$
- $\boldsymbol{e}(t)$ is an n-vector of IID standardised Gaussian random variates.

In this model the serial dependence is restricted to each site, while the cross-correlation comes through the noise term $\boldsymbol{a}(t)$. The factorisation of G is achieved by singular value decomposition (SVD) (see for example Press et al., 1992), which is stable even if G is ill-conditioned. Briefly, SVD decomposes G into $G = VWU^T$, where W is diagonal and comprises the singular values. U and V are orthonormal and in the particular case that G is square (the case of a covariance matrix), $U^T V = I$. B may be (not uniquely) computed as $VW^{1/2}U^T$, where $W^{1/2} = diag\{w_i^{1/2}\}$, so that $B = B^T$. In this formulation B is full, not triangular as in Cholesky decomposition which fails when G is ill-conditioned.

The relationship between the *hidden* $y_i(t)$ values and the corresponding rainfall amounts $z_i(t)$ (including the dry days) is developed through the following transform; this relationship is used both for estimation and for simulation.

The transform for the copulas works as follows: y is Gaussian; we define an intermediate variable $s = g(y)$, where

$$\begin{aligned} g(y) &= m - y \qquad \text{if } y < m \\ &= k(y-m)^{\alpha} \text{ otherwise} \end{aligned} \tag{7}$$

Note that if $m = 0$ and $k = \alpha = 1$, then $g(y) = |y|$, so in that case, the non-linear transform $g(.)$ is based on a non-shifted modulus.

Figure 5 shows the relationship, for $m = 1$ and $k = \alpha = 2$ in three stages: the standard normal in the upper panel, then the transformation from y to s in the middle panel and then in the lower panel, a comparison of the densities of the two distributions, the shifted normal and the V-copula-based one: $\phi(m-y)$ and $f_g(s)$. Note the long tail in s in the case of the latter.

Returning to the development, the distribution function of $S = g(Y)$ is thus

$$\begin{aligned} F_g(s) = P[S < s] &= P\left[m - s < Y < (s/k)^{1/\alpha} + m\right] \\ &= \Phi\left[(s/k)^{1/\alpha} + m\right] - \Phi[m-s] \end{aligned} \tag{8}$$

and the density function of S is

$$f_g(s) = \frac{1}{k\alpha}\left(\frac{s}{k}\right)^{\frac{1}{\alpha}-1}\phi\left[\left(\frac{s}{k}\right)^{\frac{1}{\alpha}} + m\right] + \phi[m-s] \tag{9}$$

shown in comparison with the normal density $\phi(y-m)$, with $m = 1$ in Fig. 5.

To picture the relationship, the points where $s = |y|$ are computed and shown explicitly in the middle panel of Fig. 5: we get $P[S < 2.2808] = P[-2.2808 < Y < 2.2808]$, as shown by the intersections of the horizontal and vertical lines.

When $m > 3$, then $\Phi[(s/k)^{1/\alpha} + m] \approx 1$ thus $F_g(s) \approx 1 - \Phi[m-s] = \Phi[s-m]$, so that s approaches Gaussianity in this case.

The quantile defined by the cdf $F_g(s)$ is related to the corresponding rainfall amount through the transform:

$$z_i(t) = F_{z_i}^{-1}\left[F_g(s)\right] \tag{10}$$

where $F_{z_i}[.]$ is the marginal (mixed discrete and continuous) cdf of $z_i(t)$ at the ith site, whose parameters are dependent on the month into which t falls.

The V-transform explicitly models the difference between two rainfall modelling/generating processes; it separates the positive (wet) rainfall values into two distinct component distributions which are separated by a zone corresponding to no precipitation. These lower and upper arms of the V-transform can be interpreted as describing advective/stratiform and convective precipitation respectively. Figure 6 shows the conditional distributions corresponding to the two components – (the parameters are m=1.5, k=2, and α=2). One can

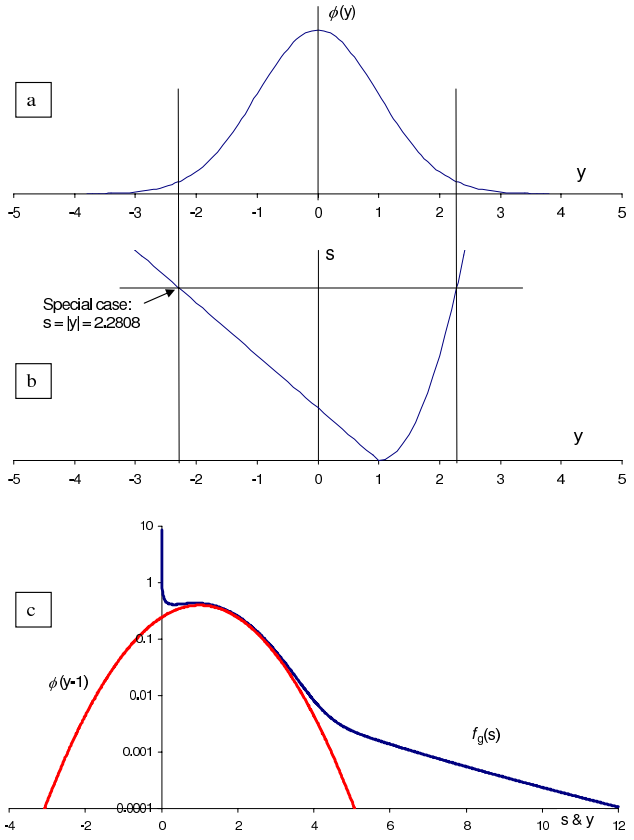

Fig. 5. The copula variable s transformed from the Gaussian y. The value of s indicated by the horizontal line in the middle panel **(b)** of the figure is 2.2808. The vertical lines, through the intersection of the $s = 2.2808$ line with the $s = g(y)$ curves, are the two values of y defining s, and in the upper panel **(a)** they bound the segment of the standard normal distribution integrated to give $P[S < s]$, leading to the cdf $F_g(s)$ of s. The densities $\phi(y-m)$ & $f_g(s)$ appear in the lowest panel **(c)**.

see that the blue distribution corresponding to the high nonlinear transformed values on the right arm of the v-transform produces much higher precipitation values than the red distribution derived from the left arm, which is in fact a segment of the (untransformed) normal distribution. A special analysis (not detailed here) of the correlations within sets of precipitation values generated from the two arms of the V-copula transform, shows that the correlation between the series is different for the two different generating mechanisms. For the above defined parameters with P[0]=0.5 one has a correlation of 0.81 for the precipitation amounts corresponding to the lower arm of the V-transformation and 0.48 for the upper arm. Thus the lower arm represents the advective (or stratiform) processes better than the upper arm, which corresponds to scattered occasionally very intense (or convective) precipitation.

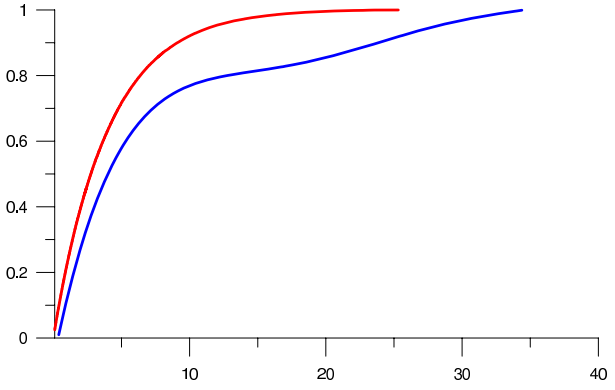

Fig. 6. Division of rainfall amount distribution corresponding to the left arm of the transformation (advective/stratiform precipitation) – red line and to the right arm (convective precipitation) – blue line. Parameters used for this transformation are $m = 1.5$, $k = 2$, $\alpha = 2$, and a Weibull distribution of the precipitation amounts is assumed.

In more detail, we can extract the wet amount and dry information directly from s depending on the dry probability at a site i on a given day t:

$$\begin{aligned} z_i(t) &= 0 && \text{if } F_g(s) < p_i(t) \\ &= F_{w_i}^{-1}\left[\left\{F_g(s) - p_i(t)\right\}/\{1-p_i(t)\}\right] && \text{otherwise} \end{aligned} \tag{11}$$

Here $F_{w_i}(z)$ is the distribution function of precipitation amounts at location i on wet days, $p_i(t)$ is the probability of a dry day at location i. The above definition leads to correct marginals at each observation location.

The interdependence of precipitation amounts is obtained through the interdependence structure of Y. The main difficulty in this representation is that the parameters of the hidden Y and g have to be estimated by searching for the correct parameter set, where Z is the only observed information set; this has to be done by maximum likelihood using nonlinear optimisation.

The spatial dependence between precipitation observations is described by the copula of the multivariate distribution defined by the transformation function $g(y)$ through the correlation matrix G of the hidden Gaussian variables $\boldsymbol{y}$. Note that this copula $C(u_1, u_2, \ldots, u_n)$ is only properly defined for $u_i > p_i(t)$, because of the pole of probability at $(0,0)$ and in addition, the 'conditional' marginal distributions $F_g(u_i, 0)$ and $F_g(0, u_j)$ for $i \neq j$, are spread over the left and bottom quadrants of the copula, as shown in Fig. 2.

To develop the model for spatial dependence, the bivariate marginal copulas describing the dependence between precipitation amounts corresponding to two different locations can be obtained from the g-transformed bivariate normal

distribution of the y-variates. The distribution function is:

$$\begin{aligned} F_g(s_i,s_j) &= P\left[S_i(t)<s_i, S_j(t)<s_j\right] = \\ &= P\left[g(Y_i)<s_i, g\left(Y_j\right)<s_j\right] \\ &= P\left[Y_i<(s_i/k)^{1/\alpha}+m, Y_j<\left(s_j/k\right)^{1/\alpha}+m\right]- \\ &\quad P\left[Y_i<m-s_i, Y_j<\left(s_j/k\right)^{1/\alpha}+m\right]- \\ &\quad P\left[Y_i<(s_i/k)^{1/\alpha}+m, Y_j<m-s_j\right]+ \\ &\quad P\left[Y_i<m-s_i, Y_j<m-s_j\right] \\ &= \Phi_2\left[(s_i/k)^{1/\alpha}+m, \left(s_j/k\right)^{1/\alpha}+m\right]- \\ &\quad \Phi_2\left[m-s_i, \left(s_j/k\right)^{1/\alpha}+m\right]- \\ &\quad \Phi_2\left[(s_i/k)^{1/\alpha}+m, m-s_j\right]+ \\ &\quad \Phi_2\left[m-s_i, m-s_j\right]. \end{aligned} \tag{12}$$

Here Φ_2 is the bivariate normal distribution function, with standard normal marginals and correlation ρ. The corresponding density required for finding the parameters of the hidden AR(1) model, for the case of two sites by maximum likelihood are found by differentiation:

$$\begin{aligned} \partial^2 F_g(s_i,s_j)/\partial s_i \partial s_j &= f_g(s_i,s_j) \\ f_g(s_i,s_j) &= \left\{1/\left(k^2\alpha^2\right)\left(s_i s_j/k^2\right)^{1/\alpha-1}\right\} \\ &\quad \phi_2\left[(s_i/k)^{1/\alpha}+m, \left(s_j/k\right)^{1/\alpha}+m\right]+ \\ &\quad \left\{1/(k\alpha)\left(s_j/k\right)^{1/\alpha-1}\right\} \\ &\quad \phi_2\left[m-s_i, \left(s_j/k\right)^{1/\alpha}+m\right]+ \\ &\quad \left\{1/(k\alpha)\left(s_i/k\right)^{1/\alpha-1}\right\} \\ &\quad \phi_2\left[(s_i/k)^{1/\alpha}+m, m-s_j\right]+ \\ &\quad \phi_2\left[m-s_i, m-s_j\right]. \end{aligned} \tag{13}$$

The n-dimensional joint density can be derived similarly and with large n this soon becomes computationally demanding in the general case. However, we chose the parameters α,m and k of the copula to be the same for all sites which eases the computational burden, but the marginal $P[0]$ values for each station are different; the fitting is done simultaneously for all stations. We note that the estimation of the copula parameters can be obtained by maximising the n-dimensional likelihood function numerically. In order to take the discrete continuous character into acount the likelihood function is built from the n-dimensional version of Eq. (13) for days with precipitation at all sites, while for days with one or more dry stations the corresponding marginals have to be integrated to the limit defined by the dry day probability. Due to the fact that this procedure requires the summation of 2^n terms for each day with observed precipitation, the procedure had to be simplified. For a given triple of parameters m, k, and α the correlations of Y can be estimated using maximum likelihood from the joint bi-variate distributions (taking the zeros also into account). The sum of the log-likelihoods of all pairs is used as an overall likelihood and maximized by varying the triple of parameters m, k, and α. A numerical optimization scheme was used for this purpose.

Although desirable in the long run, we did not perform any estimation of precision of the parameters at this stage.

3.2 Parameter estimation – marginal distributions

The marginal distributions of the positive rainfall amounts were estimated by maximum likelihood for each month at each site. The candidate distributions were the Exponential and the Weibull and that distribution was chosen whose AIC (Akaike (1974)) was a minimum. The characteristics of the two chosen distributions are (Linhart and Zucchini (1986)):

- Exponential distribution:
  - parameter: a (scale)
  - probability density function: $f(x) = (1/a)\exp(-x/a)$
  - maximum likelihood estimator: a = sample mean
  - AIC: $= ln(a)+1+1/n$
  - generate exponential random deviate: $x = -a\ln U$
- Weibull distribution:
  - parameter: a (scale), b (shape)
  - probability density function: $f(x) = (b/a)(x/a)^{b-1}\exp\left[-(x/a)^b\right]$
  - maximum likelihood estimator: $b = \left\{\left(\sum_1^n x_i^b \ln x_i\right)/\left(\sum_1^n x_i^b\right) - (1/n)\sum_1^n \ln x_i\right\}$
  - once b is found by search: $a = \left\{(1/n)\sum_1^n x_i^b\right\}^{1/b}$
  - AIC: $= -\left\{n[\ln b - b\ln a] + (b-1)\sum_1^n \ln x_i - b\sum_1^n (x_i/a)\right\} + 2/n$
  - if $b = 1$, the Weibull reduces to Exponential
  - generate Weibull random deviate: $x = a[-\ln U]^{1/b}$

The AIC values obtained for Exponential and Weibull fits were 2.21 and 2.17, respectively, so the Weibull distribution was chose to model the wet values.

3.3 Parameter estimation – spatial and temporal copulas

The parameters of the copula are defined through the non-linear V-copula transform if the parameters of the multivariate distribution which are used for its definition are defined.

Due to the fact that the distribution function of the precipitation is mixed discrete-continuous, the corresponding copula is non unique. The fit of the copula represents the wet-wet corner well and provides reasonable marginals for the dry/wet distributions. One has to estimate the (global) transformation parameters m, k, and α (one set for each season, for both spatial and temporal interdependence), the individual $12n(n-1)/2$ monthly pairwise spatial correlation coefficients of the underlying multivariate normal random variable and the $12n$ monthly station temporal correlations of the ARMA process.

It is appropriate at this stage to show an example of an empirical copula linking rainfall temporally for a selected gauge; it appears in Fig. 7. In comparison with the spatial copulas of Fig. 2, it appears that the very wet and very light rainfalls are more strongly related than the moderate ones, so that the copula density of the wet quadrant exhibits a saddle-shaped surface.

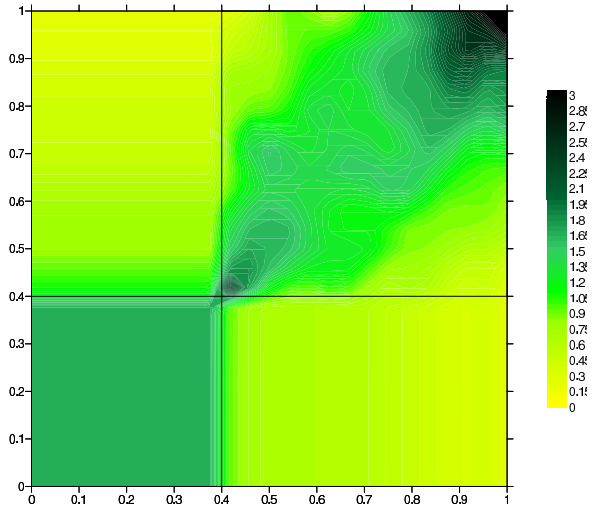

Fig. 7. Sample copula for the temporal structure of daily precipitation at station 14 (December–February); horizontal axis corresponds to day t, vertical axis to day t+1.

3.4 Validation using Entropy of the triple observations

The verification of statistics by intercomparison is a standard prerequisite, but validation of the model by using statistics not employed in the specification needs thought. To determine whether a model produces the same spatial congregation of rainfall values as the observations, something more than pair wise correlation is required for validation. Pair wise correlations can be used for verifying a model but not for independent validation, since they are used in the model definition.

The minimum construct for spatial dependence in the plane is between the values observed at three points at the vertices of a triangle. The shape of the triangle is important and as its size increases one expects the dependence between the observations at the vertices to drop. The triangle cannot be too long and thin, else there will be two points close together or one near the middle of a line joining the other two. The ideal of an equilateral triangle will ensure that there are no ambiguities in mutual distance, since orientation is of secondary importance. However, in randomly scattered sites a compromise is necessary. An approximation to equal sides needs to be allowed for in randomly spaced data points, hence a reasonable constraint on the sides of the triangle needs to be imposed. Because we used Heron's formula for the calculation of the triangle area A, we chose the following criterion for the acceptance of a suitable triple of points.

For each of the three pairs of sides in an adopted triangle, we chose that the maximum difference in a pair must be less than 10% of the perimeter of the triangle; i.e. for sides s_1, s_2, and s_3, $p = s_1 + s_2 + s_3$ and the criterion is: accept triple if $\max|s_i - s_j|/p < 0.1$ for all i and j not equal. Once the triangle has been identified, A is calculated.

To determine the level of association between the values at the vertices in a chosen triangle, we use all the contemporaneous data at the sites to compute the joint wet/dry probabilities. We then determine a given threshold to divide the quantiles into binary sets. The three thresholds chosen for this paper were (i) the wet/dry probabilities (set at 0.5) and (ii) the 0.9 and (iii) 0.975 quantiles, noting that the average of $P[0]$ values determined for the 32 stations used in this study over all months is 0.495.

In more detail, for each triple, the eight binary probabilities $p(i,j,k)$, for $i,\ j,\ k = 1,2$ were calculated over all days of the record, where the states 1 and 2 are the lower and upper partition of the probabilities by the threshold. Thus, for example, the probability that all three gauges on a given day are dry or wet are $p(1,1,1)$ and $p(2,2,2)$ respectively.

The entropy H of each of the sets of 8 probabilities was calculated as a measure of dependence in a given triple; thus

$$H = -\sum\{p(i,j,k)\cdot\ln[p(i,j,k)]\} \quad \text{summed over all } i,j,k \text{ for each triple} \tag{14}$$

The lower the entropy, the greater will be the association between the variables at a given threshold. We checked to see whether either H or $p(2,2,2)$ gave better discrimination between the sets (observations and simulations) and it turned out that H was qualitatively better, benefitting from the extra information contained in the other 7 probabilities. To determine how well the simulations compare with the observations, H was plotted against the square root of the area of the triangle as a measure of mean spacing, $h = A^{1/2}$, for each permissible triangular triple. These results appear in Sect. 5.

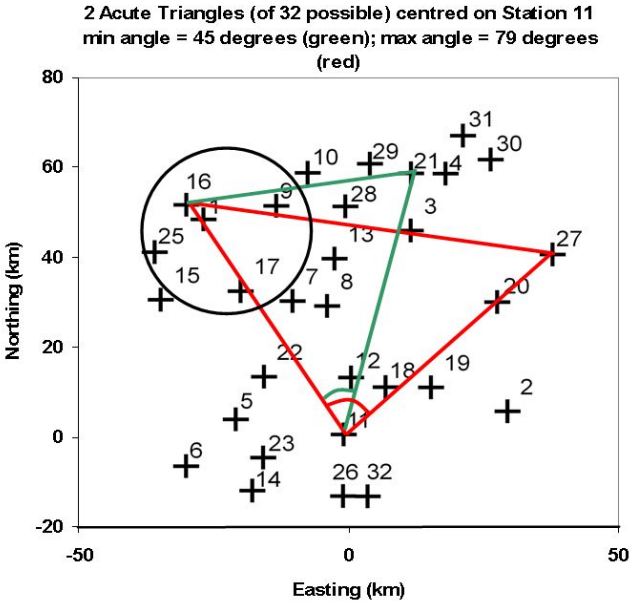

Fig. 8. The locations of the rain gauge stations (with station 11 at the origin) in the Black Forest used in this study, indicated by crosses. The two triangles based on station 11 are those with extreme angles selected from the 32 including station 11; these are used in the entropy calculation for validation, discussed in Sect. 3.3. The circled region includes gauges 1, 9, 17, and 25, whose means appear in Fig. 9.

4 Application of the methods

4.1 Data analysis – derived statistics of the observations

The model described above was applied to the set of selected stations in Baden-Würtemberg in South-West Germany, whose locations are shown in Figs. 1 and 8. Mean precipitation amounts are highly variable in this region. Table 1 shows the estimated January and July dry-day probabilities and the mean precipitation amounts on wet days.

Figure 9 shows the annual cycle of the average daily precipitation by month for a few stations. Note the different annual cycles exhibited by the relatively close stations, which are strongly linked to altitude. Locations 1 and 25 with low elevations both on the west side of the Black forest have a similar cycle, with highest precipitation amounts occurring in early summer. Station 9 at a higher elevation located on the top of the mountains is wetter and has an annual maximum in winter; station 17 at an intermediate altitude has a nearly uniform mean. These differences present a challenge for multisite precipitation modelling.

Marginal distributions and the copula parameters were estimated for each month separately. Parameters of the marginal distributions describing wet amounts were estimated using the maximum likelihood method.

Table 1. Selected rainfall stations; extremes in each column in bold.

			January		July	
Nr	Location	Elevation (m)	$p0$	Mean (mm)	$p0$	Mean (mm)
1	Achern	138	0.41	3.7	0.55	6.7
2	Albstadt Burgfelden	**911**	0.48	4.1	0.55	7.5
3	Altensteig Wart	594	0.43	4.6	0.53	5.3
4	Althengstett Ottenbronn	530	0.47	3.8	0.55	5.4
5	Elzach Oberprechtal	480	0.45	7.0	0.54	9.0
6	Gutach i,Br, Bleibach	302	0.47	5.2	0.54	8.7
7	Freudenstadt Kniebis	875	0.39	10.0	**0.49**	9.6
8	Freudenstadt (WST)	797	**0.36**	8.7	0.50	7.7
9	Forbach Herrenwies	750	0.45	**10.4**	0.52	**10.4**
10	Weisenbach	200	0.46	6.7	0.54	7.1
11	Eschbronn Mariazell	716	0.53	4.9	0.58	6.3
12	Fluorn Winzeln	660	0.52	7.0	0.61	7.4
13	Freudenstadt Igelsberg	757	0.37	7.0	0.51	7.0
14	Furtwangen	870	0.40	9.2	0.50	8.8
15	Offenburg	153	0.47	3.5	0.57	6.3
16	Rheinau Memprechtshofen	**131**	0.49	3.8	0.59	6.5
17	Oppenau	315	0.42	7.3	0.52	8.3
18	Oberndorf/Neckar	516	0.46	5.3	0.53	5.9
19	Rosenfeld	640	0.45	3.8	0.54	5.8
20	Rottenburg, Bad Niedernau	349	0.50	3.3	0.55	5.7
21	Oberreichenbach	639	0.44	5.1	0.53	5.9
22	Wolfach	265	0.46	6.7	0.54	8.4
23	Triberg Nussbach	720	0.44	7.3	0.52	7.8
24	Triberg	683	0.43	9.3	**0.49**	8.8
25	Willstätt Legelshurst	140	**0.54**	3.5	**0.62**	6.8
26	Villingen Schwenningen (NST)	715	0.51	4.7	0.56	6.7
27	Tübingen Bebenhausen	350	0.51	3.2	0.54	6.0
28	Enzklösterle	600	0.43	7.8	0.52	7.4
29	Bad Wildbad Sommerberg	740	0.41	7.2	0.50	7.0
30	Weil der Stadt	389	0.45	**3.0**	0.56	**5.2**
31	Tifenbronn	344	0.47	3.5	0.57	5.6
32	Villingen Schwenningen	720	0.47	4.7	0.53	5.7

Figures 10–14 show typical general behaviour of the data. The dry day probabilities' variation over the months of the year of four gauges introduced in Figs. 8 and 9 (gauges: 1, 9, 17, and 25), ranging from moderately dry to wet, appear in Fig. 10. They fluctuate narrowly about the ensemble mean of 0.495 and showing consistent low temporal variability in this temperate zone.

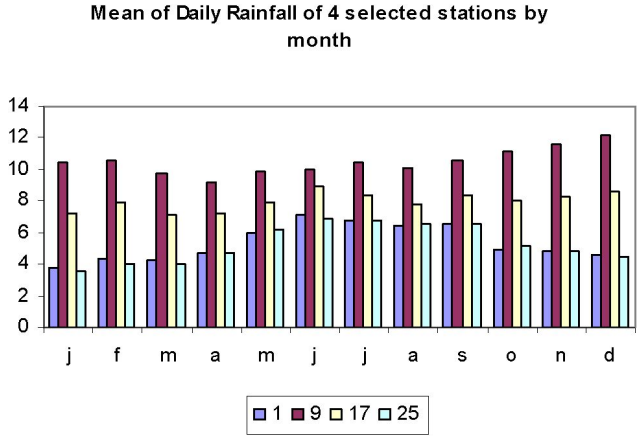

Fig. 9. Observed mean daily precipitation over the year for selected stations from the west of the cluster shown in Fig. 8; we note (by comparison with Fig. 1) that stations 1 and 25 lie in the low altitude area to the West, and that stations 9 and 17, although close, are sited at high and intermediate altitudes, clearly affecting the amount and patterns of rainfall.

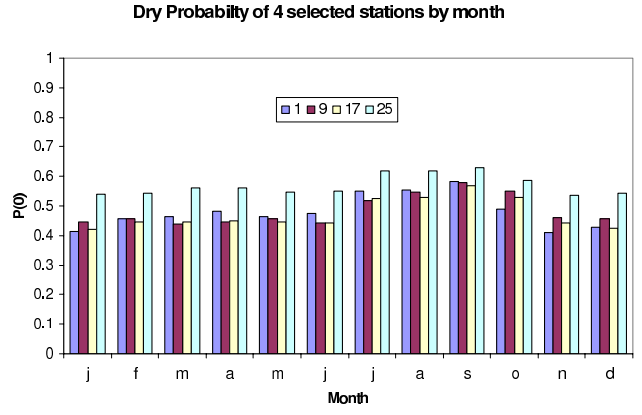

Fig. 10. The variation of the dry probabilities for 4 selected stations by month.

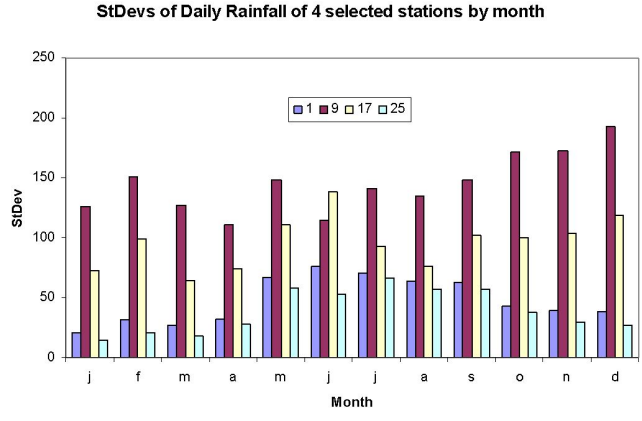

Fig. 11. The standard deviations of daily rainfall for selected stations by month.

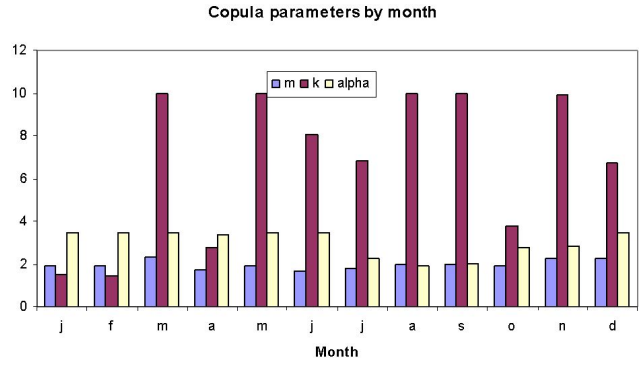

Fig. 12. Parameters of the copula transformation function g.

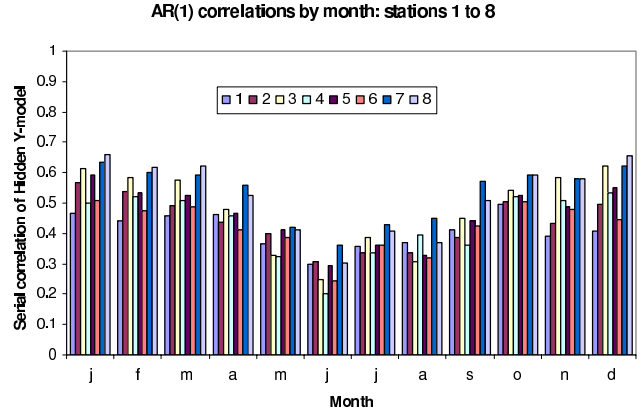

Fig. 13. Serial correlations of the copula transformed y-variates determined from the data for stations 1 to 8.

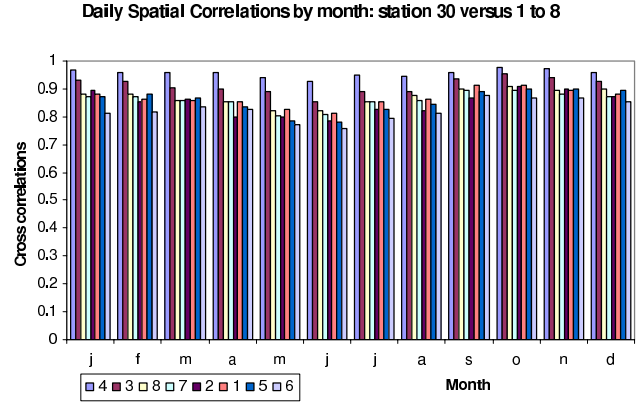

Fig. 14. Selected cross-correlations of the copula-transformed y-variates for station 30 against stations 1 to 8, ranked by distance from station 30.

The standard deviations of wet values of the same stations, shown in Fig. 11, vary in much the same way as the means shown in Fig. 9.

The spatial copula parameters (m, k, and α) are estimated for all gauges in the region each month using the multivariate copula; these appear in Fig. 12. This generalised treatment may need to be refined to allow the behaviour of individual gauges to be reflected in the model, but will add to complexity at this stage. Note the near constancy of the shift parameter m, with all months other than June and July having values of 1.91 and above. June and July experience more thunder-storms than the other months and have m=1.69 and 1.81, respectively. By contrast, the slope k of the right arm of the v-copula shown in Fig. 5, adjusts over the year to accommodate the proportions of stratiform and convective rain, being largest in the summer months. α needs to be constrained and reaches its maximum of 3.5 over the first 6 months and December.

The serial correlations driving the y-variates of the hidden AR(1) multisite model are determined by maximum likelihood through the inverse of the V-copula transform. These temporal copulas are made to model the empirical temporal copulas like that shown in Fig. 7. The stations selected to display these hidden correlations appear in Fig. 13 and lie along the SW-NE diagonal in Figs. 1 and 8, except for stations 1 and 2 which lie respectively in the extreme NW and SE. There is thus a good spread of behaviour as also indicated by the statistics in Table 1. Sites 7 and 8 are close to each other near the middle of Fig. 8 and show the highest (and most similar) correlations.

Spatial correlations of the copula-transformed y-variates seem to depend on distance, as well as altitude. We compute the monthly spatial cross-correlations of the Gaussian y-variates between sites 1–8 against site 30, as shown in Fig. 14, where the sites (1–8) have been ordered by their distance from site 30 (the furthest NE site of the set of 32) which ranges from 10 to 97 km. We note that there is not much dependence of this statistic on annual variability; it varies more between sites.

The values in Fig. 14, which range from 0.76 to 0.98, are considerably higher than the correlations for the rain amounts obtained by the Srikanthan-Pegram (henceforth "covariance") model, which range from 0.40 to 0.89 over the same set of station inter-associations exhibited in the figure. The (hidden) normal correlations for the occurrences (wet/dry) process in the covariance model are much closer to those of the corresponding stations for the copula-based model, ranging from 0.80 to 0.96. These sets of values are not strictly comparable because, in this paper we are using the multinormal correlations of variables reverse-transformed through the copula relationship, however, they are interesting.

It is not strictly fair to compare these with the intersite correlations of the hidden covariance model of Srikanthan and Pegram (2009). However, we will use that model in evaluating the efficacy of the entropy congregation criterion in the next subsection, so it worth outlining its philosophy here. The relevant passage from the abstract of that paper describes the model as a

> multisite two-part daily model nested in multisite monthly, then annual models. A multivariate set of fourth order Markov chains is used to model the daily occurrence of rainfall; the daily spatial correlation in the occurrence process is handled by using suitably correlated uniformly distributed variates via a Normal Scores Transform (NST) obtained from a set of matched multinormal pseudo-random variates, ... a hidden covariance model. A spatially correlated two parameter gamma distribution is used to obtain the rainfall depths; these values are also correlated via a specially matched hidden multinormal process.

Figure 15 shows the scatterplots of the observed and simulated interstation correlations and rank correlations of the rainfall amounts for winter. As one can see, the simulated correlations are slightly lower than the observed ones. In contrast, in the rank correlations there is no systematic difference between the simulated and the observed series. The reason for this is that the copula approach is based on the rank correlations fitted in the transformed domain.

5 Results: comparison of simulations with observations

20 replicates of the historical series (effectively 860 years) were generated and certain statistics of interest were determined from these and compared to the corresponding statistics of the historical series.

5.1 Calculation of distributional statistics

The performance of the model is demonstrated using different uni- and multi-variate statistics. Figure 16 shows the (smoothed) annual cycles of the averages of the historical and 20 simulated mean daily precipitation sequences, for a pair of selected stations (8 and 30), where it is seen that the values are satisfactorily recaptured, thus verifying this aspect of the model behaviour.

To illustrate the performance of the model in the multivariate sense, sets of unconditional and conditional cumulative probability distributions were derived for station 7 with reference to two representatively different stations 8 and 30 (station 7 is close to station 8 but far from station 30) for winter and summer conditions. Due to the fact that small precipitation amounts are usually not as important as large ones, conditionals with thresholds were also considered. These comparisons appear in Figs. 17 and 18, where a detailed comment is given in the figure captions.

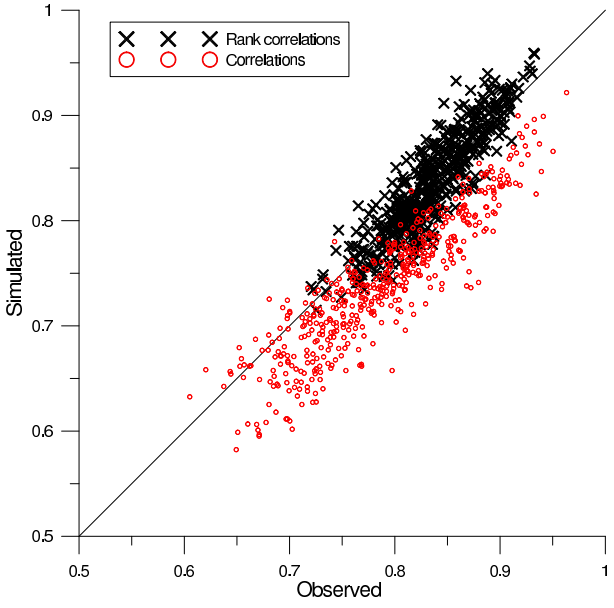

Fig. 15. Cross-correlations (circles) and rank cross correlations (crosses) for the observed and simulated precipitation series.

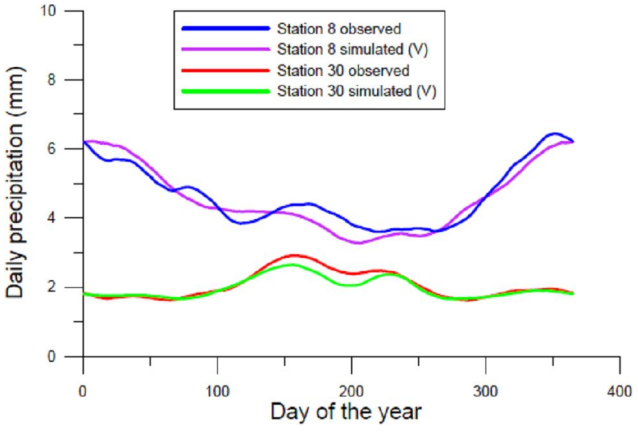

Fig. 16. Comparison of the averages of 20 simulated and the observed mean daily precipitation over the year for stations 8 and 30 – the average of the simulations is smoother, as expected.

The comparisons of observed and simulated conditional and unconditional cumulative distribution functions appearing in Figs. 17 and 18 show that the copula-based procedure captures the joint modelling of wet/dry and amounts processes well. Figure 19 displays the frequency of the numbers of gauges that are wet on any given day during winter; the results for summer are equally good. Again, because we are more concerned about heavier than lighter rainfall, the computations are performed by conditioning the observations on the higher 10 mm threshold. This will mean that there are likely to be more zero counts than if the threshold is set at the observation precision of 0.1 mm.

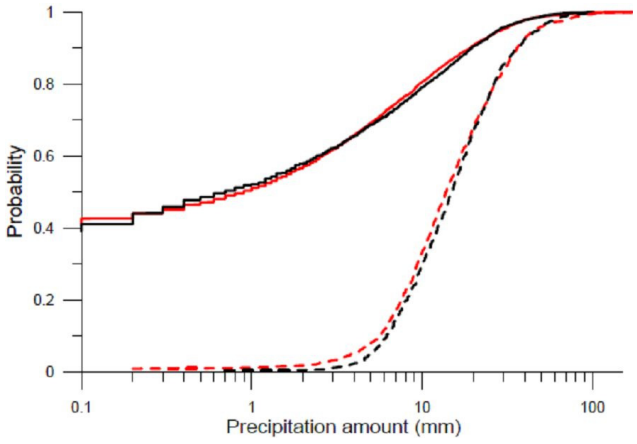

Fig. 17. Observed and (average of 20) simulated cumulative distributions of daily precipitation for station 7 in winter (December–February). The upper solid lines are the unconditional cdfs. The dashed lines show the conditional distributions where the daily precipitation at station 8 is not less than 5 mm. Black lines are the observed and the red lines are the simulated distribution functions using the V-copula. Note that in the upper line (unconditional cdf), the $P[0]$ value of 0.39 for January given in Table 1 is closely captured by the simulated series, obscured slightly by the step at the axis.

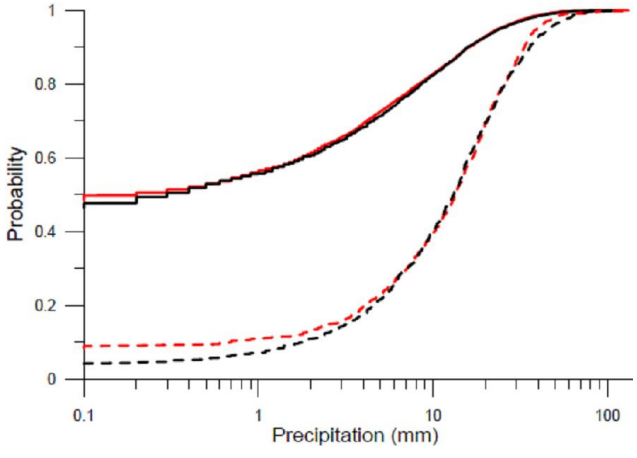

Fig. 18. Observed and (average of 20) simulated cumulative distributions of daily precipitation for station 7 in winter (December–February). The upper solid lines are the unconditional cdfs. The dashed lines show the conditional distributions where the daily precipitation at station 30 is not less than 5 mm. Black lines are the observed and the red lines are the simulated distribution functions using the V-copula. The conditional curves meet the unconditional ones near 70 mm; both the observed and simulated cdfs show a greater probability of lighter rainfall than in Fig. 17, the simulated ones approaching $P[0]$=0.09.

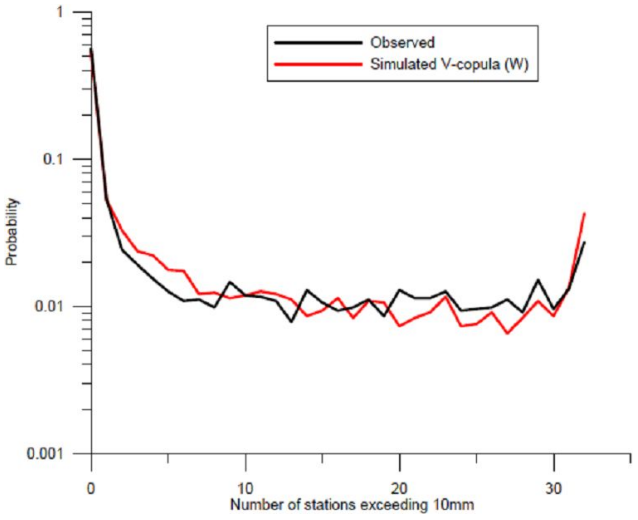

Fig. 19. Distribution of the number of 32 stations exceeding the threshold of 10 mm on a winter day: P[all stations < 10 mm]=0.55.

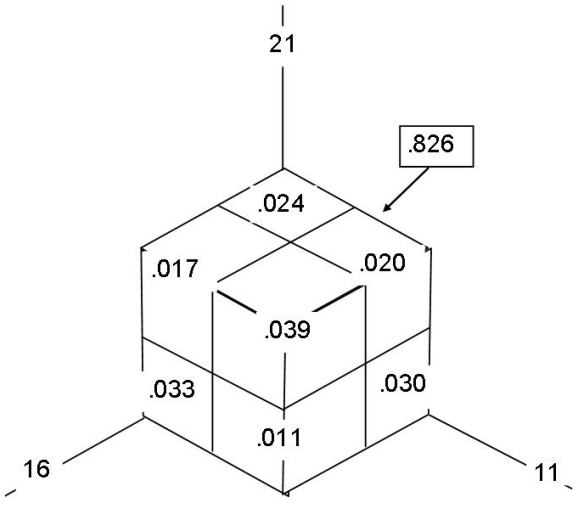

Fig. 20. Isometric view of the 3-D probability space of the three stations (11, 16, and 21) being wet or dry, with the hidden all-dry probability indicated as 0.826.

Table 2. Joint triple probabilities of observations at stations 11, 16, and 21 being below (1) and above (2) quantile 0.90.

Stations			
11	16	21	$p(i,j,k)$
1	1	1	0.8258
1	1	2	0.0239
1	2	1	0.0333
1	2	2	0.0170
2	1	1	0.0300
2	1	2	0.0203
2	2	1	0.0110
2	2	2	0.0387

5.2 Validation using entropy of triples

The calculation of the entropy H, as described in Sect. 3.4, for a given set of data is demonstrated in this section and then applied to the data-sets.

Recall that $H = -\sum\{p(i,j,k).\ln[p(i,j,k)]\}$ summed over all $i,j,k = 1,2$ for a given triple.

The red triangle in Fig. 8 has vertices at the points (11, 16, and 21) and over the record, the observed probabilities of the triple being jointly below or above the 0.90 quantile, (indicated by 1 or 2) in various combinations, are given in Table 2.

The probabilities are obtained by counting the number of occurrences of the three gauges' quantiles being jointly above or below 0.9 on all days for the record of 42 years in the eight patterns.

The diagram in Fig. 20 indicates the arrangement, with the value of the hidden (all dry) probability $p(1,1,1) = 0.826$ as indicated. If the gauged data were independent, $p(2,2,2)$ would equal $0.1^3 = 0.001$ instead of the observed 0.039. The observed relative value is a multiple of 39 greater than the independent one, indicating strong dependence between the high quantiles.

The entropy for the above triple is 0.790, less than the figure of 0.975 which would be calculated in the independent case for a quantile threshold of 0.90; this comparison indicates greater association between the data than independence. In contrast, the entropy of the completely dependent case would be 0.325, indicating maximum association between the data.

To compare the performances of the models in terms of wet/dry discrimination, noting that the average $P[0]$ for the sites is 0.495, Fig. 21 presents the (h, H) plots for the following three sets: (i) data, (ii) copula model, and (iii) covariance model, for the near average wet/dry threshold quantile of 0.5. The points plotted are for each acute triangle which satisfies the condition in Sect. 3.4 that for each of the three pairs of sides, the maximum difference in a pair must be less than 10% of the perimeter of the triangle. The horizontal axis is $h = A^{1/2}$ (A is the area of the triangle) and the vertical axis is H. The entropy has a complete interdependence lower bound of 0.693. This comparison shows that the copula model is only better than the covariance model (with respect to capturing the wet-dry occurrences) at smaller distances. They both fail to capture the inter-association at larger mutual distances approaching 60 km (when $h = A^{1/2}$ approaches 40 km). Although this is an imperfection, it is not a bad as missing the dependence between the high rainfall rates (>90% quantile), because a congregation of large rain-rates has serious hydrological consequences.

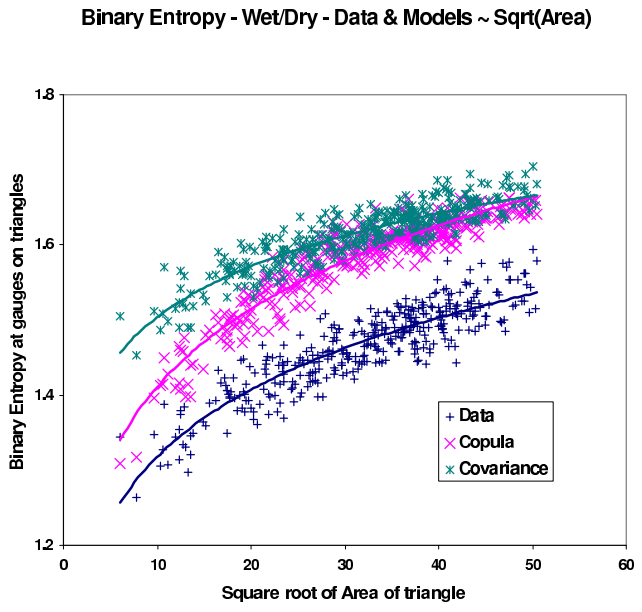

Fig. 21. Entropies of Data and two models, the Copula model and the Covariance model, under the assumption that $P(0)$=0.5, for the purpose of comparing models' ability to model the wet/dry process. The entropy value for complete interdependence is 0.693 when h=A=0.

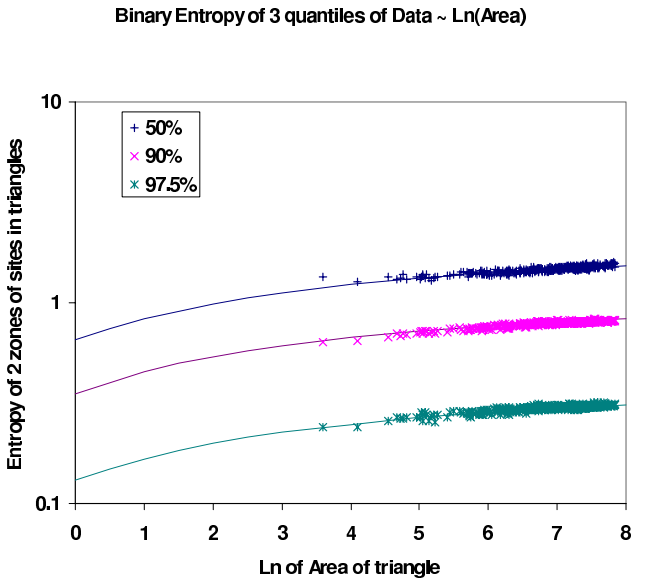

Fig. 22. Entropies of the observed Data at three quantile thresholds: 0.50, 0.90 and 0.975, and suggested convergence to the corresponding limits of complete interdependence.

Turning to the data in isolation from the simulations, Fig. 22 presents the entropy behaviour over the same network at three thresholds: dry/wet assumed to have quantiles of 0.5; 0.90 (repeated from Fig. 21) and 0.975. The entropy lower bounds are 0.693; 0.325; 0.117 for values of marginal quantiles 0.5; 0.9; 0.975, respectively. The fitted trend lines are guides, not suggesting structure, however the convergence of the (h, H) plots towards the corresponding dependence limits is more convincingly demonstrated by using $\ln A$ in place of $A^{1/2}$. The nearly equidistant behaviour of the entropy plots of the data in log-space suggest some underlying useful structure which we have yet to explore.

Finally, in Fig. 23, we present a comparison between the entropy behaviour of the three sets: (i) the data, (ii) the copula model simulations and (iii) the covariance model simulations for the 90% threshold. The Area of the red triangle in Fig. 8 joining points (11, 16, and 21) is 1411 km^2. In Fig. 23 the corresponding point $(h, H) = \left(\sqrt{\text{Area}}, \text{entropy}\right) = (37.56, 0.790)$ appears as a red dot. It is clear that the copula model is closer to the data across the range of h than the covariance model, but is still not close enough for a complete match, indicating that the congregating behaviour of the copula simulations is an improvement over the covariance ones, but still needs attention. In particular, this interdependence would be modelled by a high dimensional copula. However, the discrimination between the sequences offered by the entropy measure is encouragingly sharp.

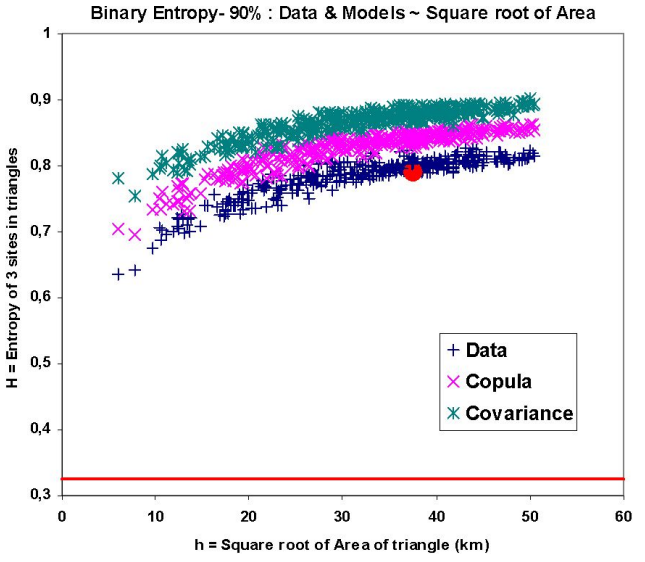

Fig. 23. Entropies of the data and the Copula model and Covariance models, plotted as (h, H). All appear to converge to the value 0.325 of perfect dependence, indicated by the red line lower bound, when $h = A = 0$. The red dot is (h, H)=(37.56, 0.790) calculated at the start of this section, corresponding to the red triangle in Fig. 8.

In summary, the entropy of the eight 3-state binary probabilities calculated by thresholding quantiles at various levels, when applied to the values at the vertices of nearly equilateral triangles, holds promise as a criterion for spatial dependence quantisation. It is clear that in all cases the Entropy statistic indicates stronger interdependence for smaller interstation distances, as expected.

The method indicates that the copula model is superior in this respect to the covariance model to which it was compared, but not yet good enough to suggest that it fully captures the spatial structure of rainfall recorded in networks of gauges. To achieve this will require the application of a truly multi-dimensional copula, not one predicated on a combination of bivariate relationships.

6 Conclusions

The paper set out to establish the nature of the interdependence between rainfall sequences in an inhomogeneous region, to determine an appropriate model. It was found that the classical Normal Scores Transform is not rich enough to capture the range of correlations being strong at high rainfall values and weak at low rainfall amounts. Multinormal variables defined by their correlation structure were nonlinearly transformed, from which set the copulas were derived. The parameters of the transforms and the hidden correlations were obtained by using numerical optimisation to obtain the maximum likelihood. The resulting parameters were used with success in modelling both the rainfall amounts and the occurrences simultaneously.

A novel technique using entropy for determining the degree of congregation of wet gauges was devised which shows that the copula-based model is an improvement over the traditional multisite covariance models, but that it still needs improvement to match the data. Other statistics used for validation, such as cumulative distribution functions conditioned on neighbouring sites experiencing rainfall above a relatively wet threshold of 10 mm, show that these distributions are well mimicked by the copula-based multisite rainfall model.

Acknowledgements. Research leading to this paper was supported by the joint DFG-NRF program, project number Ba-1150/13-1.

Edited by: E. Todini

References

Akaike, H.: A new look at the statistical model identification, IEEE T. Automat. Contr., 19(6), 716–723, 1974.

Apipattanavis, S., Podestá, G., Rajagopalan, B., and Katz, R. W.: A semiparametric multivariate and multisite weather generator, Water Resour. Res., 43, W11401, doi:10.1029/2006WR005714, 2007.

Bárdossy, A.: Copula-based geostatistical models for groundwater quality parameters, Water Resour. Res., 42, W11416, doi:10.1029/2005WR004754, 2006.

Bárdossy, A. and Li, J.: Geostatistical interpolation using copulas, Water Resour. Res., 44, W07412, doi:10.1029/2007WR006115, 2008.

Fang, H.-B., Fang, K.-T., and Kotz, S.: The meta-elliptical distribution with given marginals, J. Multivariate Anal., 82, 1–16, 2002.

Herr, D. and Krzysztofowicz, R.: Generic probability distribution of rainfall in space: The bivariate model, J. Hydrol., 306, 237–264, 2005.

Joe, H.: Multivariate models and dependence concepts, Chapman Hall, Boca Raton, 1997.

Linhart, H. and Zucchini, W.: Model Selection, Wiley, New York, 1986.

Matalas, N.: Mathematical assessment of synthetic hydrology, Water Resour. Res., 3(4), 937–945, 1967.

Mehrotra, R. and Sharma, A.: Preserving low-frequency variability in generated daily rainfall sequences, J. Hydrol., 345, 102–120, 2007.

Mehrotra, R., Srikanthan, R., and Sharma, A.: A comparison of three stochastic multi-site precipitation occurrence generators, J. Hydrol., 331, 280–292, 2006.

Nelsen, R.: An introduction to copulas, Springer Verlag, New York, 1999.

Pegram, G. and James, W.: A Multivariate Multi-lag Autoregressive Model for the Generation of Operational Hydrology, Water Resour. Res., 8, 1074–1076, 1972.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B.: Numerical recipes in Fortran, Cambridge University Press, Cambridge, England, 1992.

Salvadori, G., Michele, C. D., Kottegoda, N., and Rosso, R.: Extremes in nature. An approach using copulas, Springer, New York, 2007.

Serinaldi, F.: Copula-based mixed models for bivariate rainfall data: an empirical study in regression perspective, Stoch. Env. Res. Risk. A., 23, 677-693, 2009.

Serinaldi, F.: Multisite generalisation of a daily stochastic precipitation generation model, Stoch. Env. Res. Risk A., 22, 671–688, 2008.

Sklar, A.: Fonctions de répartition à N dimensions et leurs marges, Publ. Inst. Stat. Paris, 8, 229–131, 1959.

Srikanthan, R.: Stochastic Generation of Daily Rainfall Data at a Number of Sites, CRC for Catchment Hydrology, Technical Report 05/7, 7, 66, Monash University, 2005.

Srikanthan, R. and McMahon, T. A.: Stochastic generation of annual, monthly and daily climate data: A review, Hydrol. Earth Syst. Sci., 5, 653–670, 2001, http://www.hydrol-earth-syst-sci.net/5/653/2001/.

Srikanthan, R. and Pegram, G.: A Nested Multisite Daily Rainfall Stochastic Generation Model, J. Hydrol., 371(1–4), 142–153, 2009.

Thomas, H. and Fiering, M.: Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, Design of Water Resources Systems, Harvard University Press, Cambridge, MA, USA, Chapter 12, 1962.

Wilks, D.: Multisite generalisation of a daily stochastic precipitation generation model, J. Hydrol., 210, 178–191, 1998.