Noname manuscript No.
(will be inserted by the editor)

Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process

Denis Allard · Marc Bourotte

Abstract A problem often encountered in agricultural and ecological modeling is to disaggregate daily precipitations into vectors of hourly precipitations used as input values by crop and plant models. A stochastic model for rainfall data, based on transformed censored latent Gaussian process is described. Compared to earlier similar work, our transform function provides an accurate fit for both the body and the heavy tail of the precipitation distribution. Simple empirical relationships between the parameters estimated at different time scales are established. These relationships are used for the disaggregation of daily values at stations where hourly values are not available. The method is illustrated on two stations located in the Paris basin.

Keywords Stochastic weather generator, composite likelihood, pairwise likelihood, truncated Gaussian process.

1 Introduction

Many crop and plant models used in agriculture and plant pathology require meteorological data at the hourly time scale as inputs (Huber and Gillespie 1992; Caubel et al. 2012). Since hourly variables are costly to measure, record and archive, it is often the case that only daily records are available. A problem often encountered in agricultural and ecological modeling is thus to disaggregate daily values into vectors of hourly values. Hourly temperatures can easily be synthesized from daily maximum and minimum temperatures. For precipitation, disaggregation is less straightforward. In temperate regions such as Western Europe, to account for the random occurrence of precipitations during the day, temporal disaggregation of daily rainfall is best addressed with stochastic approaches.

There is a growing literature on stochastic precipitation generators and stochastic weather generators. They are now key features of climate impact studies, particularly in hydrology and agriculture. Stochastic weather generators are statistical models that aim at simulating quickly and realistically random sequences of atmospheric variables, such as precipitation, temperature, radiation, wind speed and relative humidity. Introduction and history of stochastic weather generators, for which precipitation is one component, have been covered by Wilks and Wilby (1999), Wilks (2010) and Ailliot et al. (submitted). Historically, rainfall occurrence has been described by Markov chains (Katz 1977; Lennartsson et al. 2008; Chen et al. 2012), including non-homogeneous Markov chains (Katz and Parlange 1995; Furrer and

INRA, UR546 Biostatistics and Spatial Processes (BioSP)
Site Agroparc, 84914 Avignon, France
E-mail: {allard,bourotte}@avignon.inra.fr
Fax: +33 4 32 72 21 82

Katz 2007; Ailliot and Monbet 2012), semi-Markov models (Racsko et al. 1991; Semenov et al. 1998) and multi-state Markov chains (Flecher et al. 2010). Another approach is to consider hidden Markov models (HMM) for occurrence (see e.g. Hughes et al. 1999; Zheng and Katz 2008). In all cases, interpretability of the states remains a key feature of these models. During wet days, a large class of distributions can be fitted to rainfall amounts. The Gamma (Richardson 1981; Flecher et al. 2010; Kleiber et al. 2012) and power transform (Ailliot et al. 2009; Zheng and Katz, 2008) are among the most widely used transform functions.

Another type of approach (see e.g. Allcroft and Glasbey, 2003; Thompson et al., 2007; Ailliot et al. 2009, Kleiber et al. 2012) is to define a latent Gaussian variable for which dry conditions correspond to censored values below a given threshold. Positive rainfall are generated by a transform of the Gaussian value above the threshold. This approach is parsimonious because a single latent variable models simultaneously the occurrence of rainfall and its intensity. To transform the Gaussian values above threshold, Kleiber et al. (2012) used a Gamma density. Thompson et al. (2007) and Ailliot et al. (2009) used a power function, but Allcroft and Glasbey (2003) reported that a power transformation is not adequate to achieve normality. They use a quadratic function of the power transformed rainfall. We found that this model is adequate for low and moderate rainfall, but not quite adequate for the most extreme amounts. In Lennartsson et al. (2008), a generalized Pareto distribution (GPD) modeled heavy rainfall above a high level. In Furrer and Katz (2008), a stretched exponential distribution was used as an alternative to the GPD. These models are promising, but they include many parameters.

Our ultimate goal is the disaggregation of daily precipitations into hourly values. Statistical models and their associate simulation algorithms, which constitute precipitation stochastic generators, are clearly well suited to the stochastic disaggregation of precipitation data. Hansen and Ines (2005) disaggregated monthly rainfall into daily values which were modeled using Markov chains for occurrence and mixture of exponential distribution for rainfall amounts. Bürger et al. (2008) proposed to include a temperature dependence in the multiplicative cascade model of Olsson (1998) for reproducing the Clausius-Clapeyron relation between heavy short-term precipitation and temperature. Hasan and Kunn (2010) used a simple Poisson-gamma model for modeling rainfall occurrence and amount simultaneously. Allcroft and Glasbey (2003) proposed to disaggregate spatio-temporal daily rainfall data with a Gaussian censored latent variable.

Here, we seek a model that can fit the data at these different time scales and for which adequate scaling laws from one time scale to the other can be found. To this end, censored latent models seem the most adequate. We will propose a power-exponential function for transforming the Gaussian values into rainfall amounts. It will be shown that this transformation provides a very good fit from the hourly to the daily time scale, including for the highest values. A striking feature of this model is that we found simple empirical relationships between the parameters estimated at different time scales. We illustrate our model and the disaggregation method on precipitation data collected in two stations located in France: Grignon (about 30 km West of Paris; records from 1996 to 2011) and Chartres (about 80 km South-West of Paris; records from 2001 to 2010).

These two stations are located in the main wheat production region in France. They enjoy a typical Western Europe temperate climate, characterized with a fair amount of precipitation during all seasons. We restrict the illustration on spring data, collected from March to May. The model is introduced in Section 2. Parameter estimation, fitting procedures and scaling relationships are given in Section 3. Disaggregation algorithm and results are shown in Section 4. Some discussion is in Section 5.

2 The model

Let $\{Z_i\}$, $i \in 1, \dots, N$ denote a random process modeling precipitation data, denoted $\{z_i\}$, $i \in 1, \dots, N$. Precipitation data are measured at discrete time intervals, $T, 2T, \dots, NT$, where T denotes the time interval between two consecutive measures. It is thus also the time interval during which precipitation is accumulated in the gauge. If $T = 1$h, the data represent hourly measurements. Daily measurements correspond to $T = 24\text{h} = 1$ day. Obviously, each daily rainfall amount must be equal to the sum of the corresponding 24 consecutive hourly measurements. We consider a discrete latent $(0, 1)$ Gaussian process $\{Y_i\}$, $i \in 1, \dots, N$ defined at the same discrete time intervals, with temporal correlation function $c(\cdot)$. The model is fully characterized by two functions: i) the function that transforms the Gaussian values into precipitations and ii) the correlation function.

2.1 The transform function

We will assume that the transform function is of the form $\psi(y)I(y \geq y_0)$ where $I(y \geq y_0)$ is the indicator function equal to 1 if $y \geq y_0$ and equal to 0 otherwise, and where ψ is a strictly increasing function: $Z_i = \psi(Y_i; \theta)$ if $Y_i > y_0$, and $Z_i = 0$ otherwise, for $i \in 1, \dots, N$. θ is the vector of parameters of ψ. In Ailliot et al. (2009) the transform is a power function, thus ensuring a 1-1 mapping between Y_i and Z_i. In Allcroft and Glasbey (2003), a quadratic form of the power transform is chosen.

We shall take a different route and we will consider a power-exponential transformation of the censored Gaussian values. It was found to offer a very good fit to the data at all time scale considered, see Figure 1. In particular, there is a fairly good accordance for extreme values. Let us denote $\phi(y)$ the probability density function (pdf) of a $(0, 1)$ Gaussian random variable and $\Phi(y)$ its cumulative probability function (cpf). We shall further denote the complementary cpf $\bar{\Phi}(y) = 1 - \Phi(y)$. The model is thus the following

$$Z = \begin{cases} 0, & Y \leq y_0 \\ z_m + b(e^{a[Y-y_0]^c} - 1), & Y > y_0, \end{cases} \tag{1}$$

where z_m is the resolution of the rain gauge, i.e. it is the minimum quantity a rain gauge is able to measure. typically $z_m = 0.5$ mm. The parameter y_0, not included in θ, plays a special role in Eq. (1) because it is directly related to the frequency of dry intervals:

$$P(\text{dry interval}) = P(Z = 0) = P(Y \leq y_0) = \Phi(y_0). \tag{2}$$

The threshold depends on the time scale of the time series. Since dry days are less frequent than dry hours, the threshold y_0 corresponding to daily values will be lower than the threshold corresponding to hourly values.

In the special case $c = 1$, closed form expressions can be derived for the first two moments of rainfall, given that the day is not dry:

$$E[Z \mid Z > 0] = z_m + b\left(e^{-ay_0}E[e^{aY} \mid Y > y_0] - 1\right) = z_m + b\left(e^{a^2/2 - ay_0}\frac{\bar{\Phi}(y_0 - a)}{\bar{\Phi}(y_0)} - 1\right), \tag{3}$$

and

$$\text{Var}(Z \mid Z > 0) = b^2\text{Var}(e^{a[Y-y_0]} \mid Y > y_0) = b^2 e^{a^2 - 2ay_0}\left[e^{a^2}\frac{\bar{\Phi}(y_0 - 2a)}{\bar{\Phi}(y_0)} - \left(\frac{\bar{\Phi}(y_0 - a)}{\bar{\Phi}(y_0)}\right)^2\right]. \tag{4}$$

Equations (3) and (4) were used in Allard (2012) to derive a method of moments estimator for the parameters a and b. The variogram of $Y(\cdot)$ conditional on $Y(\cdot) > y_0$ can also be computed when $c = 1$. Recall that $c(\tau)$ is the correlation function of the latent $(0, 1)$ Gaussian process. Then, from Tallis (1961), technical but otherwise straightforward computations lead to

$$\gamma_{Y|Y \geq y_0}(\tau) = \gamma(\tau) + \frac{\gamma(\tau)^2}{\bar{\Phi}_2(y_0, y_0; c(\tau))}\left\{y_0 f(y_0)\bar{\Phi}(y_0^*) - [2 - \gamma(\tau)]f_2(y_0, y_0; c(\tau))\right\}, \tag{5}$$

with $y_0^* = y_0\sqrt{(1 - c(\tau))/(1 + c(\tau))}$, and ϕ_2, and $\bar{\Phi}_2$ being the pdf and the complementary cpf of a $(0, 1)$ bivariate Gaussian vector with correlation $c(\tau)$.

2.2 The correlation function

We will consider several models of temporal correlation, with the requirement that parameters should be easy to identify, and that we should be able to establish relationships between parameters at different time scale. We will thus stick to well established models with a low number of parameters. Our first model will be a Matérn correlation function

$$c_1(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{|\tau|}{r}\right)^{\nu} \mathcal{K}_\nu\left(\frac{|\tau|}{r}\right), \quad \tau \in \mathbb{R},$$

where $\mathcal{K}_\nu$ is the Bessel function of second kind, with smoothness parameter ν and range parameter r. In temperate regions, the regularity of the rain process, and hence that of the latent process, is close to continuity with absence of differentiability. We thus expect ν to be less than or close to 0.5, the value that corresponds to an exponential covariance function. We also expect the rain process to exhibit different time ranges, the shorter one corresponding to typical rain events within rainy conditions, while the longer one would correspond to the succession of rainy conditions. Our second model is thus a weighted sum of two exponential models:

$$c_2(\tau) = w \exp(-|\tau|/r_1) + (1-w)\exp(-|\tau|/r_2), \quad \tau \in \mathbb{R},$$

where r_1 and r_2 are respectively the short and long range parameters and $w \in [0,1]$. Note however that this work being focused on the disaggregation of daily data, we are mainly concerned with time scales shorter than 24h. It is therefore possible that a single range is enough. Our third model is thus a simple exponential model:

$$c_3(\tau) = \exp(-|\tau|/r), \quad \tau \in \mathbb{R}.$$

3 Estimation of the parameters

The threshold y_0 is estimated by simply inverting Eq. (2): $\hat{y}_0 = \Phi^{-1}(p)$, where p is the proportion of dry measurements. The other parameters are estimated with a two stage procedure. The parameters of the transform function are first estimated with a marginal likelihood procedure, i.e. by considering the rainfall values as independent. Precipitation amounts are then transformed into Gaussian values by inverting Eq. (1). In the second step, the parameters of the correlation function are estimated using a composite likelihood approach. It was found that ignoring the temporal dependence in the first step did not change the estimates by more than 0.1% of their values while allowing a faster and more accurate convergence at the second step, in particular for the more complex models c_1 and c_2.

3.1 Estimation of the transform with marginal likelihood

For a given interval T and a given discretization z_m, the parameters $\theta = (a,b,c)$ of the model in Eq. (1) are estimated by simple maximum likelihood. The log-likelihood

$$l(\theta) = \sum_{z_i > z_m^*} \left\{ -\frac{\psi_\theta^{-1}(z_i)^2}{2} - \frac{1}{c}\log a - \log c - \log(z_i - z_m^* + b) + (\frac{1}{c} - 1)\log\left[\log(\frac{z_i - z_m^*}{b} + 1)\right]\right\}, \quad (6)$$

with $\psi_\theta^{-1}(z) = \{a^{-1}\log[b^{-1}(z - z_m^*)] + 1)\}^{\frac{1}{c}} + \hat{y}_0$, is numerically maximized by calling the function `dfoptim` in `R`. For very low precipitations, this discretization effect is proportionally very important (see top plots in Figure 1). Some values recorded at 0 correspond in reality to precipitations less than or equal to z_m. Conversely, a recorded value equal to z_m can correspond to successive rainfalls less than or equal to z_m, but whose sum is larger than z_m. The value $z_m^* \leq z_m$ is adjusted to account for the discretization effect.

Rain gauges record cumulative amounts that are multiples of z_m. It was found that the best fits were obtained for $z_m^* = 0.375$. From now on, we will set z_m^* to this value.

For the spring series in Grignon, the estimates are $\hat{y}_0 = 1.68$ and $\hat{\theta} = (3.35, 0.075, 0.47)$ for hourly values and $\hat{y}_0 = 0.41$ and $\hat{\theta} = (5.15, 0.026, 0.33)$ for daily values. Note that the parameter c decreases as the time interval increases, as a consequence of the more extreme behavior of daily data resulting from the aggregation of time correlated hourly values. The top plots in Figure 1 show the QQ-plots of the Gaussian values obtained for the spring series in Grignon on hourly values (left panel) and daily values (right panel). The agreement is very good, up to the highest value, which lies below the diagonal. In order to validate these fits and to assess the significance of this departure to the diagonal, an ensemble of 99 series of same length was simulated according to the fitted models. The corresponding QQ-plots are depicted in the lower row of Figure 1. The observed values lie on the first diagonal (open circles). The gray lines correspond to the QQ plots of the simulated series. On both panels, the first diagonal lies in the envelope of simulates lines, which demonstrates that the departure of the highest value from the first diagonal observed on the top plots are within the range of the statistical fluctuations. In addition, considering that our application is oriented towards the simulation of hourly rainfalls conditional on daily ones, these slight departures are not of great concern.

3.2 Estimation of the correlation function

As shown in Allcroft and Glasbey (2003), the correlation coefficients between pairs $\{Z(t), Z(t+\tau)\}$, with $\tau \in \{T, 2T, \dots\}$, can be estimated by maximizing the pairwise log-likelihood $\ell_\tau(z_1, \dots, z_n; \theta, \rho_\tau) = \sum_{i=1}^{n-\tau} \ell_{ij}(z_i, z_{i+\tau}; \theta, \rho_\tau)$, where ρ_τ is the correlation coefficient between Y_i and $Y_{i+\tau}$ and where $\ell_{ij}(z_i, z_j; \theta, \rho_\tau)$ takes one of the three forms, depending on whether neither, one or both measures are dry:

$$\ell_{ij}(\theta; \rho_\tau) = \begin{cases} \log \Phi_2(y_0, y_0; \rho_\tau) & \text{if } z_i = z_j = 0 \\ \log\{\Phi(y_0) - \Phi_2(y_0, y_0; \rho_\tau)\} + \log\{\phi(\psi^{-1}(z_j; \theta))\} & \text{if } z_i = 0 \text{ and } z_j > 0 \\ \log\{\phi_2(\psi^{-1}(z_j; \theta), \psi^{-1}(z_j; \theta); \rho_\tau)\} & \text{otherwise.} \end{cases} \quad (7)$$

Estimating independently all correlation coefficients ρ_τ by maximizing independently the pairwise likelihood for each τ would not automatically lead to a valid correlation function $c(\cdot)$. We thus estimate directly the parameters of the correlation models presented above using a likelihood approach. A full likelihood approach would necessitate heavy computations of multiple integrals in order to take into account all possible successions of dry and wet measurements. We will instead use the weighted pairwise log-likelihood approach presented in Bevilacqua et al. (2013), which is a specific case of the composite likelihood approach (Lindsay 1998). The Weighted Pairwise Log-likelihood (WPL) for the parameters of the correlation function is:

$$\mathrm{WPL}(\eta) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{ij} \ell_{ij}(z_i, z_j, \hat{\theta}; \eta) \quad (8)$$

where the values w_{ij} are non-negative weights and η is the vector of parameters of the correlation model. As compared to Eq. (7), there is a slight change of notation, since now the correlation coefficient ρ is obtained from a correlation function with parameters η. There are many possible choices for the weights. The simplest case, which will be chosen here, is when $w_{ij} = 1$ if $|T_i - T_j| \le \Delta$ and $w_{ij} = 0$ otherwise, in which Δ controls both the efficiency of the estimation and the computational efficiency. We will follow Bevilacqua et al. (2013) for selecting the adequate value Δ by seeking the value

$$\hat{\Delta} = \arg \min_{\Delta \in \mathbb{N}^*} \mathrm{tr}(\mathbf{G}_\Delta^{-1}(\eta_{\mathrm{WLS}}); \Delta). \quad (9)$$

The matrix $\mathbf{G}_\Delta$ is the Godambe information matrix associated to the WPL in Eq. (8):

$$\mathbf{G}_\Delta(\eta) = \mathbf{H}_\Delta(\eta) \mathbf{J}_\Delta^{-1}(\eta) \mathbf{H}_\Delta(\eta)', \quad (10)$$

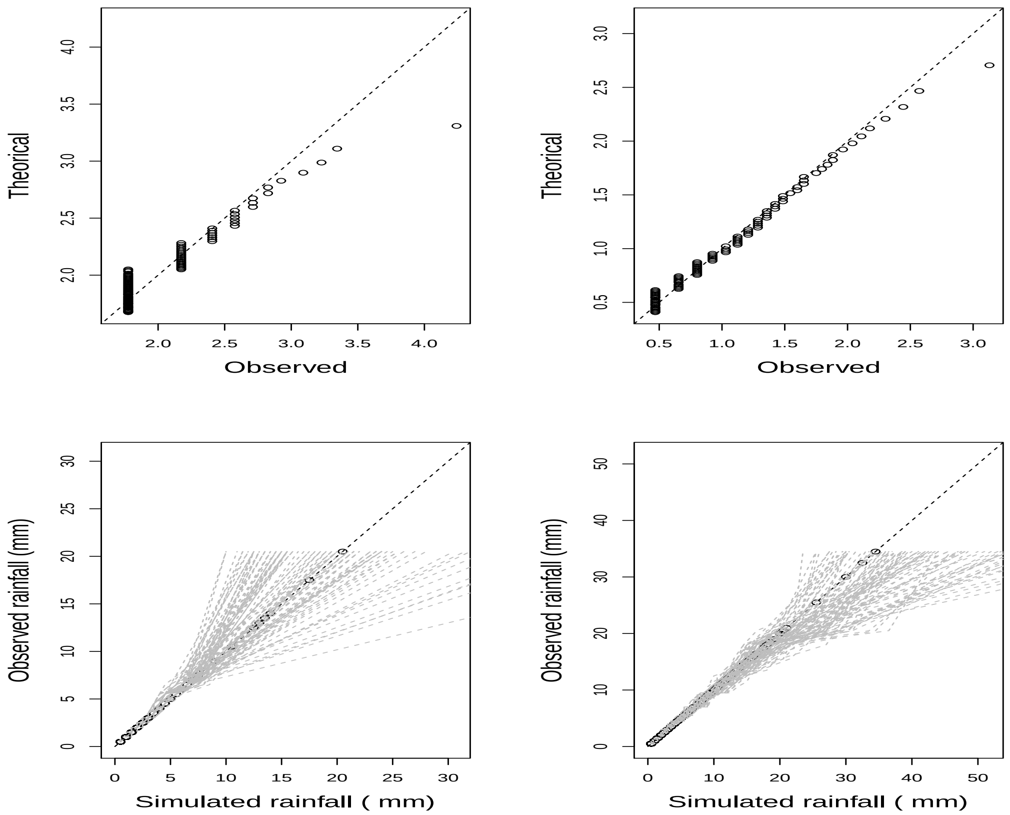

Fig. 1 Top: Q-Q plots of spring rainfall data at Grignon; X-axis: Gaussian quantiles corresponding to the empirical distribution; Y-axis; Gaussian values as given by Eq. (1). Bottom: observed vs. simulated QQ plots of 99 series of data simulated according to the fitted model. Open circles: observed values. Grey lines: simulated series. Left panels: hourly rainfall. Right panels: daily rainfall

where

$$\mathbf{H}_{\Delta}(\eta) = -\mathrm{E}[\mathrm{WPL}(\eta)^{(2)}] \text{ and } \mathbf{J}_{\Delta}(\eta) = \mathrm{E}[\mathrm{WPL}(\eta)^{(1)}\mathrm{WPL}(\eta)^{(1)'}],$$

and $\mathbf{H}_{\Delta}(\eta)'$ is the transpose matrix of $\mathbf{H}_{\Delta}(\eta)$. Here, $f^{(1)}$ means the gradient of f (with respect to the parameters in η) and $f^{(2)}$ its Hessian matrix. The inverse of $\mathbf{G}_{\Delta}(\eta)$ is an approximation of the asymptotic variance of the WPL estimator, $\mathbf{H}_{\Delta}(\eta)$ is the sensitivity matrix of WPL(η) and $\mathbf{J}_{\Delta}(\eta)$ is its variability matrix. The minimum of the trace of the Godambe matrix is sought starting with a classical weighted least square estimation η_{WLS} based on the empirical variogram, as advocated in Cressie (1993) and Chilès and Delfiner (2012). We found consistently that the trace of the inverse of the Godambe matrix first decreases rapidly and then reaches a lower floor value usually located between 12 and 36 hours for hourly precipitations and between 4 and 6 days for daily ones. Then the trace increases slightly, but remains very close to the minimum value. Since this indicator is essentially flat around its minimum and because finding the optimal value is computer intensive we decided to set $\Delta = 24$ hours for all hourly series and $\Delta = 5$ days for daily ones.

Selecting the correct model of correlation is, as it is the case for all model selection problem, a major issue in statistics. Generally speaking, adding parameters to a model leads to better fitting, and thus to increased values of the log-likelihood. But doing so may lead to over-fitting. Bevilacqua et al. (2013), following Varin and Vidoni (2005), proposed to use a Composite Likelihood Information Criterion (CLIC) for selecting the correlation model. The CLIC penalizes the log-likelihood at the maximum with a negative term, equal to $-2\text{tr}(\hat{\mathbf{J}}_\Delta \hat{\mathbf{H}}_\Delta^{-1})$, which increases with the dimension of the matrices, i.e. with the number of parameters in η. On spatial and spatio-temporal simulations, Bevilacqua et al. (2013) showed that CLIC identified correctly the true model among three in a vast majority of cases, from 82% to 96% depending on the model chosen for the simulations, when the number of temporal repetition is equal to 200. One drawback of the CLIC is that the penalization term does not depend on the number of data at hand. The Bayesian Information Criterion (BIC) introduced in Schwartz (1978) is a selection criterion that accounts for the number of data used in the estimation procedure. The penalization term is $-\sharp\eta.\log M$ where $\sharp\eta$ is the number of parameters in η and M is the number of repetitions. Since BIC was proposed in situations where the data are independent, we must adjust the number of data in order to account for temporal dependency. For doing this, we will use the integral range (Lantuéjoul, 2002) defined as the integral of the correlation function

$$A(\eta) = \int_{\mathbf{R}} c(\tau;\eta) d\tau \simeq \sum_{k=-\infty}^{\infty} c(kT;\eta). \tag{11}$$

Then, it can be shown that the equivalent number of independent data is equal to $M = N/A(\hat{\eta})$. Finally, we obtain $\text{BIC} = 2\text{WPL}(\hat{\eta}) - \sharp\eta.\log M$. Table 1 shows the estimated parameters and the selection criteria for hourly and daily rainfalls on the Grignon spring series. For hourly values, the most adequate model is the double exponential according to both selection criteria. Note that under this model, the correlation at 24 and 48 hours are respectively equal to 0.17 and 0.03. For daily rainfall, the Matérn and the double exponential models yield very similar scores. The double exponential is preferred by CLIC (by one unit), whereas the Matérn model is preferred by BIC (by 7 units).

Model	$\hat{r}_1$	$\hat{r}_2$	$\hat{w}$	$\hat{\nu}$	$\hat{A}$	WPL($\hat{\eta}$)	CLIC	BIC
				Hourly rainfall				
Matérn	7.6	—	—	0.28	10.1	-251601	-503218	-503233
Two Exp.	2.5	15.7	0.60	—	15.6	-251574	**-503173**	**-503192**
Single Exp.	5.4	—	—	—	10.8	-251724	-503457	-503463
				Daily rainfall				
Matérn	2.7	—	—	0.17	2.7	-10971	-21942	**-21966**
Two Exp.	0.7	4.7	0.63	—	4.5	-10970	**-21941**	-21973
Single Exp.	1.6	—	—	—	3.3	-10983	-21966	-21977

Table 1 Estimates and selection criteria for the Grignon spring rainfalls.

In order to further validated the estimation procedure, we simulated 49 series of realizations of hourly values of same length as the original series under two models: a single exponential model and the double exponential model selected by CLIC. We then computed the variograms of Z and $Z \mid Z \geq 0$ for the simulated series and for the genuine rainfall measurements. Results are shown in Figure 2. The left panels correspond to the series simulated with the single exponential model, which is unable to reproduce the very short correlations. A double exponential model is able to better account for short and long correlations at the same time. Simulated variograms of Z seem less variable than that of the measured series, while simulated variograms of $Z \mid Z \geq 0$ tend to be slightly more variable than the variogram computed on the positive rainfall amounts.

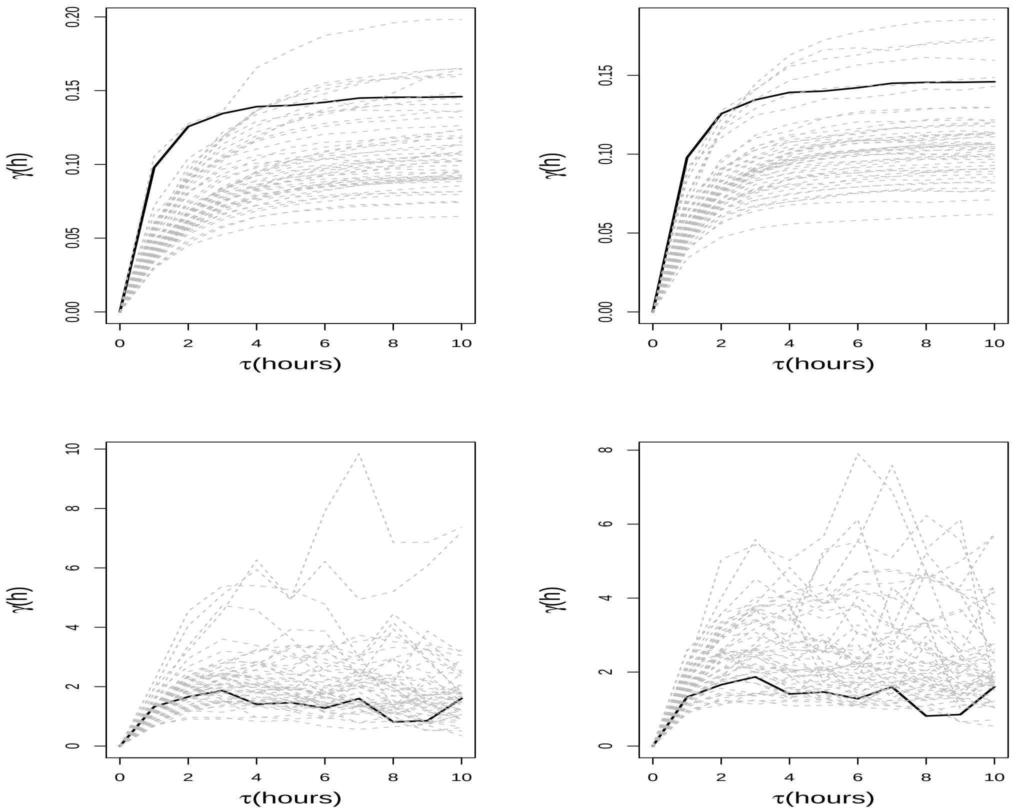

Fig. 2 Top panels: variograms of hourly rainfall amounts of the Grignon spring series (solid black lines) and 49 simulated series with fitted models (dashed grey lines); left panel, single exponential model; right panel, double exponential model; parameters are given in Table 1. Bottom panels: same, computed on hourly non-null rainfall amounts.

3.3 Scaling relationships

When hourly data are available, the data set can be transformed at different time scales. For example, bi-hourly data sets can be generated by adding two consecutive hourly values. The same procedure can be easily applied for other time intervals, thus generating data sets synthetically measured at the following time intervals $T = \{2, 3, 6, 12\}$ hours. Figure 3 illustrates the variations of the corresponding estimated parmaters y_0, a, b, c and r, as a function of T or $\log T$, for the spring series in Grignon and Chartres. As expected, the threshold y_0 decreases as T increases, since the proportion of dry days is lower than the proportion of dry hours. As shown in Figure 3, we found simple empirical relationships between the parameters and T. Similar results were consistently observed on other sites analyzed and for all seasons.

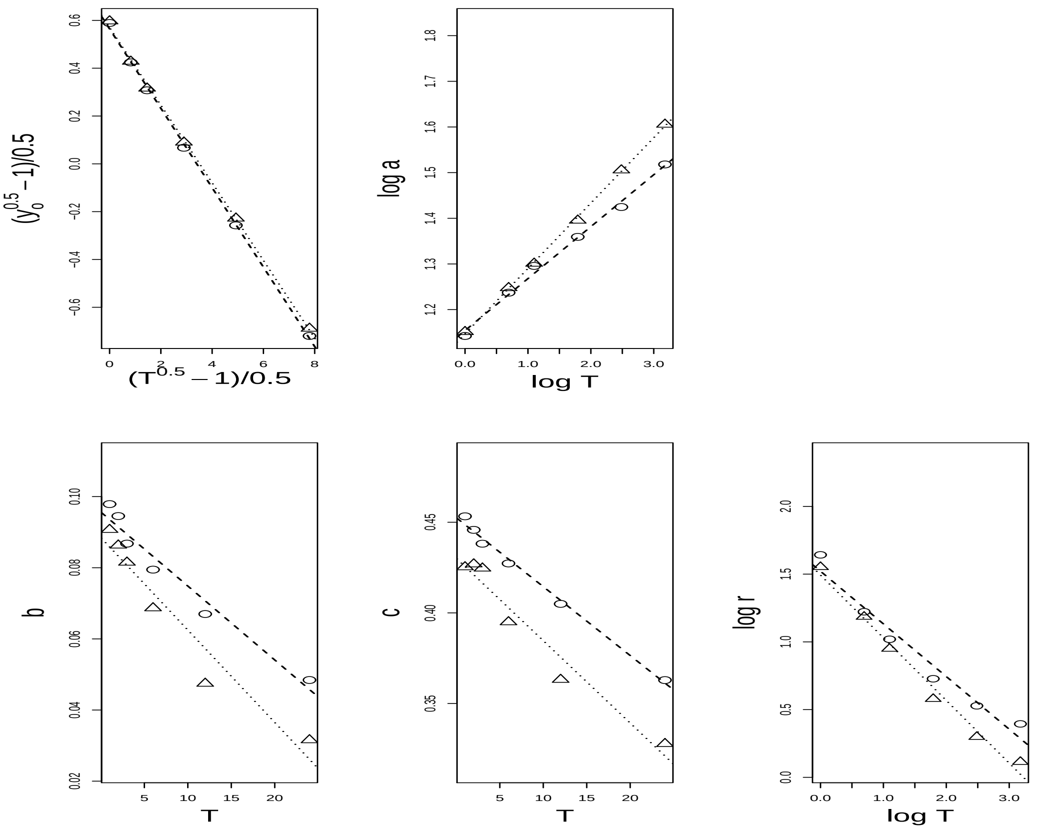

Fig. 3 Empirical Relationships between the parameters y_0, a, b, c and r of the model (1) with respect to the time lag T (in hours). Circles: Grignon. Triangles: Chartres. Dashed line: fitted linear regressions

4 Disaggregation of daily values

4.1 Algorithm

In order to respect the time arrow, daily values, (z_d), $d = 1, \ldots, N$ are disaggregated into hourly values sequentially. Let $\mathbf{z}_d = (z_{d,1}, \ldots, z_{d,24})$ denote the vector of 24 hourly rainfall values of day d and let denote $\mathbf{y}_d$ the corresponding Gaussian values, according to model (1). For day 1, the 24 hourly Gaussian values are simulated conditional on the sum of their transforms being equal to z_1. Then, from day 2 to day N, hourly rainfall amounts are simulated each day d conditionally on the previous ones. The distribution of $\mathbf{y}_d$ conditional on all previous simulated values can be approximated by a less demanding conditioning on the hourly values of the few previous days concatenated in a vector $\mathbf{y}_p = (\mathbf{y}_{d-1}, \ldots, \mathbf{y}_{d-k})$, where k is a small integer. Several values were tested and we found that in our application setting $k = 1$ was sufficient. The distribution $\mathbf{y}_d \mid \mathbf{y}_p$ is multivariate normal:

$$\mathbf{y}_d \mid \mathbf{y}_p \sim \mathrm{MVN}(\mathbf{R}_{dp}\mathbf{R}_{pp}^{-1}\mathbf{y}_p, \mathbf{R}_{dd} - \mathbf{R}_{dp}\mathbf{R}_{pp}^{-1}\mathbf{R}_{pd}), \tag{12}$$

where $\mathbf{R}_{fg}$ is the correlation matrix derived from the estimated correlation function computed between vectors $\mathbf{y}_f$ and $\mathbf{y}_g$, with $f, g \in \{d, p\}$. We need to ensure that the total simulated rainfall each day is close to the measured one. The transformation (1) being non linear, there is no straightforward way of simulating these Gaussian values conditional on the sum of their transform being equal to a given value. The simplest algorithm is a rejection technique: we simply generate vectors from the conditional distribution (12) and accept the proposal if

$$z_d = \sum_{i=1}^{24} z_{d,i} = \sum_{i=1}^{24} \psi(y_{d,i}) I(y_{d,i} \geq \hat{y}_0)$$

is equal to the measured value, up to a pre-specified tolerance ϵ. On dry days, i.e. when $z_d = 0$, the condition becomes $y_{d,i} \leq y_0$, for $i = 1, \ldots, 24$. In other words, a truncated Gaussian vector must be simulated on the orthant $\mathcal{O}_{24} = \otimes_{i=1}^{24}(-\infty, y_0)$ with probability density proportional to that of the multivariate Gaussian vector $\mathbf{y}_d \mid \mathbf{y}_p$ in Eq. (12). The implemented simulation algorithm is thus the following.

```
Algorithm disagg

1. Read the hourly parameters: $a$, $b$, $c$, $y_0$
2. Read the parameters of the covariance function and compute the matrices $\mathbf{R}_{dd}$, $\mathbf{R}_{dp}$ and $\mathbf{R}_{pp}$ from this covariance function
3. Simulate $\mathbf{y}_1 \sim \mathrm{MVN}(\mathbf{0}, \mathbf{R}_{dd}) \mid \{\sum_{i=1}^{24} \psi(y_{1,i}) I(y_{1,i} \geq \hat{y}_0) = z_1\}$
4. For $d = 2, \ldots, N$ :
       Until acceptance:
        i. Simulate $\mathbf{y}_d \sim \mathrm{MVN}(\mathbf{R}_{dp}\mathbf{R}_{pp}^{-1}\mathbf{y}_p, \mathbf{R}_{dd} - \mathbf{R}_{dp}\mathbf{R}_{pp}^{-1}\mathbf{R}_{pd}) \mid \{\sum_{i=1}^{24} \psi(y_{d,i}) I(y_{d,i} \geq \hat{y}_0) = z_d\}$.
       ii. If $|\sum_{i=1}^{24} \psi(y_{d,i}) I(y_{d,i} \geq \hat{y}_0) - z_d| \leq \epsilon$ accept $\mathbf{y}_d$; otherwise reject
```

The acceptance rate can sometimes be very low (as low as 10^{-6}) but it is compensated by fast computations. To accelerate the algorithm, a slightly more complex algorithm can be implemented. First, parameters at intermediate time intervals such as $T = 12, 6$ or 3 hours are estimated, either directly or by using relationships illustrated in Figure 3. The disaggregation is then performed at each time scale in turn, from the coarsest to the finest, by applying the above technique. From one scale to the next, the length of the vectors to be simulated is dramatically reduced (length 2 or 3). At a given time scale, the earlier vectors must always be simulated before the later ones. Conditioning is on all vectors previously simulated (within the current day and the k previous days) at the same time scale. By doing so, the acceptance rate is multiplied by orders of magnitude and the simulation is accelerated.

4.2 Results

As part of the research project CLIMATOR (Brisson and Levrault, 2010), this model was used for disaggregating daily precipitations into hourly precipitations for both measured values (in the past) and large scale climatic model outputs (in the future). These disaggregated precipitations were used as input values for agricultural and plant models. We were faced with two situations, one being much easier than the other. In the easy situation, some hourly values were available, because they had been measured during some period in the past. In this case, all we had to do was to estimate hourly parameters from the hourly time series, and use these estimates for disaggregating daily rainfalls when hourly ones were not measured.

In the more difficult situation, there were no hourly measurements available. In order to obtain hourly parameters from the daily ones, we used "analog" hourly rainfall time series which were measured either at a nearby station, or at a station considered to belong to a very similar climate. To illustrate this case, daily rainfall at Chartres are disaggregated pretending that hourly rainfall are not available. On the Grignon series, parameters were estimated at different time intervals with an exponential covariance function, i.e. according to model $c_3(\tau)$ for $T = 1, 2, 3, 6, 12, 24$ hours. The regression lines shown in Figure 3 were then fitted. On this Figure, the genuine parameters in Chartres, along with the corresponding regression lines are shown. It can be seen that the regression lines at the two stations are very close.

Hourly parameters at the Chartres station are obtained by extrapolating the daily parameters in Chartres, $(\hat{y}_0^d, \hat{a}^d, \hat{b}^d, \hat{c}^d, \hat{r}^d)$, with the slopes of the regression lines computed at Grignon. We obtained the following predicted hourly parameters $(\tilde{y}_0^h, \tilde{a}^h, \tilde{b}^h, \tilde{c}^h, \tilde{r}^h) = (1.70, 3.47, 0.08, 0.42, 3.47)$, which are to be compared to the parameters directly estimated on the hourly precipitation: $(\hat{y}_0^h, \hat{a}^h, \hat{b}^h, \hat{c}^h, \hat{r}^h) = (1.69, 3.16, 0.09, 0.43, 4.72)$.

QQ plots and wet spells of an ensemble of 99 stochastic disaggregations of the spring daily data are represented in Figure 4. Simulated values have been gridded to the same 0.5 mm scale for comparison with measured data. The resulting QQ plots behave as step functions with fixed increments for low rainfall amounts. Nonetheless, observed data are within the envelope of the 99 simulated series. On this example, the fit is better with the predicted parameters (left panel) than with those directly estimated of the hourly data (right panel). This is not always the case, and in general the envelope contains the diagonal. The distribution of the length of rain events is pretty well simulated, except for a slight underestimation of very short rainfalls. Overall, the distribution and the length of the rain events are well respected. Note that neither on the observed data nor on the simulated series there are rain events longer than 18 hours.

5 Summary and Discussion

A method for the stochastic disaggregation of daily rainfall values into hourly ones is presented. It has been routinely applied for disaggregating daily values at 12 sites representative of the French climatic variability that were part of the CLIMATOR project (Brisson and Levrault, 2010). The method relies on a Gaussian latent process, which is transformed into a precipitation process with a power-exponential function. It was found that this transformation function models adequately the body and the tail of rainfall distributions. For space considerations it was only illustrated on the Grignon spring rainfall data, but we found a very good fit on all series analyzed so far. Another consistent finding is that, when analyzing the data at different time scales, we found very good empirical relationships between the parameters and the time scale, see 3. These relationships allow us to predict the parameters of the model at fine time scales that were not measured. Figure 4 illustrates these performances.

We used a sequential algorithm for several reasons. First, from a modeling point of view, it seems natural to respect the time arrow and therefore to simulate one day after the other. Second, taking advantage of this natural order leads to very efficient algorithms. Conditioning each simulated day to all other days (past and future) necessitates to implement an iterative MCMC algorithm which would be orders of magnitude slower.

This model can be extended in several directions. It is often the case that rainfall is not evenly distributed during the day. To account for this, a straightforward generalization is to allow the threshold and the other parameters to vary within the day. External covariates, if available, could also be included in the statistical model. Even though the CLIC selected the double exponential model as the best model for the auto-correlation of the Gaussian process, the scaling relationships were obtained for the single exponential model. Exploring if such relationships hold for more complex correlation function is another possible extension.

More fundamentally, further work must be undertaken to provide theoretical grounds supporting these empirical findings. In particular, the extreme value properties of this transformation must be assessed and the scaling relationships must be explored under a theoretical point of view. The change of support theory, well-known in geostatistics, is certainly a good starting point for this, see e.g. Chilès and Delfiner (2012, ch. 6).

As already pointed out in the Introduction, stochastic disaggregation of precipitation is one particular instance of stochastic weather generators. Thanks to the promising results obtained in this work,

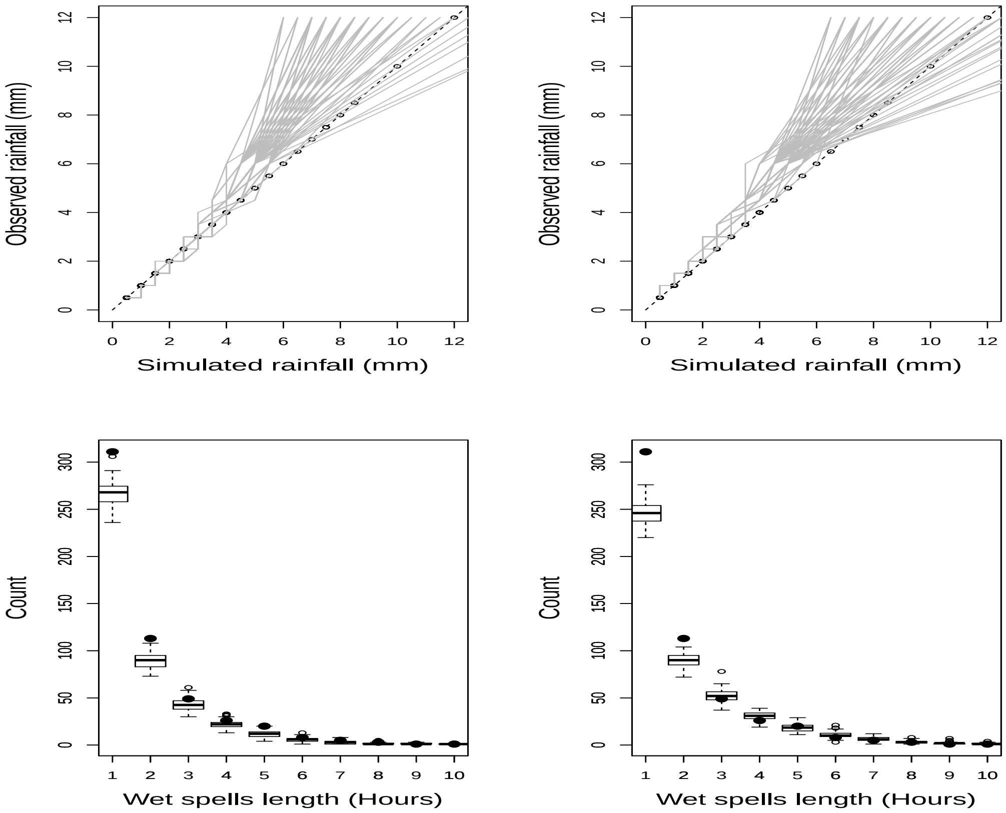

Fig. 4 Top: QQ plot of the 99 disaggregated precipitations vs. observed hourly precipitations, spring precipitations at the Chartres station. Open circles: observed values. Grey lines: simulated series. Bottom: distribution of the wet spells. Black circles: observed series. Boxplot on the 99 disaggregated hourly series. Left: predicted parameters obtained from the regression curves shown in Figure 3. Right: estimated parameters on daily rainfall data.

we believe that latent Gaussian processes with a power-exponential transform are adequate models for precipitation for weather generators.

Acknowledgments We thank Emilio Porcu and Alessandro Fassò for inviting us to join this special issue. Among others, we thank Pierre Ailliot, Valérie Monbet, Philippe Naveau and Carlo Gaetan for fruitful discussions about this work. This work was part of the project CLIMATOR (2007-2010) (`http://w3.avignon.inra.fr/projet_climator/`) funded by ANR (The French National Research Agency). CLIMATOR was lead by Nadine Brisson who left us in Octobre 2011. We dedicate this work to her memory.

References

1. Ailliot P, Allard D, Monbet V, Naveau P (submitted) Stochastic weather generators : a review of weather type models.
2. Ailliot P, Monbet V, (2012) Markov-switching autoregressive models for wind time series. Environ Model Softw 30: 92-101.
3. Ailliot P, Thompson C, Thomson P (2009) Space time modeling of precipitation using a hidden Markov model and censored Gaussian distributions. J Roy Stat Soc, C 58: 405-426.

4. Allard D (2012) Modelling spatial and spatio-temporal non Gaussian processes. In: Porcu E, Montero JM and Schlather M (eds) Advances and Challenges in Space-time Modelling of Natural Events, Lecture Notes in Statistics 207. Springer, pp 141-164.
5. Allcroft DJ, Glasbey CA (2003) A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. J Roy Stat Soc C 55: 1952-2005.
6. Bevilacqua M, Gaetan C, Mateu J, Porcu E (2012) Estimating space and space-time covariance functions for large data sets : a weighted composite likelihood approach. J Am Stat Assoc 107: 268-280.
7. Brisson N, Levrault F (2010) Changement climatique, agriculture et forêt en France: simulations d'impacts sur les principales espèces. Le Livre Vert du projet CLIMATOR (2007-2010). ADEME, Orléans.
8. Bürger G, Heistermann M, Bronstert A (2014) Towards sub-daily rainfall disaggregation via Clausius-Clapeyron. J. Hydrometeorol. doi:10.1175/JHM-D-13- 0161.1, in press.
9. Caubel J, Launay M, Lannou C, Brisson N (2012) Generic response functions to simulate climate-based processes in models for the development of airborne fungal crop pathogens. Ecol Model 242: 92-104.
10. Chen J, Brissette FP, Leconte R (2012) WeaGETS–a Matlab-based daily scale weather generator for generating precipitation and temperature. Procedia Environ Sci 13: 2222-2235.
11. Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Second Edition. Wiley, New York.
12. Cressie NAC (1993) Statistics for Spatial Data. John Wiley & Sons, New York.
13. Flecher C, Naveau P, Allard D, Brisson N (2010) A stochastic daily weather generator for skewed data. Water Resour Res 46:W07519.
14. Furrer EM, Katz RW (2007) Generalized linear modeling approach to stochastic weather generators. Clim Res 34: 129-144.
15. Hansen JW and Ines AVM (2005) Stochastic disaggregation of monthly rainfall data for crop simulation studies. Agr Forest Meteorol 131: 233-246.
16. Huber L, Gillespie, TJ (1992) Modeling leaf wetness in relation to plant-disease epidemiology. Annu Rev Phytopathol 30: 553-577.
17. Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. Appl Stat 48: 15-30.
18. Katz RW (1977) Precipitation as a chain-dependant process. J Appl Meteorol 16:671-676.
19. Katz RW, Parlange MB (1995) Generalizations of chain-dependent processes: Application to hourly precipitation. Water Resour Res 31:1331-1341.
20. Kleiber W, Katz RW, Rajagopolan B (2013) Daily Spatio-Temporal Precipitation Simulation Using Latent and Transformed Gaussian Processes. Water Resour Res 48:W01523.
21. Hasan MM, Kunn PK (2010) A simple Poisson-gamma model for modelling rainfall occurrence and amount simultaneously. Agr Forest Meteorol 150: 1319-1330.
22. Lennartsson J, Baxevani A, Chen D (2008) Modelling precipitation in Sweden using multiple step Markov chains and a composite model. J Hydrol 363: 42-59.
23. Lantuéjoul C (2002) Geostatistical Simulations. Springer, Berlin.
24. Lindsay, B (1988) Composite likelihood methods. Contemp Math:80, 221-239.
25. Olsson J (2008) Evaluation of a scaling cascade model for temporal rainfall disaggregation. Hydrol Earth Syst Sc 2: 19-30.
26. Racsko P, Szeidl L, Semenov M (1991) A Serial Approach to Local Stochastic Weather Models. Ecol Model 57: 27-41.
27. Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17: 182-190.
28. Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:261-464.
29. Semenov AM, Brooks RJ, Barrow EM, Richardson CW (1998) Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Clim Res 10:85-107.
30. Tallis GM (1961) The moment generating function of the truncated multi-normal Distribution. J Roy Stat Soc B 23: 223-229.
31. Thompson C, Thomson P, Zheng X (2007) Fitting a multisite rainfall model to New Zealand data. J Hydrol 340:25-39.
32. Varin C, Vidoni P (2005) A note of composite likelihood inference and model selection, Biometrika, 92: 519-528.
33. Wilks, DS (2010) Use of stochastic weather generators for precipitation downscaling. Wiley Interdisciplinary Reviews: Climate Change 1:898-907.
34. Wilks DS, Wilby RL (1999) The weather generation game: a review of stochastic weather models. Prog Phys Geogr 23: 329-357.
35. Zheng X, Katz RW (2008) Simulation of spatial dependence in daily rainfall using multisite generators. Water Resour Res 44:W09403.