A stochastic daily weather generator for skewed data

Cedric Flecher, P. Naveau, Denis Allard, Nadine N. Brisson

WATER RESOURCES RESEARCH, VOL. 46, W07519, doi:10.1029/2009WR008098, 2010

A stochastic daily weather generator for skewed data

C. Flecher,[1,2] P. Naveau,[1] D. Allard,[3] and N. Brisson[2]

Received 9 April 2009; revised 21 October 2009; accepted 10 December 2009; published 16 July 2010.

[1] To simulate multivariate daily time series (minimum and maximum temperatures, global radiation, wind speed, and precipitation intensity), we propose a weather state approach with a multivariate closed skew-normal generator, WACS-Gen, that is able to accurately reproduce the statistical properties of these five variables. Our weather generator construction takes advantage of two elements. We first extend the classical wet and dry days dichotomy used in most past weather generators to the definition of multiple weather states using clustering techniques. The transitions among weather states are modeled by a first-order Markov chain. Second, the vector of our five daily variables of interest is sampled, conditionally on these weather states, from a closed skew-normal distribution. This class of distribution allows us to handle nonsymmetric behaviors. Our method is applied to the 20 years of daily weather measurements from Colmar, France. This example illustrates the advantages of our approach, especially improving the simulation of radiation and wind distributions.

Citation: Flecher, C., P. Naveau, D. Allard, and N. Brisson (2010), A stochastic daily weather generator for skewed data, *Water Resour. Res.*, *46*, W07519, doi:10.1029/2009WR008098.

1. Introduction

[2] Stochastic weather generators [*Katz*, 1996; *Semenov and Barrow*, 1997; *Qian et al.*, 2005] aim at reproducing the statistical distributional properties of meteorological variables. They have been applied to a wide range of hydrological, ecological, and agricultural studies. For example agronomical models and more specifically crop models need a large variety of daily weather data as inputs [*Wilks*, 1997; *Brisson et al.*, 2003, 2009], to model past, present and future variability for yields. Such daily inputs have to be simulated quickly and easily for long time periods at a given station. In this paper we focus on five variables: minimum and maximum temperatures (T_n and T_x), precipitation P, wind speeds at two meters V and radiation R. The choice of these variables was motivated by the inputs required for the crop models used in a research project (french CLIMATOR project) aimed at exploring the impact of climate change on agriculture in the 21st century. Most other variables that hydrological, ecological and agronomical models may need can be computed from these variables using physically based relations, e.g., relative humidity and potential evapotranspiration. One year typical time series are presented for these variables in Figures 1 and 2.

[3] Conceptually, the majority of statistical weather generators [*Richardson*, 1981; *Richardson and Wright*, 1984; *Semenov and Barrow*, 1997; *Rajagopalan et al.*, 1997] can be classified into two categories. The first one consists in pooling out analog days from a database of past observations according to a given criterion, e.g., with a k-nearest neighbors algorithm [*Rajagopalan and Lall*, 1999]. The main advantage of this nonparametric approach is that the statistical properties of the given database are adequately reproduced. An important drawback resides in the incapability of creating new time series, i.e., unobserved meteorological situations. To alleviate this undesirable feature, the second category of weather generators is based on stochastically drawing random realizations from a statistical model whose parameters have been estimated on a database of past observations. If such parametric or semiparametric models are well built, then most of the distributional characteristics of the studied variables can be reproduced. For example, WGen and LARS-WG, introduced by *Richardson* [1981] and *Semenov and Barrow* [1997], respectively, belong to this class of weather generators. *Apipattanavis et al.* [2007] attempted to combine both categories in a single semiparametric approach. By construction, analog or nonparametric methods are not well adapted to the climate change context. We thus decided to opt for a parametric approach, in which climate change could be accounted for by making the parameters varying. In this paper we present the weather generator for a stationary climate. By this we mean that, even though the parameters of probabililty distributions depend upon the season, they do not change from year to year. Adaptation of the weather generator to climate change is left to further works.

[4] Most parametric weather generators work by defining two daily precipitation states: dry or wet days. The state transitions are classically modeled by a Markov chain [*Semenov et al.*, 1998]. Conditionally on the precipitation state, the other meteorological variables are often assumed to be independently and identically distributed (iid) (e.g., CLIMGEN [*Stockle et al.*, 1998]). More complex models have also been proposed. For example, *Furrer and Katz* [2007] studied a generalized linear model conditioned on rainfall occurrences in order to integrate the ENSO index as a prior information. In contrast to these models for which

[1]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, UVSQ, IPSL, CNRS, CEA, Gif-sur-Yvette, France.
[2]AgroClim, INRA, Avignon, France.
[3]Biostatistiques et Processus Spatiaux, INRA, Avignon, France.

0043-1397/10/2009WR008098

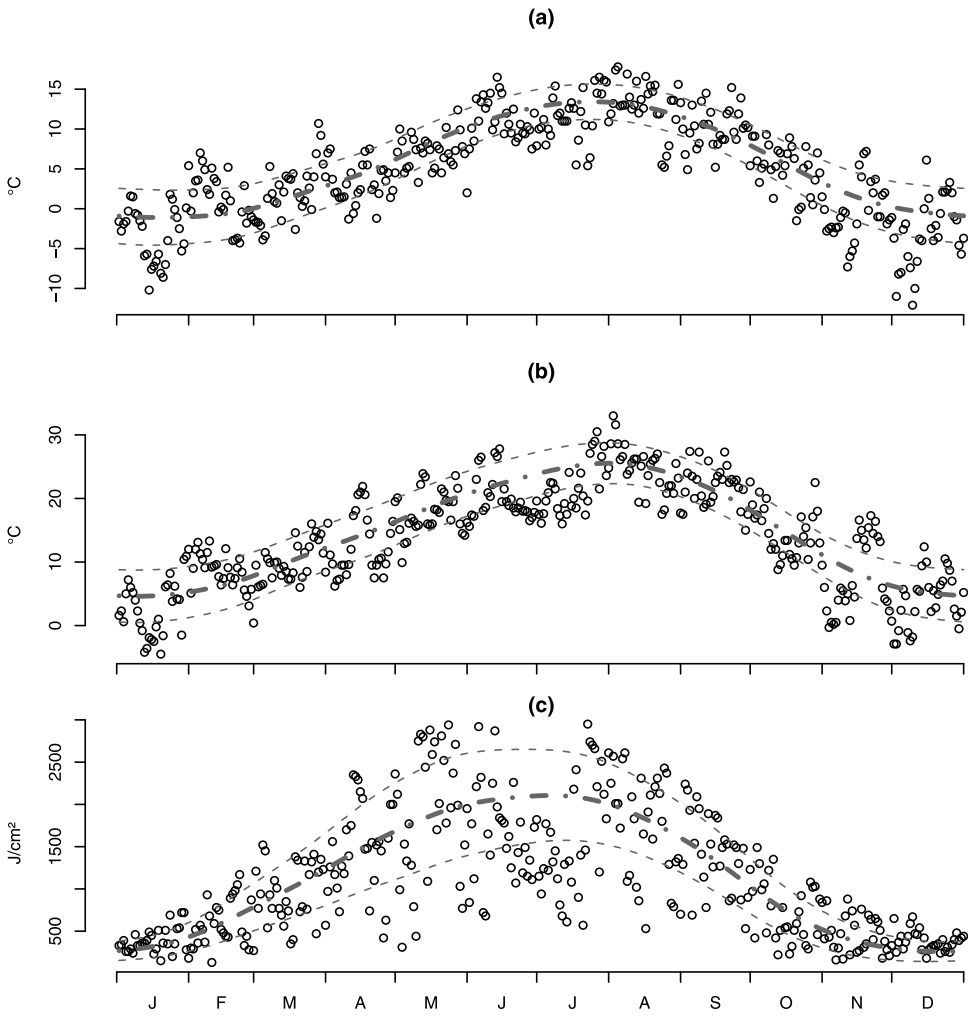

Figure 1. Measured (a) minimal temperature, (b) maximal temperature, and (c) radiation time series in Colmar from 1 January to 31 December 1980. Thick grey lines show the medians. Thin grey lines show median plus or minus absolute deviation.

only two states were defined, we have chosen to extend the number of daily states. This strategy allows us to better capture the complexity of weather changes. This concept of daily states has been successfully applied in downscaling large information to local scales. *Boé et al.* [2006] and *Boé and Terray* [2008] used for example weather types defined in terms of large-scale circulation similarities based upon the 500 hPa geopotential height resulting from the downscaled ARPEGE atmospheric model [*Gibelin and Déqué*, 2003]. *Vrac and Naveau* [2007] built precipitation-related patterns from a set of observed local precipitation records. In order to differentiate our approach from the large scale, we will use the term of weather state.

[5] Concerning the distribution of the variables of interest, daily precipitation amounts have been either fitted by a gamma or an half-normal distribution [e.g., *Semenov et al.*, 1998]. Gaussian distributions generally model temperatures and radiations. *Semenov et al.* [1998] emphasized that some variables such as radiations can strongly depart from Gaussianity (see, e.g., Figures 8g and 9g). To overpass this problem, *Young* [1994] implemented a mixture of distribution. In this paper we also propose a mixture of distribution but with two major differences. First, each cluster of the mixture corresponds to one weather state and, second, the distribution within each cluster (i.e., within each weather state) belongs to the family of multivariate closed skew-normal (CSN) distributions [*Genton*, 2004; *Pewsey and González-Farías*, 2007]. This class of distribution offers a general framework to fit both non-Gaussian and Gaussian variables. Conditionally to weather states, CSN distributions will be fitted to our five variables.

[6] The present paper describes in section 2 the general structure of our weather state approach with a multivariate closed skew-normal generator (named WACS-Gen) and briefly recalls the main properties of the closed skew-normal distribution. In section 3 an algorithm is proposed to estimate the parameters of the model and then in section 4 a real meteorological series measured in Colmar (France) is compared to series simulated by WACS-Gen with parameters estimated on a subset of this series.

2. WACS-Gen: A Weather Generator Based on Weather States and Skew-Normal Distributions

[7] We first explain how seasonality is accounted for. When a within-year trend is detected on a variable, the

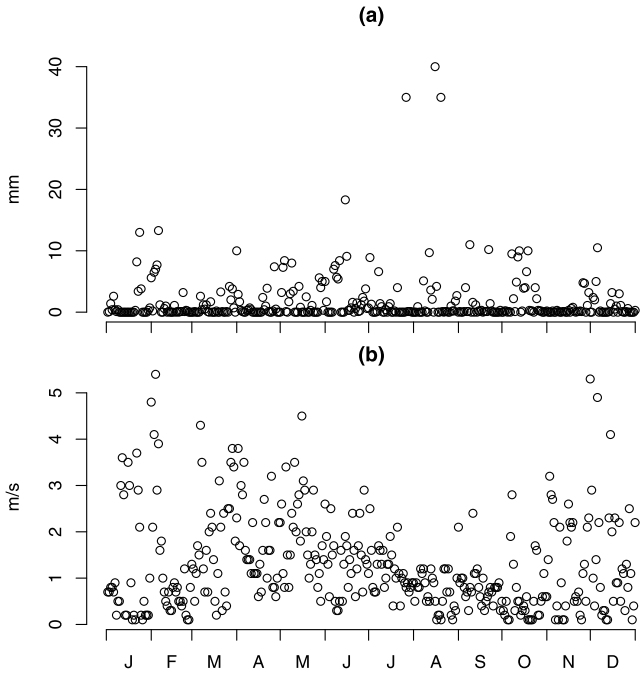

Figure 2. Measured (a) precipitation and (b) wind speed time series in Colmar from 1 January to 31 December 1980.

median and the average absolute deviation (defined as the mean of absolute difference between the variable and its median) are computed for each day and smoothed by a spline function [*Green and Silverman*, 1994]. This smoothed median is then subtracted to the studied variable and the difference is rescaled by the smoothed average absolute deviation. This normalization procedure is preferred to the classical mean and standard deviation based technique because rank statistics like median are more robust in presence of a departure from symmetry (see, e.g., the radiation). For the example of the Colmar series studied below, temperatures and radiations depend highly upon the day in the year (Figure 1). Figures 1a, 1b, and 1c correspond to temperature minima and maxima and radiations, respectively. No significant trend could be detected for precipitation intensity and wind speed (Figure 2). After transformation, and given a season and a weather state (see below), these temperature and radiation residuals are assumed to be stationary. They are the main object of this study. They will be studied independently within the four following seasons: December-January-February (DJF), March-April-May (MAM), June-July-August (JJA) and September-October-November (SON) [*Semenov et al.*, 1998].

[8] In the last decade, weather types have been frequently used to analyze various physical and stochastic climate models outputs at large scale [*Boé et al.*, 2006; *Boé and Terray*, 2008; *Vrac and Naveau*, 2007]. Weather types are classically defined for each season and their number varies from eight to ten types per season [*Bubnova et al.*, 1995].

[9] The proximity among meteorologically similar days can be determined by clustering methods. The two "wet" and "dry" weather states defined in earlier versions of weather generators can be viewed as a special case with only two weather states. At the other end of the spectrum, analog methods correspond to an extreme case in which there are as

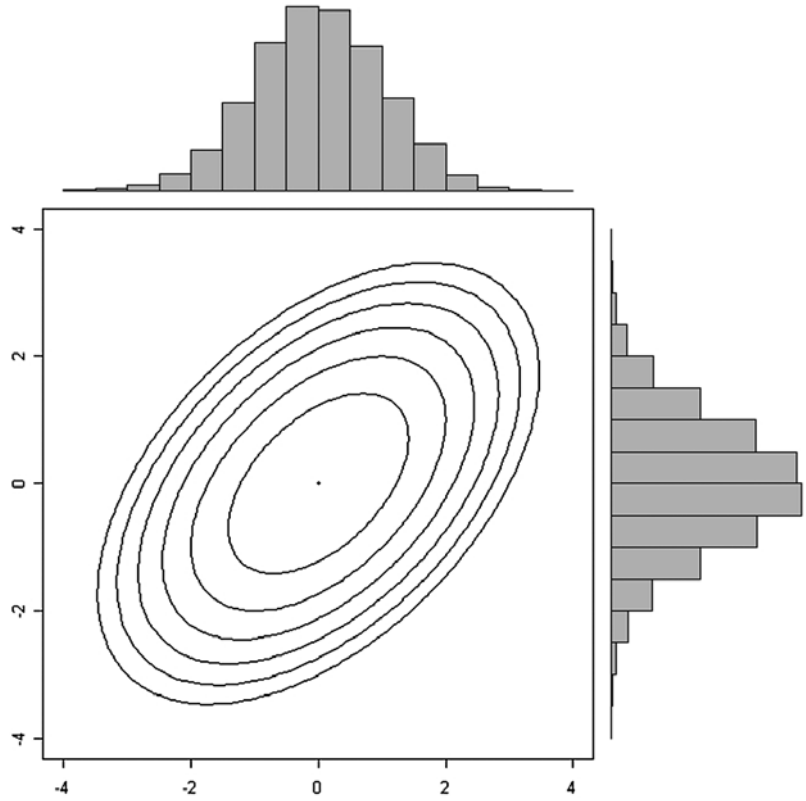

Figure 3. Density of a standard bivariate Gaussian vector with correlation parameter 0.5. Marginal histograms are plotted on corresponding sides.

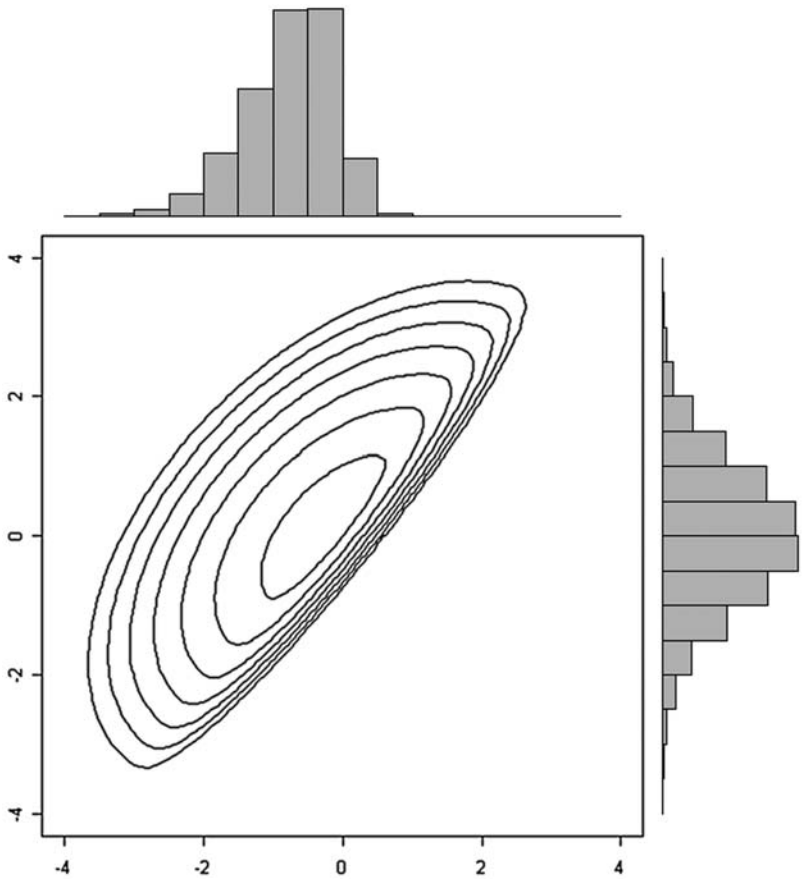

Figure 4. Density of a standard bivariate closed skew-normal vector CSN_2^* with skewness parameter (−0.94, 0.70) and correlation parameter $\rho = 0.5$ simulated from equation (3). Marginal histograms are plotted on corresponding sides.

many weather states as days. In this paper we strike a middle ground in which the number of states is not fixed a priori but is inferred from the data under study. More precisely, our multivariate clustering in N_W weather states is obtained by separately running a clustering algorithm on wet days and on dry days, respectively. The task is implemented with the Mclust function from the MClust package developed by *Fraley and Raftery* [2003, 2006] for R, an open source statistical software. The likelihood increases when the number of parameters, hence the number of clusters, increases. Choosing a large N_W also has to be viewed as a cost, otherwise the number of clusters will always be equal to the number of days. The number of weather states N_W is thus derived by optimizing the Bayesian information criterion (BIC) [*Schwarz*, 1978]. BIC penalizes the log likelihood of the model with a term equal to the number of free parameters times the logarithm of the data number. The transitions between successive weather states are modeled by a first-order homogeneous Markov chain with N_W states. This simply means that (1) the weather state of a day t, say $W(t)$, only depends on the weather state $W(t-1)$ at day $(t-1)$ and (2) the probability of such transitions is assumed to be independent of time and can be written as

$$p_{w,w'} = p(W(t+1) = w' | W(t) = w), \text{ with } (w, w') \in \{1, .., N_W\}^2. \quad (1)$$

[10] Skew-normal distributions are extensions of the normal distribution which admit skewness while retaining most of the interesting properties of the Gaussian distribution. An overview of theoretical and applied developments related to skewed distributions is provided in the book edited by *Genton* [2004]. More recently, a special issue in the journal *Communications in Statistics* edited by *Pewsey and González-Farías* [2007] was centered on the different aspects of skew distributions. Concerning its definition, a k-dimensional random vector $\mathbf{Y}$ is said to have a multivariate closed skew-normal distribution, denoted by $\mathrm{CSN}_{k,l}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\nu}, \boldsymbol{\Delta})$, if its density function is of the form

$$f_{k,l}(\mathbf{y}) = c_l \, \phi_k(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \Phi_l(\mathbf{D}^t(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Delta}),$$
$$\text{with } c_l^{-1} = \Phi_l\big(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Delta} + \mathbf{D}^t \boldsymbol{\Sigma} \mathbf{D}\big), \quad (2)$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\nu} \in \mathbb{R}^l$ are both location vectors, $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ and $\boldsymbol{\Delta} \in \mathbb{R}^{l \times l}$ are both covariance matrices, $\mathbf{D} \in \mathbb{R}^{k \times l}$, $\phi_k(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_k(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the probability distribution function (pdf) and cumulative distribution function (cdf), respectively, of the k-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{D}^t$ is the transpose of the matrix $\mathbf{D}$. In the particular case $\mathbf{D} = \mathbf{0}$ then $\mathbf{Y}$ is the usual k-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}$. The difference between Gaussian and skew-normal densities are illustrated Figures 3 and 4. Clearly, adding a skewness parameter through the skew-normal distribution provides flexibility for

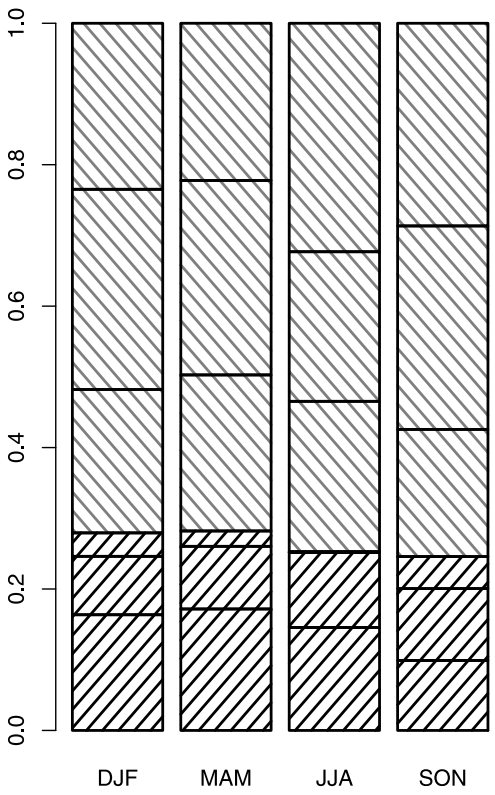

Figure 5. Number and frequencies of weather states for each season: wet weather states (black hatching) and dry weather states (grey hatching).

modeling skewness on the margins but also in the bivariate density. *González-Farías et al.* [2004] noticed that the CSN distributions defined by (2) are overparameterized and that without loss of generality $\boldsymbol{\nu}$ can be set equal to $\mathbf{0}$. In practice, the normalizing constant c_l^{-1} defined in (2) can be difficult to compute. To simplify its expression, we assume, without loosing the skew-normal flexibility, that $k = l$, $\mathbf{D} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{S}$ and $\boldsymbol{\Delta} = \mathbf{I}_k - \mathbf{S}^2$ where $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-1}$, $\mathbf{I}_k$ is the k-dimensional identity matrix and $\mathbf{S}$ is a diagonal matrix with elements in $[-1, 1]$. With this parameterization, equation (2) becomes

$$f_{k,k}(\mathbf{y}) = 2^{-k}\ \phi_k(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\ \Phi_k\left(\mathbf{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu});\mathbf{0},\mathbf{I}_k-\mathbf{S}^2\right),$$

which will be denoted $CSN_k^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{S})$ and is similar to the homotopic framework described by *Allard and Naveau* [2007]. On all climate series considered as part of the research project CLIMATOR (11 sites in France), and in particular on the Colmar series studied in this paper, we observed that mixtures of skew-normal densities could adequately be fitted on (R, V, T_n, T_x). This hypothesis is however not necessarily reasonable for precipitations. Daily rainfalls were properly fitted in most cases by a Gamma distribution, which does not belong to the class of CSN. The gamma distribution was chosen for its flexibility to model distributions of precipitation amount encountered the 11 locations in France. Other choices, like generalized Pareto distributions (GPDs) are of course possible and can easily be implemented. GPD was not chosen because it is theoretically constructed to fit very high values above a high fixed threshold. This paper does not focus on extreme rainfalls. The aim is rather to model the entire distribution. The variable P is thus transformed for all seasons into $\tilde{P} = \Phi^{-1}(G(P))$ where G represents the fit by a Gamma cdf and Φ^{-1} corresponds to the inverse of the standardized Normal cdf. $\tilde{P}$ is thus modeled as a Gaussian random variable; for a given season and a given weather state, $\tilde{P}$ will be considered as a CSN in order to account for possible asymmetries within clusters. This allows us to assume that, for a given season and a given weather state w, the vector ($\tilde{P}$, R, V, T_n, T_x) follows a $\text{CSN}_5^*(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w, \mathbf{S}_w)$ with

$$\boldsymbol{\mu}_w = \begin{bmatrix} \mu_{w,1} \\ \vdots \\ \mu_{w,5} \end{bmatrix}, \boldsymbol{\Sigma}_w = \begin{bmatrix} c_w^{1,1} & c_w^{1,2} & \cdots & c_w^{1,5} \\ c_w^{2,1} & c_w^{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ c_w^{5,1} & \cdots & \cdots & c_w^{5,5} \end{bmatrix},$$

$$\mathbf{S}_w = \begin{bmatrix} \delta_{w,1} & 0 & \cdots & 0 \\ 0 & \delta_{w,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \delta_{w,5} \end{bmatrix},$$

where the subscript w indicates that this vector is defined conditionally to the weather state w and the subscript for the season is dropped to simplify notations. Under this assumption, correlations among our five variables are modeled within a unique and common multidimensional distributional framework.

[11] Concerning the representation of temporal persistence [*Rajagopalan and Lall*, 1999], let us consider two consecutive days t and $t + 1$ and their associated weather states, say w and w'. The main question here is how the five dimensional structure of the CSN* vector $\mathbf{X}_w$ at time t can be changed to become a five dimensional CSN* vector $\mathbf{X}_{w'}$ at time $t + 1$. This is achieved by first rotating the vector $\mathbf{X}_w$ in order to make its margins independent. Lemma 1 (see Appendix) provides the elements to perform this first step. The transformed vector

$$\tilde{\mathbf{X}}_w = \boldsymbol{\Sigma}_w^{-1/2}(\mathbf{X}_w - \boldsymbol{\mu}_w) \sim \text{CSN}_5^*(\mathbf{0}, \mathbf{I}_5, \mathbf{S}_w), \tag{3}$$

whose margins are independent and distributed as $\text{CSN}_1^*(0,\ 1,\ \delta_{w,i})$. For the particular case of dry days the dimension of $\tilde{\mathbf{X}}_w$ is reduced to four since precipitation is always equal to zero. Equation (3) allows to model the temporal evolution from $\tilde{\mathbf{X}}_w$ to $\tilde{\mathbf{X}}_{w'}$ only throughout their marginals. As a second step, the pairwise structure between the i^{th} components of $\tilde{\mathbf{X}}_w$ and $\tilde{\mathbf{X}}_{w'}$ is assumed to be a bivariate $\text{CSN}_{2,2}(\mathbf{0}, \boldsymbol{\Sigma}_{w,w'}^{(i)}, \mathbf{D}_{w,w'}^{(i)}, \mathbf{0}, \boldsymbol{\Delta}_{w,w'}^{(i)})$ where

$$\boldsymbol{\Sigma}_{w,w'}^{(i)} = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}, \mathbf{D}_{w,w'}^{(i)} = \frac{1}{1-\rho_i^2}\begin{bmatrix} \delta_{w,i} & -\rho_i\delta_{w,i} \\ -\rho_i\delta_{w',i} & \delta_{w',i} \end{bmatrix},$$

and

$$\boldsymbol{\Delta}_{w,w'}^{(i)} = \mathbf{I}_2 - \frac{1}{1-\rho_i^2}\begin{bmatrix} \delta_{w,i}^2 & -\rho_i\delta_{w,i}\delta_{w',i} \\ -\rho_i\delta_{w,i}\delta_{w',i} & \delta_{w',i}^2 \end{bmatrix}.$$

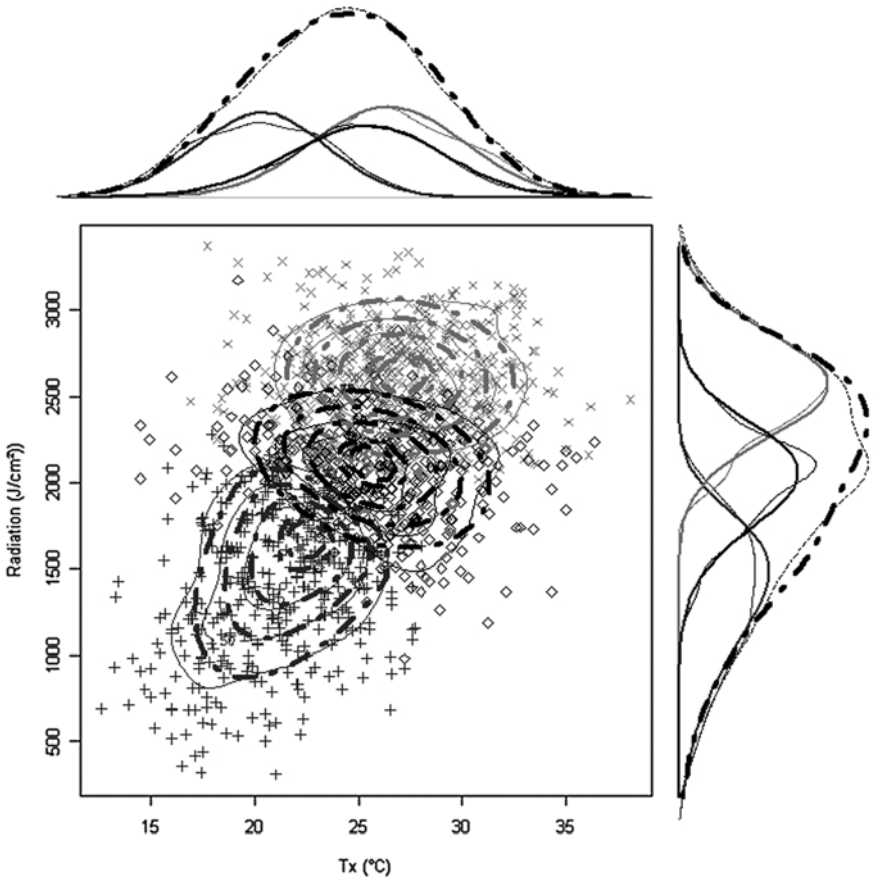

Figure 6. Bivariate plot (T_x, R) with estimated densities for the three dry weather states during JJA season. The plot inside the box shows the following: thin solid lines, kernel estimates of the bivariate densities; thick dashed lines, estimated CSN* densities. Plots along the top and right of the box show marginal densities: thin solid lines, kernel densities; thick solid lines, CSN* densities; thin dashed lines, mixture of kernel densities; thick dashed lines, mixture of CSN* densities.

Lemma 2 ensures that the margins of this bivariate vector are indeed $CSN_1^*(0, 1, \delta_{w,i})$.

3. Parameter Estimation and Weather Generator Scheme

[12] The weather state transition probability defined by (1) is simply estimated by

$$\hat{p}_{w,w'} = \frac{\sharp\left[W(t) = w, W(t+1) = w'\right]}{\sharp[W(t) = w]}, \tag{4}$$

where $\sharp$ denotes the cardinal.

[13] Concerning the inference of the marginal CSN* parameters, *Azzalini and Capitanio* [1999] studied the classical maximum likelihood estimation (mle) approach and *Flecher et al.* [2009] proposed a weighted moment method. Conditionally to the weather state w, a mle approach ignoring temporal dependence is implemented to estimate the parameters of $\text{CSN}_5^*(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w, \boldsymbol{S}_w)$. The estimates are only slightly changed if the temporal dependence is taken into account in the estimation procedure. It has a larger impact on the covariance matrix of the estimators of the parameters, but since this matrix is not used in the weather generator, this point is simply ignored for the sake of ease of use.

[14] Then the correlation coefficient ρ_i in $\boldsymbol{\Sigma}_{w,w'}^{(i)} = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}$ is estimated via a weighted moment approach [*Flecher et al.*, 2009], i.e., by solving the following equation in ρ_i

$$\mathbb{E}\left(\Phi_2\left(\left(\tilde{X}_w^{(i)}(t), \tilde{X}_{w'}^{(i)}(t+1)\right); \mathbf{0}, \mathbf{I}_2\right)\right) = 4\Phi_4\left(\mathbf{0}; \mathbf{0}, \mathbf{M}_{w,w'}^{(i)}\right),$$

where $\tilde{X}_w^{(i)}(t)$ corresponds to the i^{th} component of the vector $\tilde{\mathbf{X}}_w$ at time t,

$$\mathbf{M}_{w,w'}^{(i)} = \frac{1}{1-\rho_i^2}\begin{bmatrix} 2(1-\rho_i^2) & \rho_i(1-\rho_i^2) & \delta_{w,i} - \rho_i^2\delta_{w',i} & \rho_i\left(\delta_{w',i} - \delta_{w,i}\right) \\ \rho_i(1-\rho_i^2) & 2(1-\rho_i^2) & \rho_i\left(\delta_{w,i} - \delta_{w',i}\right) & \delta_{w',i} - \rho_i^2\delta_{w,i} \\ \delta_{w,i} - \rho_i^2\delta_{w',i} & \rho_i\left(\delta_{w,i} - \delta_{w',i}\right) & (1-\rho_i^2) & 0 \\ \rho_i\left(\delta_{w',i} - \delta_{w,i}\right) & \delta_{w',i} - \rho_i^2\delta_{w,i} & 0 & (1-\rho_i^2) \end{bmatrix},$$

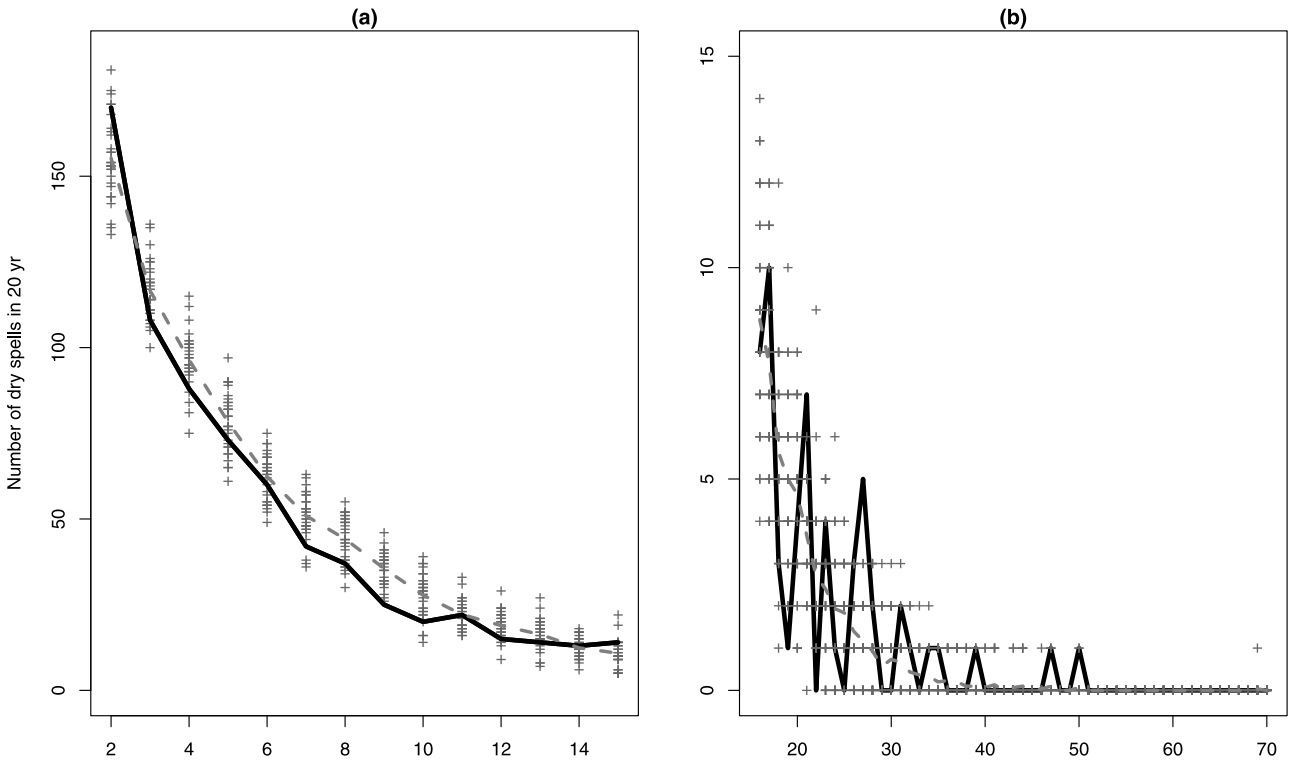

Figure 7. Dry spell lengths (a) from 2 to 15 days and (b) from 16 to 70 days. Solid line, measurements; crosses, number of dry spells in 30 simulations; black line, mean of these simulations. Note the scale change in the y axis between Figures 7a and 7b.

and $\mathbb{E}(\Phi_2((\tilde{X}_w^{(i)}(t), \tilde{X}_{w'}^{(i)}(t+1)); \mathbf{0}, \mathbf{I}_2))$ is replaced by its empirical estimator.

[15] After estimating the parameters, the following algorithm simulates realizations of the five dimensional vector of interest.

1. A season is chosen and one $\tilde{\mathbf{X}}(0)$ is randomly chosen (e.g., with an analog method).

2. The transition probabilities estimated with (4) are used to generate a Markov chain sequence of weather states.

3. Given $\tilde{\mathbf{X}}(t) = \mathbf{x}_t$ and two consecutive weather states, w and w', a realization of the vector $\tilde{\mathbf{X}}(t+1)$ defined by (3) is drawn according to (see Lemma 2)

$$\left[\tilde{X}_{w'}^{(i)}(t+1) | \tilde{X}_w^{(i)}(t) = x_t^{(i)}\right] \sim CSN_{1,2}\left(\rho_i\, x_t^{(i)}, 1-\rho_i^2, \boldsymbol{D}_c^{(i)}, \boldsymbol{\nu}_c^{(i)}, \boldsymbol{\Delta}_c^{(i)}\right),$$

where

$$\mathbf{D}_c^{(i)} = \frac{1}{1-\rho_i^2}\begin{bmatrix}\rho_i\delta_{w,i}\\ \delta_{w',i}\end{bmatrix}, \boldsymbol{\nu}_c^{(i)} = -\frac{1}{1-\rho_i^2}\begin{bmatrix}\delta_{w',i}\\ -\rho_i\delta_{w,i}\end{bmatrix}x^{(i)}(t),$$

$$\boldsymbol{\Delta}_c^{(i)} = \mathbf{I}_2 - \frac{1}{1-\rho_i^2}\begin{bmatrix}\delta_{w,i}^2 & -\rho_i\delta_{w,i}\delta_{w',i}\\ -\rho_i\delta_{w,i}\delta_{w',i} & \delta_{w',i}^2\end{bmatrix}.$$

4. To invert relation (3), $\tilde{\mathbf{X}}$ is multiplied by $\boldsymbol{\Sigma}_{w'}^{1/2}$ and $\boldsymbol{\mu}_{w'}$ is added.

5. To add back trends and seasonal effects, we inverse the steps of the standardization based on the median and the absolute deviation described in the first paragraph of section 2.

4. Weather Data in Colmar, France

[16] Colmar, a city in the north east part of France, is located at 48°05′N latitude, 7°21′E longitude and has an altitude of 175 m. A 20 year series is available from 1973 to 1992 for the five daily variables under study. Annual precipitation amounts are about 530 mm and the frequency of rainy days is about 1/4. The climate is characterized by warm summers from June to September and cold winters (the annual temperature cycle is well marked with a 25°C mean in July and 2°C in January). Both oceanic and continental climate trades can affect this site. This produces an important variability on the daily meteorology.

[17] For each season a maximum of eight different weather states is allowed. The BIC criterion provides a number of regime clusters that is equal to five for the JJA season and six for the other seasons. The repartition per season appears to be fairly homogeneous throughout a year (see Figure 5).

[18] The estimation of the parameters of the CSN* distributions is illustrated during dry days of the JJA season on the pair of variables (T_x, R) (Figure 6). The bivariate distributions in each cluster (i.e., weather state) are well modeled by their corresponding CSN* densities. The two marginal densities resulting from the mixture are fairly well

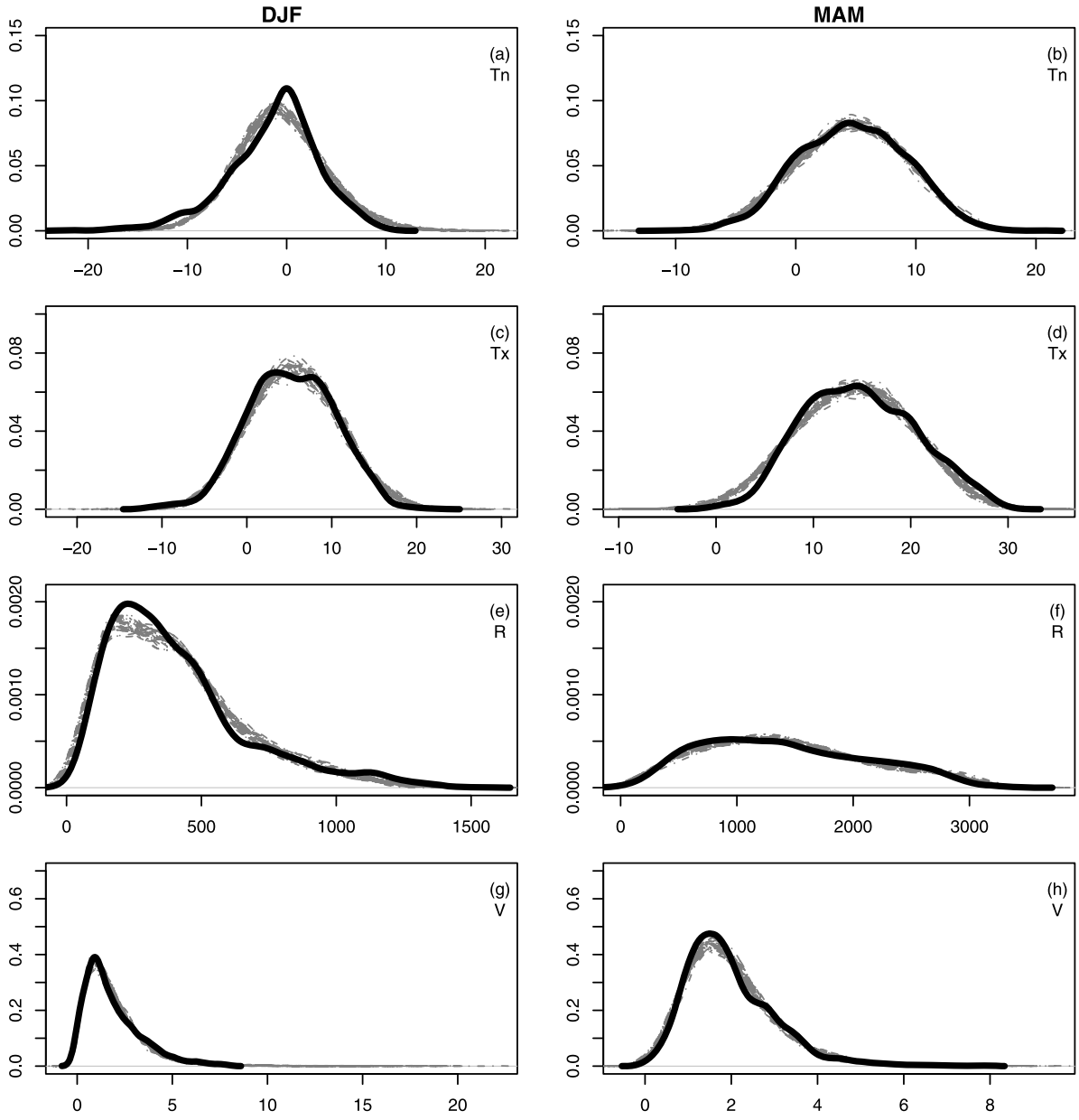

Figure 8. Densities of (a and b) minimal temperatures, (c and d) maximal temperatures, (e and f) radiations, and (g and h) wind speed for (left) DJF and (right) MAM. Black line, measurements; grey dashed lines, simulations.

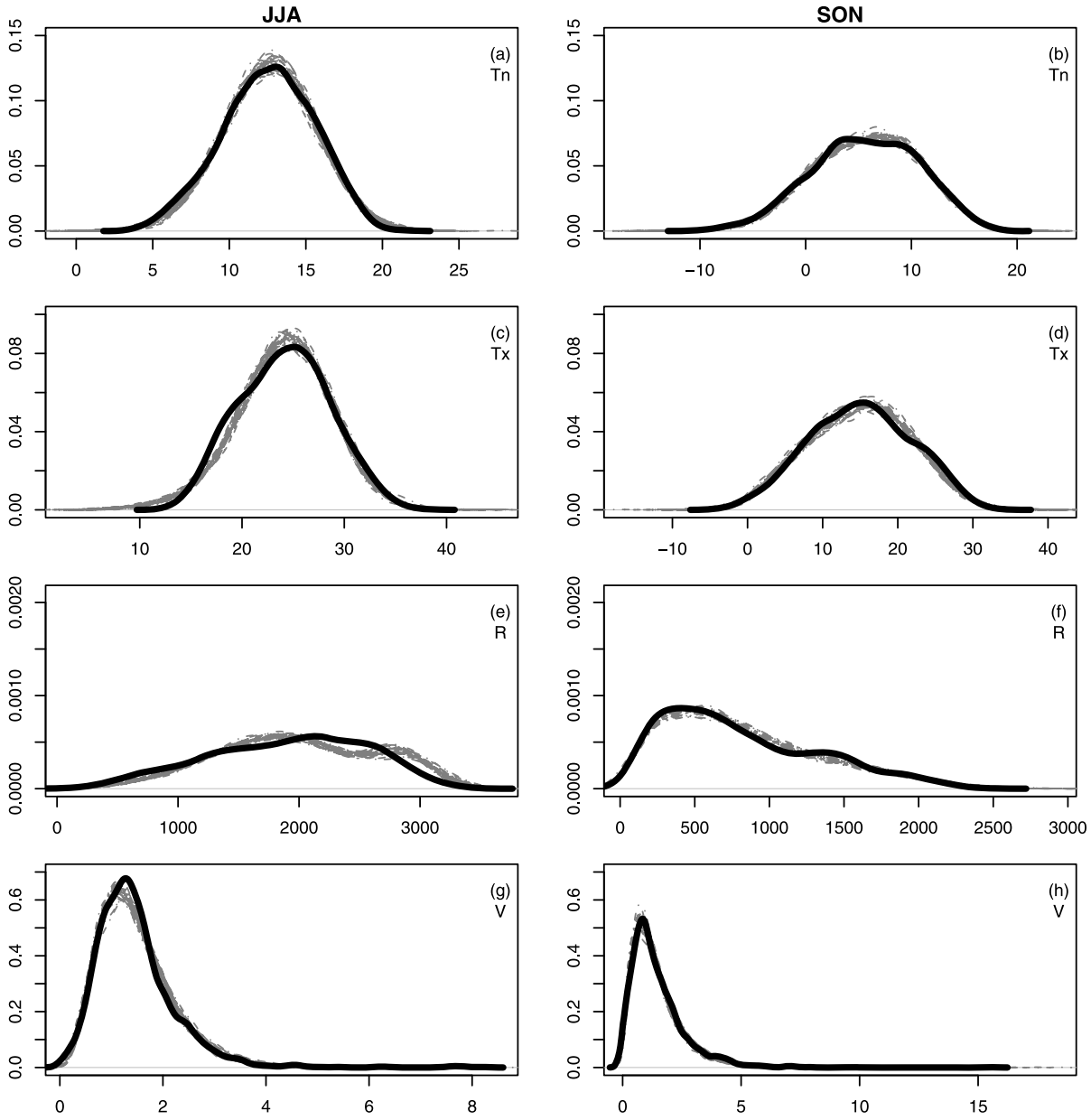

Figure 9. Same as Figure 8 but for (left) JJA and (right) SON.

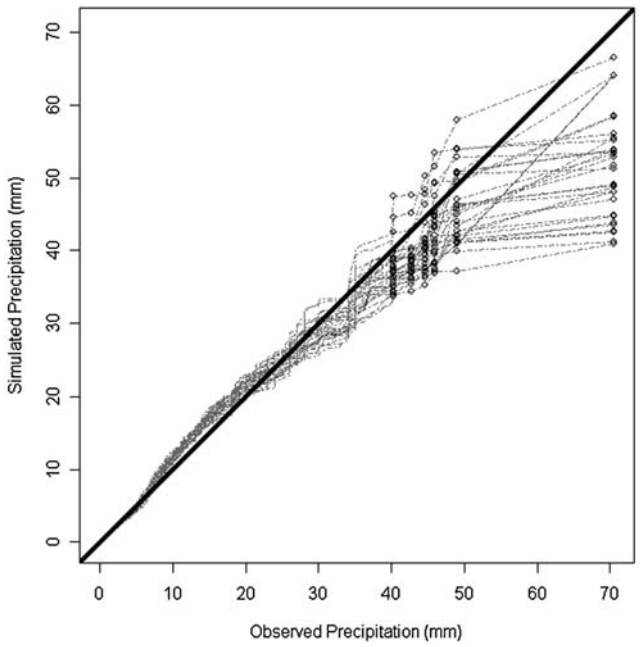

Figure 10. QQ plots of precipitations (all seasons pooled together). Black line, measurements; grey dashed lines, simulations. The five largest values are represented by points for each simulation.

reproduced. Although presented here with a pair of variables for the ease of representation, similar results are obtained for the full vector of 5 variables, and for all seasons.

[19] After estimating our CSN* model parameters, thirty runs of 20 years are simulated. As a first step to compare our simulated time series to Colmar measurements, Figure 7 shows the dry spell length distribution computed from the Colmar observations (solid lines), the average on thirty simulations (dashed line) and the values of the thirty simulations (crosses), for all seasons pooled together. Similar results were obtained when considering each season separately. The left plot focuses on dry spell lengths shorter than 15 days and the right plot zooms on longer dry spells. Both graphs indicate that this variable is fairly well reproduced by our Markov chain. Concerning the five variables of interest, their densities are shown in Figures 8 and 9. For each season, Figures 8a, 8b, 9a, and 9b display minimal temperatures, Figures 8c, 8d, 9c, and 9d display maximal temperatures, Figures 8e, 8f, 9e, and 9f display radiation, and Figures 8g, 8h, 9g, and 9h display wind speed densities. The black solid lines correspond to the measurements and the grey dashed lines represent the thirty realizations obtained with our weather generator. In Figure 8, the left and right plots correspond to DJF and MAM, respectively. Figure 9 shows the same information for JJA and SON, respectively. These plots exemplify the advantage of the CSN distribution, which is able to capture the skewness exhibited in radiations and wind speed distributions. Note that the probability density function of radiation is not very well fitted during JJA (Figure 9e). The fit can be significantly improved by defining two additional weather states (figure not shown). The difference between the BIC criterion for 6 and 8 clusters is positive, but small. In this paper we focus on the general presentation of the generator; the problem of finding other criteria than BIC for selecting the number of clusters will be the subject of further work. We therefore maintain the current fit with 6 clusters.

[20] Concerning precipitation intensity, its distribution is represented by a quantile-quantile plot (QQ plot) in Figure 10. This QQ plot is defined as the sorted simulated rainfalls versus the sorted historical record. A good fit corresponds to the first diagonal. This graph indicates that precipitation amounts are well reproduced by the generator. On this site, the highest observed precipitation is 70 mm, while the highest simulated precipitations are in the range 40–68 mm. Note however that the mean precipitation is well reproduced and that on other sites the opposite situation (higher simulated highest precipitations than measured ones) can be observed (results not reported here).

[21] In Figure 11, star plots represent correlations between our five variables for each season. Large positive correlations are near the star plot border whereas large negative one are near the center. Such graphs provide a graphical way to view a correlation matrix. The correlations between precipitation and other variables are only computed for wet days. Figure 11 shows that our model is capable to reproduce the observed cross correlations. For each variable, the correlations between two consecutive days is represented with the same star plot graph in Figure 12. The persistence between two consecutive days is well reproduced except for the winter season (DJF), which provides the most severe discrepancy between observations and simulations, mainly for temperature variables.

[22] Wind speed and precipitation are variables known to be difficult to model. To study the improvement brought by the introduction of multiple weather states, thirty additional simulations of 20 year length are also obtained by forcing our generator to only have the two classical wet and dry weather states. In Figure 13, wind speed boxplots and densities are obtained with a classical two weather states (wet and dry) and for six weather states, as defined in Figure 5, respectively. Figure 14 are quantile-quantile plots of the amount of precipitation for the Colmar series and for each of the 30 simulations in the two cases: multiple weather states case (Figure 14a) and classical wet/dry case (Figure 14b). For both variables, introducing multiple weather states improves significantly the fitting of the distribution. Current stochastic weather generators are for example known to underestimate the probability of small precipitations as explained by *Semenov et al.* [1998]. In the case of Colmar, the precision of the data is 0.1 mm. The overall frequency of precipitation less than or equal to 2 mm is 6.1% on the data. On simulations with two wet/dry weather states it ranges from 1.6% to 2.3%. On simulations with a BIC optimized number of weather states, it ranges from 5.6% to 7.1%.

[23] Figure 15 displays the wind speed autocorrelation boxplot for each season computed with a two or six weather states. The horizontal black lines correspond to the observed wind speed autocorrelation per season. Despite the incapacity for both generators to reproduce the wind speed autocorrelation in the MAM season, Figures 13 and 15 clearly show the improvement brought by the introduction of additional weather states for wind speed distribution

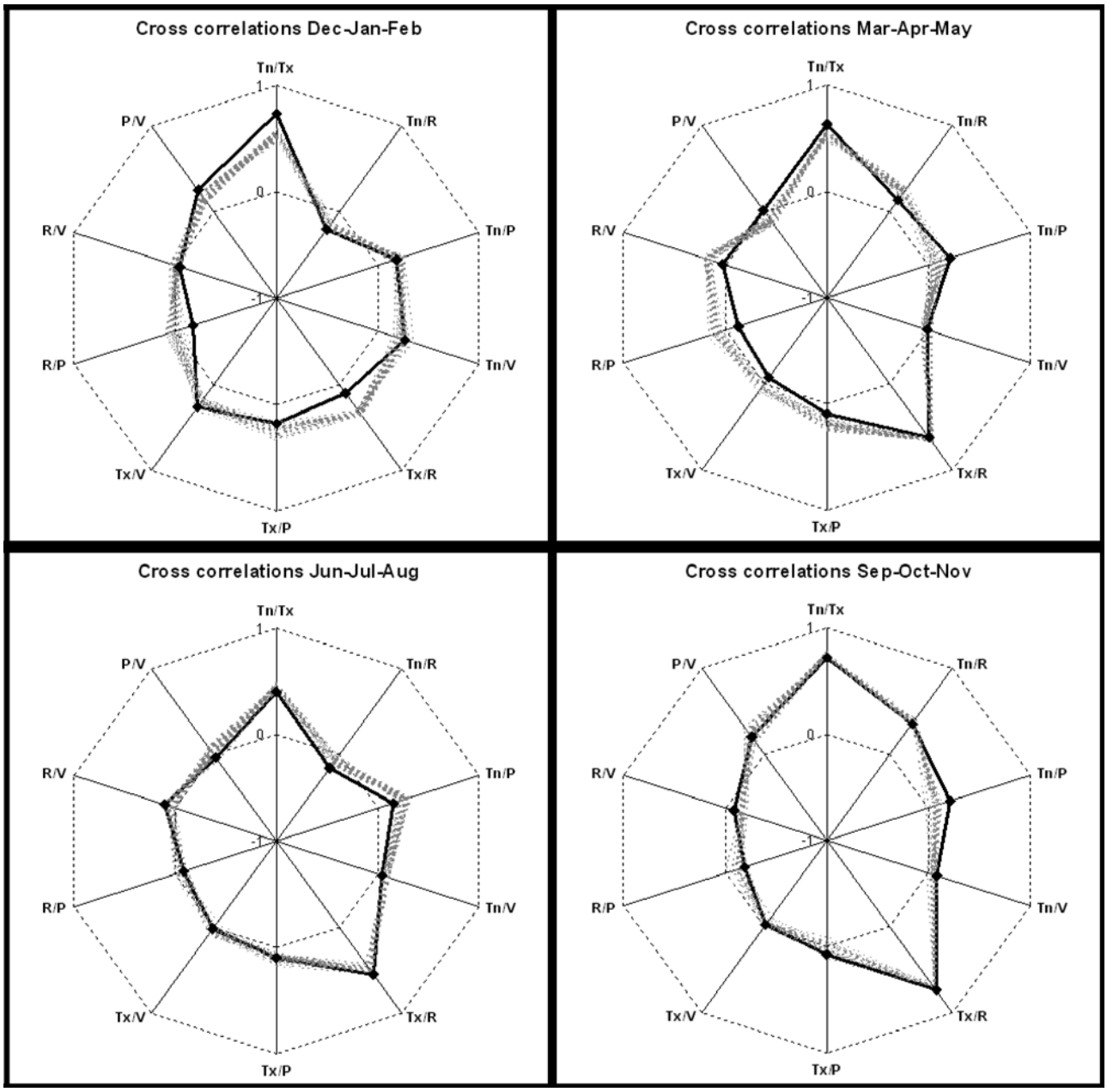

Figure 11. Correlations between variables for each season. Large positive correlations are near the border of the star plot, whereas large negative ones are near the center. Black line, measurements; grey lines, simulations.

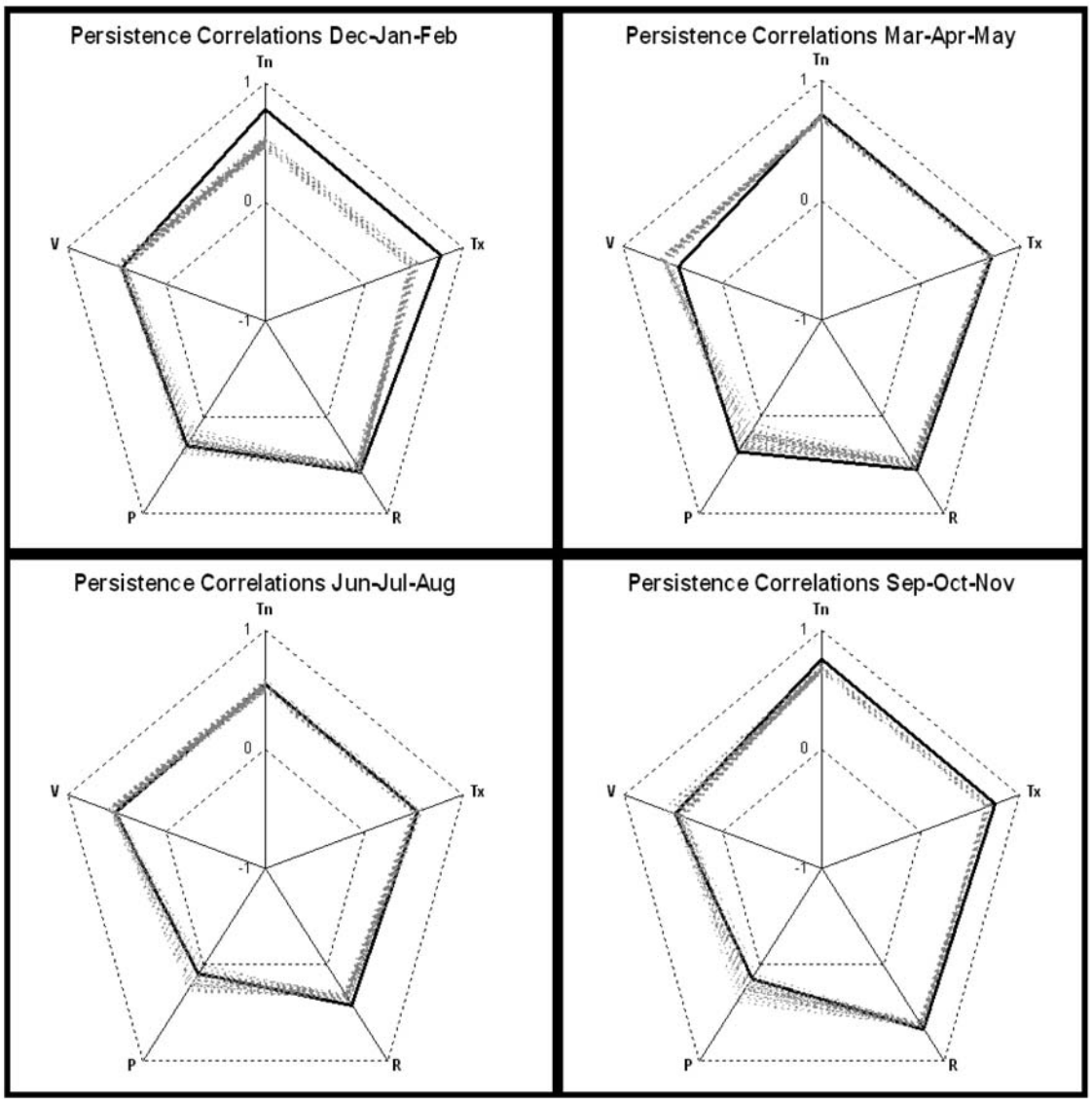

Figure 12. Correlations between two consecutive days for the same variable. Black line, measurements; grey lines, simulations.

and autocorrelation modeling. Concerning precipitation, Figure 16 indicates that generators with two or more weather states do provide decent but not excellent results.

5. Conclusion

[24] We have presented WACS-Gen, a new weather generator, which presents several improvements compared to previous ones: (1) the number of weather states is no longer limited to the dry/wet states, but is fitted to the variability of the observed data using a model-based clustering algorithm on detrended data and (2) conditionally on the season and the weather state, the multivariate data are modeled using CSN distributions, thus allowing for residual skewness; correlation between variables and along time is also modeled, including between successive days with different weather states.

[25] Allowing for multiple weather states is a major improvement, but it raises the question of defining a good criterion for selecting the correct number of clusters. Here, we have proposed to use BIC, a widely used criterion in model based clustering [*Fraley and Raftery*, 2003]. It provides most of the time a very good fit of the probability densities. In one situation (radiation during JJA), the fit could be improved by increasing the number of clusters, as compared to the BIC criterion. Finding better criteria than BIC will be the subject of future work.

[26] This generator has been tested on different weather series measured in contrasted climatic regimes across France. Although we only have reported results on the Colmar series due to space constraints, our results showed consistently that WACS-Gen substantially improves the reproduction of histograms, cross and temporal correlations as compared to generators with only dry/wet weather states. Histograms are also very well reproduced thanks to the mixture of CSN distributions inherited from the multiple weather states. Of particular interest is the ability of our generator to model the correlation between the amount of

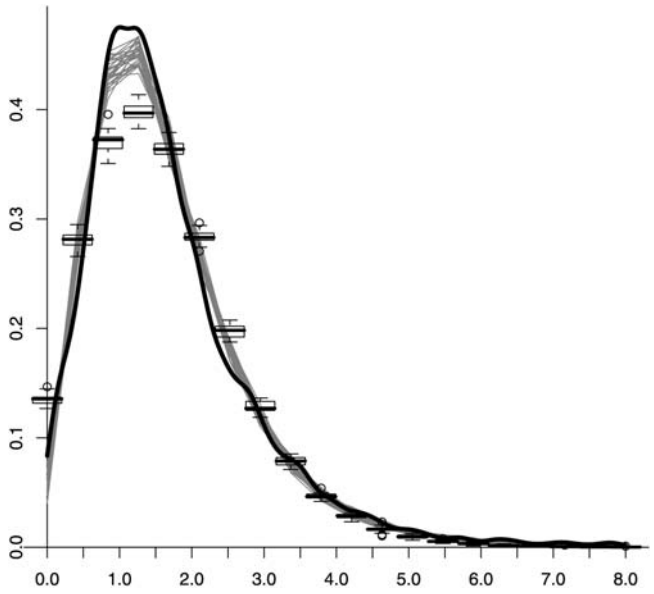

Figure 13. Wind speed. Solid line, measured wind speed density; box plots, simulated densities with dry/wet weather states; grey lines, simulated densities with multiple weather states.

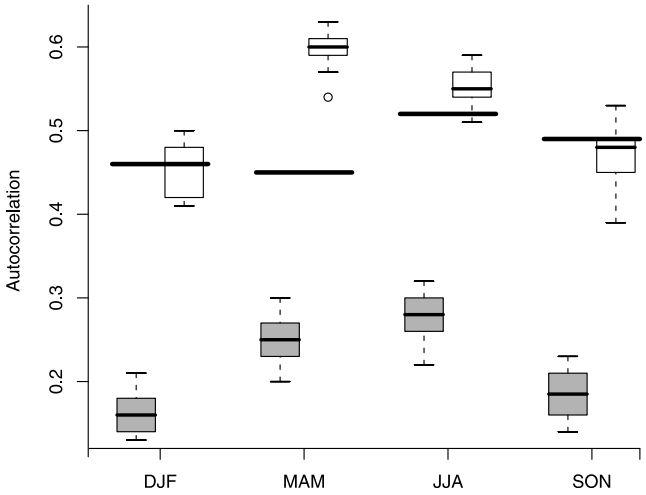

Figure 15. Wind speed autocorrelation box plots for each season. Grey boxes, dry/wet weather states; open boxes, multiple weather states.

precipitation and the other variables instead of only conditioning these variables to the precipitation event.

[27] We still have some difficulties in reproducing some statistics, in particular correlations with wind speeds and extreme events. Wind speed is known to be a difficult variable, with strongly nonlinear correlation to other variables. Although being able to account for asymmetrical distributions, CSN are not targeted at modeling extreme data. Future improvements on weather generators should be focused on integrating extreme values theory to better reproduce extreme events, and on modeling nonlinear relationship between variables. The impact of these improve-

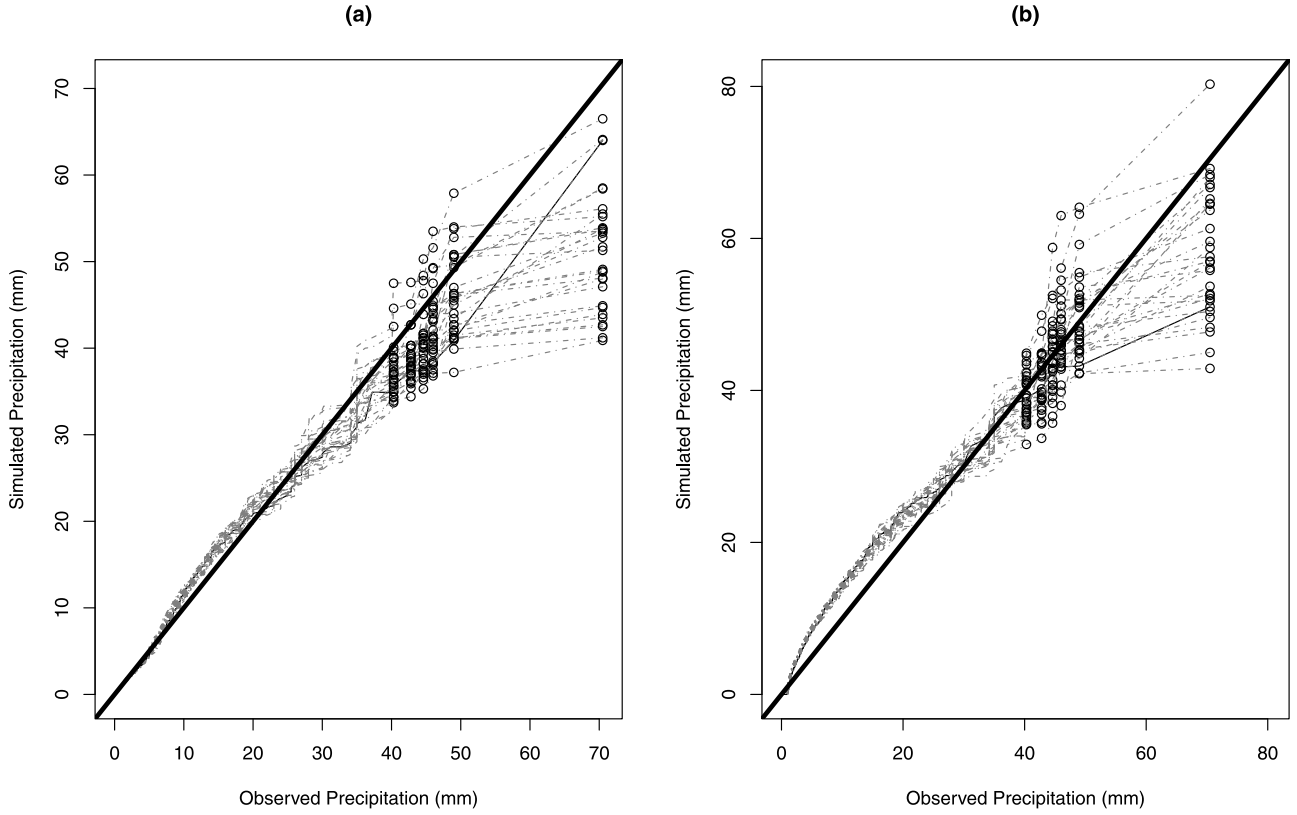

Figure 14. QQ plots of precipitations (overall) (a) with multiple weather states and (b) with dry/wet weather states. Black line, measurements; grey dashed lines, simulations.

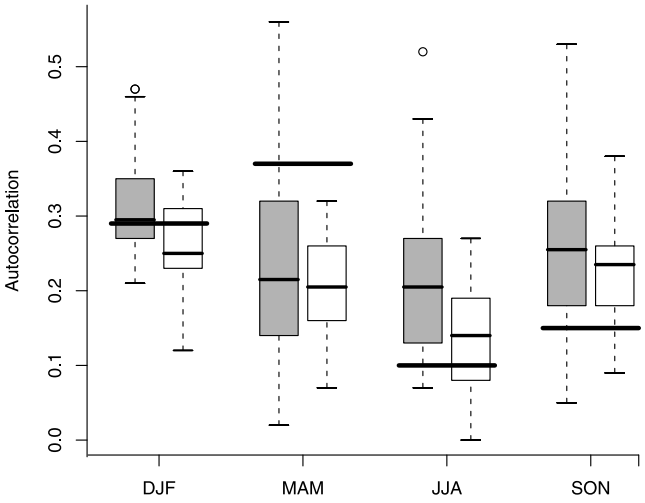

Figure 16. Precipitation autocorrelation box plots. Grey boxes, dry/wet weather states; open boxes, multiple weather states.

ments on crop models still needs to be assessed, which will be our very next task.

[28] In the framework of climate change studies, we not only need to consider that the parameters will change with time, but we also need to consider the change of support (i.e., downscaling) problem. General climate models provide output variables varying with time. at very large scale, while crop models need weather variables at very small scale; we therefore need to provide models to estimate small-scale parameters from large-scale data. This should be treated in a forthcoming paper.

Appendix A

[29] The following lemmas can be derived from the CSN properties described by *González-Farías et al.* [2004] and *Genton* [2004]. Proofs are technical, but otherwise straightforward. They are available upon request.

Lemma 1. Let $\mathbf{X}$ be a $\mathrm{CSN}_{n,n}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{S}\boldsymbol{\Sigma}^{-1/2}, \mathbf{0}, \mathbf{I}_n - \mathbf{S}^2)$, where $\boldsymbol{\Sigma}^{-1/2}$ is defined as a positive symmetric matrix such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$ and $\mathbf{S} = diag(\boldsymbol{\delta})$, with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^t$. Then $\widetilde{\mathbf{Y}} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ follows a $\mathrm{CSN}_{n,n}(\mathbf{0}, \mathbf{I}_n, \mathbf{S}, \mathbf{0}, \mathbf{I}_n - \mathbf{S}^2)$.

Lemma 1 indicates that the skewness parameter remains unchanged after standardization.

Lemma 2. Let $\mathbf{X} = (X_1, X_2)$ be a $\mathrm{CSN}_{2,2}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{D}, \mathbf{0}, \boldsymbol{\Delta})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \mathbf{D} = \frac{1}{1-\rho^2}\begin{bmatrix} \delta_1 & -\rho\delta_1 \\ -\rho\delta_2 & \delta_2 \end{bmatrix},$$
$$\boldsymbol{\Delta} = \mathbf{I}_2 - \frac{1}{1-\rho^2}\begin{bmatrix} \delta_1^2 & \rho\delta_1\delta_2 \\ \rho\delta_1\delta_2 & \delta_2^2 \end{bmatrix},$$

then

1. both margins are distributed as a $\mathrm{CSN}_{1,1}(0, 1, \delta_i, 0, 1 - \delta_i^2)$ for $i \in \{1, 2\}$.

2. the conditional distribution of X_2 given $X_1 = x_1$ is a $\mathrm{CSN}_{1,2}(\mu_c, \sigma_c^2, \mathbf{D}_c, \boldsymbol{\nu}_c, \boldsymbol{\Delta}_c)$ with $\mu_c = \rho x_1$, $\sigma_c^2 = 1 - \rho^2$,

$$\mathbf{D}_c = \frac{1}{1-\rho^2}\begin{bmatrix} -\rho\delta_1 \\ \delta_2 \end{bmatrix}, (\nu)_c = \frac{1}{1-\rho^2}\begin{bmatrix} \delta_1 \\ -\rho\delta_2 \end{bmatrix} x_1,$$
$$\boldsymbol{\Delta}_c = \mathbf{I}_2 - \begin{bmatrix} \delta_1^2 & -\rho\delta_1\delta_2 \\ -\rho\delta_1\delta_2 & \delta_2^2 \end{bmatrix}.$$

[30] **Acknowledgments.** This work was supported by the ANR-CLIMATOR project and the ANR-AssimilEx project. Part of Philippe Naveau's work has been supported by the EU-FP7ACQWA project (http://www.acqwa.ch/) under contract 212250, by the PEPER-GIS project (http://www.gisclimat.fr/projet/peper), by the ANR-MOPERA project, and by the NICE RTN project. The authors wish to express their gratitude to the editor and referees for their very helpful comments and suggestions. They would also like to credit the contributors of the R project.

References

Allard, D., and P. Naveau (2007) A new spatial skew-normal random field model, *Commun. Stat.*, *36*, 1821–1834.

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

Azzalini, A., and A. Capitanio (1999), Statistical applications of the multivariate skew-normal distribution, *J. R. Stat. Soc., Ser. B*, *61*, 579–602.

Boé, J., and L. Terray (2008), A weather type approach to analysing winter precipitation in France: Twentieth century trends and influence of anthropogenic forcing, *J. Clim.*, *21*, 3118–3133.

Boé, J., L. Terray, F. Habets, and E. Martin (2006), A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling, *J. Geophys. Res.*, *111*, D23106, doi:10.1029/2005JD006889.

Brisson, N., et al. (2003) An overview of the crop model STICS, *Eur. J. Agron.*, *18*, 309–332.

Brisson, N., M. Launay, B. Mary, and N. Beaudoin (Eds.) (2009), *Conceptual Basis, Formalisations and Parameterization of the STICS Crop Model*, 304 pp., Ed. Quae, Versailles, France.

Bubnova, R., G. Hello, P. Bernard, and J. F. Geleyn (1995), Integration of the fully elastic equations cast in the hydrostatic pressure terrain following coordinate in the framework of the ARPEGE/Aladin NWP system, *Mon. Weather Rev.*, *123*, 515–535.

Flecher, C., P. Naveau, and D. Allard (2009), Estimating the closed skew-normal distributions parameters using weighted moments, *Stat. Probab. Lett.*, *79*, 1977–1984.

Fraley, C., and A. E. Raftery (2003), Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLust, *J. Classif.*, *20*, 263–286.

Fraley, C., and A. E. Raftery (2006), MCLust version 3 for R: Normal mixture modeling and model-based clustering, *Tech. Rep. 504*, Dep. of Stat., Univ. of Wash., Seattle.

Furrer, E. M., and R. W. Katz (2007), Generalized linear modeling approach to stochastic weather generators, *Clim. Res.*, *34*, 129–144.

Genton, M. G. (Eds.) (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman and Hall, Boca Raton, Fla.

Gibelin, A.-L., and M. Déqué (2003) Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution model, *Clim. Dyn.*, *20*, 327–339.

Green, P. J., and B. W. Silverman (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, Boca Raton, Fla.

González-Farías, G., J. Domínguez-Molina, and A. Gupta (2004), Additive properties of skew-normal random vectors, *J. Stat. Plann. Inference*, *126*, 521–534.

Katz, R. W. (1996), The use of stochastic models to generate climate scenarios, *Clim. Change*, *32*, 237–255.

Pewsey, A., and G. González-Farías (2007), Skew-elliptical distributions and their application—Preface, *Commun. Stat. Theory Methods*, *36*, 1657–1659.

Qian, B., H. Hayhoe, and S. Gameda (2005), Evaluation of the stochastic weather generators LARS-WG and AAFC-WG for climate change impact studies, *Clim. Res.*, *29*, 3–21.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*, 3089–3101.

Rajagopalan, B., U. Lall, D. G. Tarboton, and D. S. Bowles (1997), Multivariate non parametric resampling sheme for generation of daily weather variables, *Stochastic Hydrol. Hydraul.*, *11*(1), 523–547.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*, 182–190.

Richardson, C. W., and D. A. Wright (1984), A model for generating daily weather variables, *Publ. ARS-8*, 83 pp., Agric. Res. Serv., U.S. Dep. of Agric., Washington, D. C.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Semenov, M. A., and E. M. Barrow (1997), Use of stochastic weather generator in the development of climate change senarios, *Clim. Change*, *35*, 397–414.

Semenov, A. M., R. J. Brooks, E. M. Barrow, and C. W. Richardson (1998), Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates, *Clim. Res.*, *10*, 95–107.

Stockle, C. O., P. Steduto, and D. A. Wright (1998), Estimating daily and daytime mean VPD from daily maximum VPD, paper presented at 5th Congress, Eur. Soc. Agron., Nitra, Slovakia.

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.

Wilks, D. S. (1997), Forecast values: Prescriptive decision studies, in *Economic Values of Weather and Climate Forecasts*, edited by R. W. Katz and A. H. Murphy, pp. 109–145, Cambridge Univ. Press, New York.

Young, K. C. (1994), A multivariate chain model for simulating climatic parameters from daily data, *J. Appl Meteorol.*, *33*, 661–671.

D. Allard, Biostatistiques et Processus Spatiaux, INRA, Site Agroparc, F-84914 Avignon CEDEX 9, France. (allard@avignon.inra.fr)

N. Brisson and C. Flecher, Agroclim, INRA, Site Agroparc, F-84914 Avignon CEDEX 9, France.

P. Naveau, Laboratoire des Sciences du Climat et de l'environnement, UMR 8212, UVSQ, IPSL, CNRS, CEA, Orme des Merisiers, Bat. 701 C. E. Saclay, F-91191 Gif-sur-Yvette CEDEX, France.