Clustering a Global Field of Atmospheric Profiles by Mixture Decomposition of Copulas

MATHIEU VRAC*

Laboratoire de Météorologie Dynamique, Ecole Polytechnique, Institut Pierre Simon Laplace, Palaiseau, and Centre de Recherche de Mathématiques de la Décision, Université Paris IX Dauphine, Paris, France

ALAIN CHÉDIN

Laboratoire de Météorologie Dynamique, Ecole Polytechnique, Institut Pierre Simon Laplace, Palaiseau, France

EDWIN DIDAY

Centre de Recherche de Mathématiques de la Décision, Université Paris IX Dauphine, Paris, France

(Manuscript received 21 October 2004, in final form 10 March 2005)

ABSTRACT

This work focuses on the clustering of a large dataset of atmospheric vertical profiles of temperature and humidity in order to model a priori information for the problem of retrieving atmospheric variables from satellite observations. Here, each profile is described by cumulative distribution functions (cdfs) of temperature and specific humidity. The method presented here is based on an extension of the mixture density problem to this kind of data. This method allows dependencies between and among temperature and moisture to be taken into account, through copula functions, which are particular distribution functions, linking a (joint) multivariate distribution with its (marginal) univariate distributions. After a presentation of vertical profiles of temperature and humidity and the method used to transform them into cdfs, the clustering method is detailed and then applied to provide a partition into seven clusters based, first, on the temperature profiles only; second, on the humidity profiles only; and, third, on both the temperature and humidity profiles. The clusters are statistically described and explained in terms of airmass types, with reference to meteorological maps. To test the robustness and the relevance of the method for a larger number of clusters, a partition into 18 classes is established, where it is shown that even the smallest clusters are significant. Finally, comparisons with more classical efficient clustering or model-based methods are presented, and the advantages of the approach are discussed.

1. Introduction

The major role played by a priori information in the problem of retrieving atmospheric variables from satellite vertical sounder observation can be understood by considering the fact that radiance observed by the sensor integrates the atmospheric thermal structure over relatively thick layers. Such an integration results in the well-known problem of the nonuniqueness of the solution. By specifying an initial guess solution as close as possible to the final correct solution, any a priori information may help to overcome this difficulty. For example, the improved initialization inversion (3I) method (Chédin et al. 1985; Scott et al. 1999) makes systematic use of available a priori information for retrieving the best possible temperature and water vapor profiles initial guess through a pattern recognition type approach. A priori knowledge of the airmass types, optimally at synoptic scale (Davis and Walker 1992; Kalstein et al. 1993), should be very useful because the average temperature or moisture profiles and associated variances may be very different for each airmass class. Providing full description of the vertical atmospheric column, operational meteorological analyses

* Current affiliation: Center for Integrating Statistical and Environmental Science, University of Chicago, Chicago, Illinois.

Corresponding author address: Dr. Mathieu Vrac, Center for Integrating Statistical and Environmental Science, University of Chicago, 5734 South Ellis Avenue, Chicago, IL 60637.
E-mail: vrac@uchicago.edu

are at the basis of the method we have developed to determine the type of the air mass observed. This method allows one to cluster a set of atmospheric profiles by characterizing each type of air mass in a statistical sense, opening a way to model the distribution of each variable (temperature, moisture) and the probability of appearance of any new atmospheric temperature and/or moisture profile. This method, dealing with probability distribution functions data, also called cumulative distribution functions data, is based on the notion of mixture densities (Diday et al. 1974; Dempster et al. 1977) and on the statistical modeling of dependencies between and among the temperature and moisture variables, through the so-called copula functions (Sklar 1959; Schweizer and Sklar 1983).

From a sample of N units (e.g., p-dimensional vectors of temperature or moisture values), the classical problem of mixture densities consists of estimating a probability density function f (e.g., of temperature) as a finite sum of K-weighted parametric densities (K given),

$$f(x_1, \ldots, x_p) = \sum_{k=1}^{K} p_k f(x_1, \ldots, x_p, \alpha_k), \qquad (1)$$

where $f(., \alpha)$ is a density with parameter α belonging to $\mathbb{R}^d$, p_k is the probability that a unit from the sample follows the law with density $f(., \alpha_k)$ with $\forall k = 1, \ldots, K$, $0 < p_k < 1$ and $\Sigma_{k=1}^{K} p_k = 1$. This problem has been investigated by many authors with two different approaches. The most widespread approach (called estimation approach) deals with the estimation of the mixture parameters (p_k, α_k) (Dempster et al. 1977; Everitt and Hand 1981). Classical methods for the estimation of these parameters maximize the likelihood. The most powerful algorithms rely on the expectation maximization (EM) method, as in Dempster et al. (1977) and studied by Redner and Walker (1984). Once the parameters are estimated, a partition $P = (P_1, \ldots, P_K)$ into K clusters can be made by applying the maximum a posteriori principle (MAP),

$$P_k = \{x; f(x, \alpha_k) \geq f(x, \alpha_j), j = 1, \ldots, K; j \neq k\}. \qquad (2)$$

The second approach (called clustering approach), looks for a partition into K clusters such that each cluster P_k can be seen as a subsample with density $f(x, \alpha_k)$. With such an approach, the currently used algorithms rely on the so-called dynamical clustering method found in Diday et al. (1974) and consist, at each step, of estimating the best parameters (p_k, α_k) according to a classifier log-likelihood criterion (Scott and Symons 1971), and then in defining the new partition $P = (P_1, \ldots, P_K)$ according to the MAP Eq. (2), and so on. For more details, see Symons (1981). Obviously, as this approach uses a different criterion, it provides slightly different estimates than the estimation approach. For both approaches, the number of clusters has to be given. Finding the right number is still a hard question in clustering analysis and in mixture models because there are no completely satisfactory methods. A very common approach consists of running the mixture models several times for various numbers of clusters and using the number that maximizes a given criterion, generally based on the likelihood. More specifically, bootstrapping methods, likelihood ratio tests, and Bayesian or classification-based information criteria for instance, have been applied to assess this number. A very nice review of some of these methods is given in McLachlan and Peel (2000).

In this study, we apply a model-based clustering approach (i.e., a clustering approach type) to a set of temperature and moisture atmospheric profiles to characterize air masses. To get a good modeling of dependencies between the variables and among each variable, we use a particular family of statistical distributions: the so-called copula functions (Sklar 1959; Schweizer and Sklar 1983; Vrac 2002). Copulas are particular distribution functions that link the joint (i.e., multidimensional) cumulative distribution functions of a multivariate random variable $\mathbf{X} = (X_1, \ldots, X_n)$, to the marginal (i.e., one-dimensional) cumulative distribution functions of each univariate random variable X_i. This link, giving an analytic modeling of dependencies is expressed through the most important result in copula theory, that is Sklar's theorem. However, copulas can also be used as statistical distribution with uniform marginal distribution.

Moreover, the model-based clustering approach by mixture copulas we present here, deals with cumulative distribution functions (cdf) data. The cumulative distribution function F (sometimes called distributions in this paper) of a real random variable X is defined by $F(x) = \mathbb{P}(X \leq x)$. Hence, the initial dataset we want to study contains vectors of cdf data (e.g., bivariate vectors of temperature and moisture cdf) instead of classical numerical values. Such data are usually studied in functional data analysis (FDA), developing data analysis, and statistical methods for functional data (Ramsay and Silverman 1997, 2002), or in symbolic data analysis (SDA), our approach in the present paper, extending data analysis methods to more complex data than classical numerical or categorical data (e.g., intervals, sequences of weighted values, histograms, density functions, etc.) See Bock and Diday (2000) for a review of SDA.

First, we present the data and detail the process we use to transform our dataset from numerical to cdf data.

Presentation and transformation are made in section 1. Developing a mixture model method for such cdf data requires adapting the concept of distribution. This is given in section 3 through the concept of distribution function of distributions (DFDs). This section explains what copulas are, associates the adapted distribution (DFD) with copulas, and presents the key Sklar's theorem. The general shape of the mixture copulas formula applied to DFDs and an algorithm for solving it are given in section 4. Finally, this algorithm is applied to our cdf data in section 5 and several results and statistical descriptions are given, showing the very good relevance of the air masses obtained. Comparisons with more classical efficient clustering and model-based methods are done on section 6, where we show that our approach gives results that are more satisfactory.

2. From numerical data to cumulative distribution function data

The atmospheric dataset we use comes from the European Centre for Medium-Range Weather Forecasts (ECMWF). Data points are realized as grid points over the earth at each latitude and longitude degree, and extended in altitude to 50 data point levels called sigma coordinates, from the surface to about 65 km. Several products (e.g., the temperature and the specific humidity) are recorded at each 3D latitude × longitude × altitude grid point every six hours (at 0000, 0600, 1200, 1800 UTC). Suppose our present specific interest is the field of world temperatures for 0000 UTC 15 December 1998 (the reasoning will be the same for specific humidity). The objective then is to partition the weather world into well-defined temperature (humidity) regions meteorologically or synoptically coherent and to include estimation of the underlying probability distribution function for each identified region.

The method we propose to apply allows us to work with cdf data, instead of having classical numerical data. A cdf modeling of the initial numerical data allows us to reduce the data, for computational purposes, in a very effective way. Indeed, knowledge contained in a single value, say $F(x)$ of a cdf F, is more important than the one contained in any single numerical data point: the single cdf value $F(x)$ gives us a piece of information about what happened before x. Then, we have to compute the cdfs associated to the profiles and to the variables we want to study. In this article, we only work with the temperature and specific humidity profiles. So, for each temperature profile, we estimate a cdf from the 37 first temperature values (surface to 30 km altitude), and for each humidity profile, a cdf is computed from the 24 first humidity values (surface to 13.5 km altitude). Values above about 30 km for the temperature and above about 13.5 km for the humidity are not considered as being accurate. Therefore, we just use the most precise forecasts to estimate two cdfs (temperature and humidity) for each profile. This is done by assuming that all the values (e.g., of temperature) in a given profile have the same distribution law: for each profile, we use the Parzen's kernel method to estimate a cdf. For instance, for the temperature (the humidity), this nonparametric method uses the $p = 37$ ($p = 24$) values $(X_1, \ldots, X_p)$ closest to the surface to compute $\hat{f}(x)$, the estimate probability density function (pdf) with

$$\hat{f}(x) = \frac{1}{ph} \sum_{i=1}^{p} \mathrm{Ke}\left(\frac{x - X_i}{h}\right), \qquad (3)$$

where Ke is called a kernel function (usually, and in this article, it is the normal pdf) and h is the window width, automatically estimated in the following according to the mean integrated square error (MISE), $h = 1.06\sigma p^{-1/5}$, with σ equal to the sample standard deviation. For details on h estimation or on cdf or pdf estimation, see (Silverman 1986) for a review of methods. From the dataset $(X_1, \ldots, X_p)$, we could have used a parametric distribution (such as a normal distribution) to model the cdfs associated with each profile. But a parametric modeling approach would have supposed knowledge about the general shape of our distributions, which we do not have. Actually, Eq. (3) can be seen as a mixture (i.e., a weighted sum) of p parametric normal distributions, each one associated with one X_i in the database $(X_1, \ldots, X_p)$. The Parzen's kernel method given in (3) does not suppose any shape for the distribution to be estimated. Obviously, this implies that we lose the notion of where the T and q appear in the vertical profiles, which would have been true also for a parametric distribution, and so, we may lose level features corresponding to specific profile characteristics (e.g., low-level inversion). However, this simple cdf modeling approach gathers information from the entire vertical profile and, consequently, reduces the number of data to be dealt with in a very effective way. By applying this approach, each profile has a temperature and a specific humidity cdf. Considering all the world profiles, we have 360 × 180 profiles. To reduce the total number of items while preserving information, we develop the temperature (or humidity) patterns for every other (i.e., 2° apart) latitude–longitude grid point. There are therefore (360/2) × (180/2) = 16 200 such locations and therefore 16 200 cumulative distribution functions.

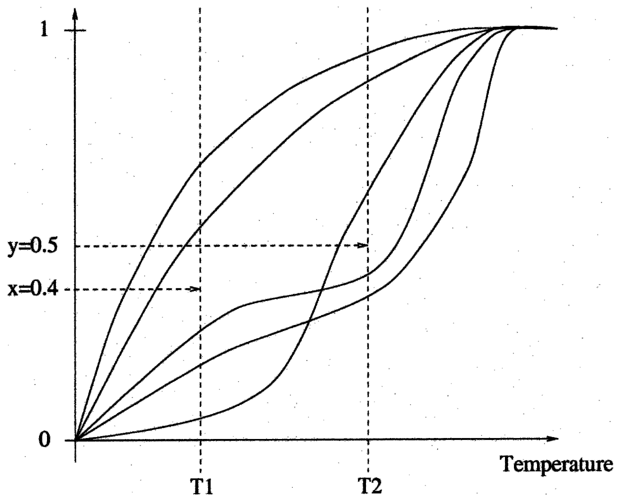

FIG. 1. An example of five temperature cumulative distribution functions.

3. Distribution of cdf data and copulas

Developing an extension of the mixture density algorithms for cdf data requires adapting the concepts of pdf or cdf. The concept of distribution of distributions has been developed for the first time in a general probabilistic context by Vrac (2002) and theoretical details may be found in (Diday and Vrac 2005). Here, we just give the essentials of this notion.

a. Distribution function of distributions

For simplicity reasons, we consider the case of only one product, such as temperature. We work on a set $\Im$ of N-estimated temperature cdfs $F_1, \ldots, F_N$, for example, the 16 200 temperature cdfs. For a given value T of temperature and a given value x of probability (i.e., $x \in [0, 1]$), we would like to know how many chances we have to get a cdf F_i (randomly chosen among $\Im$), which is lower than x at point T. That is, what is the probability to get $F_i(T) \leq x$? This probability, defined for each temperature value T, is a function of $x \in [0, 1]$: this function is what we call a distribution function of distributions (DFD). More formally, a DFD at a given T, denoted by $G_T(x)$, is a function from [0, 1] to [0, 1] defined by

$$G_T(x) = \mathbb{P}(\{F_i \in \Im | F_i(T) \leq x\}) \quad (4)$$

corresponding to the cdf of a random variable with values in a probabilistic space (i.e., instead of having numerical realizations of the random variable, we have cdfs). As an illustration, we consider in Fig. 1 five temperature cdfs estimated from five (arbitrarily chosen) temperature profiles as described in the previous section. By fixing a value T_1 of temperature, we can easily compute $G_{T_1}^{\mathrm{emp}}(x)$, an empirical version of $G_{T_1}(x)$ for all $x \in [0, 1]$. For example, in this figure, $x = 0.4$ and $G_{T_1}^{\mathrm{emp}}(x)$ corresponds to the percentage of cdfs taking a smaller value than 0.4 at the point T_1. In this case, there are only three cdfs checking it among five. So, $G_{T_1}^{\mathrm{emp}}(0.4) = 3/5$.

Obviously, we can extend DFDs to the multivariate case. For a given $\mathbf{T} = (T_1, \ldots, T_n)$ of temperature values and a given $\mathbf{x} = (x_1, \ldots, x_n) \in [0, 1]^n$ (x_i is a probability), how many chances do we have to get a cdf F_i (randomly chosen among $\Im$), which is lower than x_1 in T_1 and lower than x_2 in T_2 and . . . and lower than x_n in T_n? That is, what is the probability to get $F_i(T_1) \leq x_1$ and $F_i(T_2) \leq x_2$ and . . . and $F_i(T_n) \leq x_n$? This n-dimensional function of $(x_1, \ldots, x_n)$, denoted by $H_{\mathbf{T}}$, is defined from $[0, 1]^n$ to [0, 1] by

$$H_{\mathbf{T}}(\mathbf{x}) = \mathbb{P}(\{F_i \in \Im | F_i(T_1) \leq x_1; \ldots ; F_i(T_n) \leq x_n\}) \quad (5)$$

and corresponds to the joint cdf of a n-dimensional random variable with probabilistic values. The function $H_{\mathbf{T}}$ is called a joint distribution function of distributions (JDFD). In Fig. 1, we see that we only have two temperature cdfs that are lower than 0.4 in T_1 and lower than 0.5 in T_2. Therefore, the empirical estimation in (0.4, 0.5) of the bivariate JDFD at the given $\mathbf{T} = (T_1, T_2)$ is $H_{\mathbf{T}}^{\mathrm{emp}}(0.4, 0.5) = 2/5$. Actually, the DFD (JDFD) should be called a single (multidimensional) point of a distribution function of distributions rather than a DFD, but when the number of T_is is big enough, we can suppose that the entire continuous DFD is well-characterized and then that the multidimensional point of DFD is a good approximation of the complete DFD. Hence, for simplicity, in the following we will talk about JDFD for multidimensional point of DFD and DFD for single point of DFD. More theoretical details and considerations on DFDs and JDFDs, with several variables and products, are given in Vrac (2002) and Diday and Vrac (2005).

b. Modeling and estimating points of DFD

Although an empirical model for DFD and JDFD can be useful for graphical representation of JDFDs, for example, it does not enable one to get continuous functions. In this subsection, we present a way to model continuous DFDs and JDFDs. Modeling and estimating a DFD (respectively a JDFD) corresponds to modeling and estimating a cdf from [0, 1] ($[0, 1]^n$, respectively) to [0, 1], or a pdf and then integrating. For this purpose, several statistical distributions and methods exist. For the presented application, we only work with one statistical distribution: the beta distribution, which is the Dirichlet's distribution in one dimension. Its pdf is

$$f_{\nu_1,\nu_2}(x) = \frac{x^{\nu_1-1}(1-x)^{\nu_2-1}}{\int_0^1 y^{\nu_1-1}(1-y)^{\nu_2-1}\,dy}, \qquad (6)$$

with $\nu_1 > 0$ and $\nu_2 > 0$ the two parameters. From a temperature cdf dataset $\Im = \{F_1, \ldots, F_N\}$, and for a given temperature value T, the two parameters ν_1 and ν_2 have to be fitted to the dataset $\{F_1(T), \ldots, F_N(T)\}$, for example by the Newton–Raphson algorithm (see Press et al. 2002). The DFD at point T, obtained by integrating f_{ν_1,ν_2}, characterizes the probabilistic position of the cdfs in $\Im$ at T. With two values T_1 and T_2, the JDFD is starting to give us an idea of the shape of $\Im$. To make a functional link between positions (e.g., in T_1 and T_2) and to get a statistical description of the global shape of $\Im$ (i.e., a link between a JDFD and its two marginal DFD), the copulas' theory is used.

c. Modeling dependence between DFDs with copulas

The word copula first appeared in the journal *Statistics* in 1959 in an article by Sklar (1959). The name comes from the fact that copula functions couple together a joint cdf with its marginal cdfs (or margins). These functions can be seen as particular cumulative distribution functions with uniform marginal distributions. We do not give the formal definition of these functions (see Nelsen 1998; Schweizer and Sklar 1983), but only the most important result on copulas given by Sklar's theorem. Using a similar notation as in section 3b, Sklar's theorem says that if we have H, an n-dimensional cdf with marginal cdfs $G_1, \ldots, G_n$, then there exists a copula function C coupling together H and $(G_1, \ldots, G_n)$. More formally, there exists a copula C such that for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$H(x_1, \ldots, x_n) = C[G_1(x_1), \ldots, G_n(x_n)]. \qquad (7)$$

Then, when fixing, for example, two temperature values T_1 and T_2 and modeling the associated temperature DFDs (with beta laws), G_{T_1} and G_{T_2}, we can model the JDFD $H_\mathbf{T}$ with $\mathbf{T} = (T_1, T_2)$ as a composition of a copula function and the two DFDs. Formally, there exists a copula C such that $\forall x_1, x_2 \in [0, 1]$,

$$H_\mathbf{T}(x_1, x_2) = C[G_{T_1}(x_1), G_{T_2}(x_2)]. \qquad (8)$$

Obviously, copulas are a very specific class of functions. There exists many copulas and most studies work on a class of parametric copulas: the Archimedian copulas (see Nelsen 1998; Vrac 2002). In our application, we make use of Frank's family (Genest 1987) given by

$$C_\beta(u, v) = \frac{\log\left[1 + \frac{(\beta^u - 1)(\beta^v - 1)}{(\beta - 1)}\right]}{\log\beta}, \qquad (9)$$

with $u, v \in [0, 1]$, β strictly positive and $\beta \neq 1$. This family has the following noteworthy properties, helpful to explain the clusters of the future partitions:

- $\lim_{\beta\to 0} C_\beta(u, v) = \min(u, v)$ (copula called min copula or lower bound of Fréchet-Hoeffding),
- $\lim_{\beta\to 1} C_\beta(u, v) = uv$ (copula of independence),
- and $\lim_{\beta\to\infty} C_\beta(u, v) = \max(u + v - 1, 0)$ (copula called upper bound of Fréchet-Hoeffding).

When the marginal DFDs are known, estimation of the β parameter is based on likelihood criterion with a Newton–Raphson method. When this is not the case, β can be estimated from Kendall's τ or Spearman's ρ (see Genest and Rivest 1993). For more details on copulas' properties and Archimedian copulas, see Nelsen (1998), Schweizer and Sklar (1983), and Vrac (2002). In the following, two temperature values $\mathbf{T} = (T_1 = 225$ K, $T_2 = 265$ K) and two specific humidity values $\mathbf{q} = (q_1 = 3.10^{-5}$ kg kg^{-1}, $q_2 = 6.10^{-3}$ kg kg^{-1}) are fixed for estimating temperature and humidity DFDs. This choice can appear quite arbitrary. Most of the time, the expert knowledge is essential in choosing the values of a fixed number of T_is. However, some numerical approaches can be used, such as the triangle method or the surface of distributions, to help the expert to make or confirm this choice. These two methods are not given here but are detailed in Diday and Vrac (2005) and support the choice made in this paper, leading us to choose the above values. In this application, only two values of distribution are kept for each physical variable because the higher the dimensionality, the slower the convergence and the more complex the copulas to be explained. Indeed, n-dimensional copulas, Archimedian or not, are numerous but the simplest ones induce a high symmetry in dependencies between variables and the most complex copulas are hard to understand (see, e.g., Nelsen 1998; Hillali 1998). In the following application, all the DFDs are modeled with beta laws and the relationships between the JDFDs ($H_\mathbf{T}$ for temperature and $H_\mathbf{q}$ for humidity) and their marginal DFDs (G_{T_i} for temperature and G_{q_i} for humidity) are modeled by Frank's copulas.

4. Mixture decomposition of copulas

Our goal is dual: first, we would like to get a probabilistic model to describe our (cdf) data and to compute the probability of having each temperature and humidity cdf. Secondly, we want to cluster the dataset $\Im$ into K clusters. Each cluster must describe a coherent set of atmospheric profiles (cdfs). Profiles in the same cluster must be close to each other and profiles in different clusters must be distant enough. To reach these two goals, we only use one method: mixture decomposition of copulas. First, we consider our dataset $\Im$ to be only the 16 200 temperature cdfs from section 2. Setting $T_1 = 225$ K, $T_2 = 265$ K, and using $\Im$, we are looking for the JDFD of temperature $H_\mathbf{T}$. We assume that $H_\mathbf{T}$ is a mixture of K JDFDs, that is, for all (x_1, x_2) in [0, 1], $H_\mathbf{T}(x_1, x_2)$ is written as a sum of K parametric JDFDs $H_\mathbf{T}^k(x_1, x_2)$ $(k = 1, \ldots K)$, each one weighted by a mixture ratio p_k $(k = 1, \ldots, K)$. Sklar's theorem, through the relations (8) or (7), can be applied to each $H_\mathbf{T}^k$ with margins $G_{T_1}^k$ and $G_{T_2}^k$, and then the initial mixture of JDFDs becomes a mixture of copula functions C_k applied to a couple of (points of) DFDs. In the following, C_k is taken to be a bivariate Archimedian Frank's copula. Let $h(x_1,x_2) = \partial^2 H_\mathbf{T}/\partial x_1,\partial x_2$ denote the pdf associated with $H_\mathbf{T}$, and $h_k = = \partial^2 H_\mathbf{T}^k/\partial x_1,\partial x_2$ the pdf associated with $H_\mathbf{T}^k$. When working with pdfs instead of cdfs, the initial mixture of JDFDs becomes now

$$h(x_1, x_2) = \sum_{k=1}^{K} p_k \left[\prod_{i=1}^{2} \frac{dG_{T_i}^k}{dx}(x_i) \right] \times \frac{\partial^2 C_k}{\partial u_1 \partial u_2} [G_{T_1}^k(x_1), G_{T_2}^k(x_2)]. \quad (10)$$

From this equation, the mixture densities algorithms can be extended. We propose here to extend and apply a clustering approach algorithm: the CEM algorithm that allows us to get a model and clusters from numerical data. This algorithm (Celeux and Govaert 1992) consists of adding a clustering step (C-step) in the EM method as in Dempster et al. (1977). Theoretical and practical comparisons between the classical EM algorithm, its stochastic version (SEM) and the CEM algorithm can be found in Celeux and Govaert (1992, 1993). In our case, CEM can be summarized as follows: Given a partition $\mathbf{P} = (P_1, \ldots, P_K)$ (usually randomly generated), the clustering algorithm is defined in two steps. The first one consists of estimating the mixture parameters in (10): the mixture ratios p_k with $p_k = [\mathrm{card}(P_k)/\mathrm{card}(\Im)]$, the parameters $(b_i^k)i = 1, 2$, $k = 1, K$ of each DFD G_{T_i} in component k [$b_i^k = (\nu_1^{T_i,k}, \nu_2^{T_i,k})$, parameters of the beta laws], and the copula parameters $(\beta_k)_{k=1,\ldots,K}$ by maximizing the classifier log-likelihood (also called classification log-likelihood) CL,

$$\mathrm{CL}(\mathbf{P}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{F_j \in P_k} \log\{h_k[F_j(T_1), F_j(T_2)]\}, \quad (11)$$

with $\boldsymbol{\theta} = \{\beta_k, b_1^k, b_2^k\}_{k=1,\ldots,K}$, and $F_j(T_i)$ the value of the cdf F_j in T_i.

The second step consists of allocating the cdfs into K new clusters by

$$P_k = \{F_i; p_k h_k[F_i(T_1), F_i(T_2), \theta_k] \geq p_m h_m[F_i(T_1), F_i(T_2), \theta_m] \forall m\}, \quad (12)$$

with $k < m$ in case of equality. Then, we go back to the first step until convergence of the criterion CL defined in (11). At convergence, we obtain a partition into K clusters of the cdf dataset, a statistical model for each cluster and a global model as a mixture of copulas.

Let us define the simple decision rule

$$P_k = [F_i; d(F_i, \overline{F}_k) < d(F_i, \overline{F}_m), \forall m], \quad (13)$$

with $d(F_i, \overline{F}_k)$ a distance (the Euclidian distance for example) between the couple $[F_i(T_1), F_i(T_2)]$ $[F_i(q_1), F_i(q_2)]$ and the couple $[\overline{F}_k(T_1), \overline{F}_k(T_2)]$ $[\overline{F}_k(q_1), \overline{F}_k(q_2)]$, where $\overline{F}_k$ is the mean distribution (of temperature or humidity according to the context) in cluster k. Once the classes are defined, one can wonder to what extant the use of this simpler decision rule to assign new profiles would change the results. Given our model-based clustering approach, let us take a simple example with a two-component one-dimensional Gaussian mixture (easier than a mixture of copulas for the example). Let us assume that we have a first cluster, c_1, with a Gaussian pdf narrow around the mean m_1 (i.e., small variance σ_1^2) and a second cluster, c_2, much larger around its mean m_2 (i.e., big variance σ_2^2). We suppose that the clusters are well separated. In the case of a new value $\upsilon = (m_1 + m_2)/2 - \epsilon$ with a small $\epsilon > 0$ [i.e., υ is slightly on the left of the middle of (m_1, m_2)], the decision rule (13) implies that υ belongs to cluster c_1 with the narrow pdf. But if we take into account the pdfs, as it is done in Eq. (12), the new value υ will clearly be assigned to cluster c_2 (see Fig. 2). That means that the simple decision rule (13) does not take into account the distribution associated with each cluster. Actually, this rule implicitly assumes the same variance for the two clusters.

5. Application to the clustering of a set of thermodynamical profiles

As explained in section 1, there are no completely satisfactory methods to assess g, the number of clusters or components in clustering analysis and in mixture models. In the present application we used an approximate Bayesian solution to the choice of g using the classification maximum likelihood (ML) approach. This

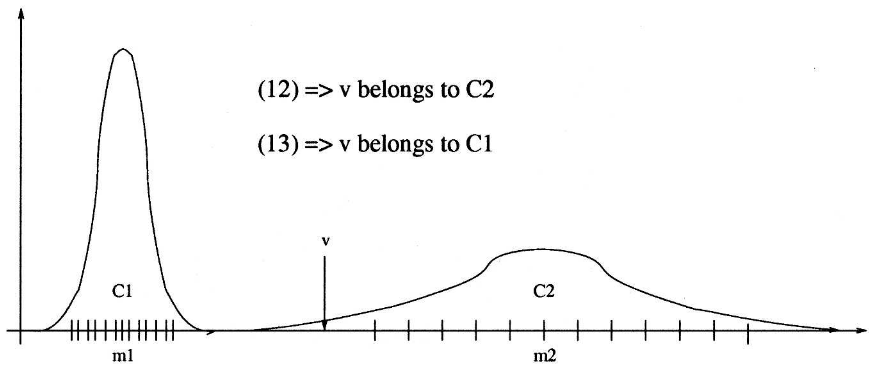

FIG. 2. An example of the decision rule.

approximation suggested and detailed by Banfield and Raftery (1993), leads to the approximate weight of evidence (AWE) criterion,

$$\mathrm{AWE}(g) = -2\log(L_C) + 2d[3/2 + \log(N)], \quad (14)$$

where d is the number of parameters to be estimated when g components are in the mixture, L_C is the classification ML associated with (11) and N is the number of atmospheric profiles. The 3I algorithm mentioned in section 1, worked with a five-cluster partition for several years, corresponding to two polar clusters, two temperate clusters and one tropical (Chédin et al. 1985; Achard 1991). Because of the small number of clusters, this partition assumed equivalent behavior (i.e., similar thermodynamic profiles) in the winter in the North Hemisphere and in the summer in the South Hemisphere (and conversely). This is obviously not true. To improve the partition used for the 3I algorithm, a mixture of copulas has been run for g ranging from 7 to 18 (six has not been tested because we wanted to favor an odd number of clusters to keep a kind of symmetry to the tropical cluster). For this range, the criterion given in (14) is maximized for seven clusters for the temperature and for the humidity. Hence, in the following sections, the seven-cluster partition is presented.

a. Clustering of temperature profiles into seven clusters

Temperature profiles for 0000 UTC 15 December 1998 are classified into seven clusters with the Frank's copula and with a DFD modeled with a beta law with parameters ν_1 and ν_2. From the 16 200 distributions, the algorithm converges in two iterations to the partition in Fig. 3a. Parameters of copulas and DFD are given in Table 1. The clusters of the partition look coherent and identify realistic climatic regions: one large tropical class (cluster 4), two polar classes, winter in the North Pole (cluster 1) and summer in South Pole (cluster 7), two temperate classes (clusters 2 and 5). Cluster 3 links midlatitude and tropical zones and cluster 6 links polar and midlatitude zones. Some high relief is identified (Himalayas, Andes), in spite of the use of the sigma coordinates. Moreover, clusters 1, 4, and 5 have similar parameters of copulas (10^{-6}) near zero, meaning that their copulas are near the so-called min copula $C(u, v) = \min(u, v)$ (see section 3c). This value implies that for these clusters, our method has grouped together distributions having a tendency to have the same shape: within each one of these three clusters, the functional data that we handle do not cross over each other (i.e.,

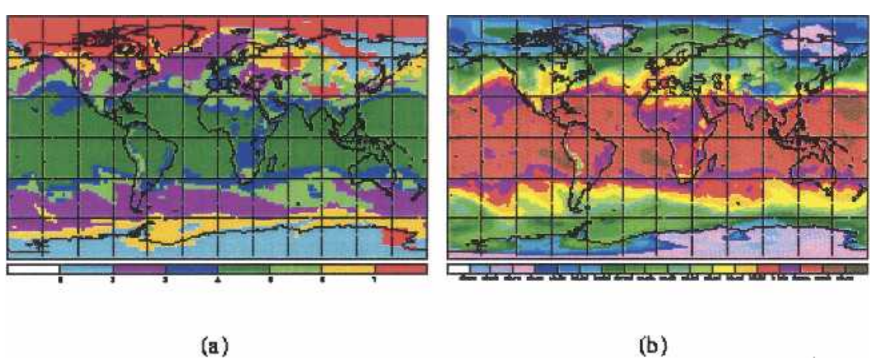

FIG. 3. (a) Clustering on temperature distributions and (b) mean temperature (500–700 hPa).

TABLE 1. Parameters of the clustering into seven clusters for temperature.

Classes	β	ν_1 in T_1	ν_2 in T_1	ν_1 in T_2	ν_2 in T_2
1	0.000 001	6.836 969	14.342 546	12.208 704	2.217 218
2	0.300 001	11.408 380	69.945 442	21.956 680	14.064 272
3	0.004 093	12.180 901	70.0	61.601 810	70.0
4	0.000 001	12.651 747	70.0	56.703 354	70.0
5	0.000 001	13.335 871	70.0	11.891 472	11.261 731
6	0.030 567	6.040 135	25.066 311	8.938 780	3.687 328
7	0.007 445	8.839 353	22.021 719	19.165 813	2.168 266

they evolve in a parallel way). Moreover, an interesting comparison can be made with the map of the mean temperature between 500 and 700 hPa (Fig. 3b). Transition shapes between midlatitude and tropical zones agree well with that of Fig. 3b. The tongues (incursion of warm air mass into colder air mass or conversely) are well identified. The red disc (cluster 7) located at 60°N, 60°E is perfectly explained with Fig. 3b, and corresponds to a depression. More generally, the fit with the synoptic analysis of the situation is good (shape of hot or cold air incursions, depression, etc.).

Another way to describe the clusters is to look at the pdfs of the temperature at different pressure levels. These pdfs, determined using the Parzen's kernel method, bring information on the distribution of the temperature values in each cluster and on their degree of discrimination. For instance, in Figs. 4a,b [corresponding to pdfs at levels 900 hPa (about 1 km altitude) and 500 hPa (5.5 km altitude), respectively], the clusters appear well discriminated. In these figures, the y axis corresponds to the values of the normalized pdfs (i.e., when minimal temperature value = 0 and maximal temperature value = 1). As expected, clusters 1 and 7 are the two coldest, cluster 4 the warmest, and clusters 3 and 6 (closer to the Tropics and closer to the poles, respectively) are transition clusters. From the result in Fig. 3a, we can sort out the clusters, from the coldest (near the poles) to the warmest (near the Tropics): cluster 6, cluster 2, cluster 5, and cluster 3. This order is verified by the pdfs. However, the higher the altitude (i.e., the smaller the pressure), the smaller is the discrimination between the clusters. It is the case above the tropopause. This is shown on Fig. 4c illustrating the results at 70 hPa, 20 km altitude (stratosphere). In this example, the tropical cluster 4 is very distinct, and becomes the coldest, as expected by the climatology.

b. *Clustering of specific humidity profiles into seven clusters*

A similar clustering into seven clusters has been carried out on specific humidity distributions. We fixed two values $q_1 = 3.10^{-5}$ kg kg^{-1} and $q_2 = 6.10^{-3}$ kg kg^{-1}. Parameters of copulas and DFDs are given in Table 2

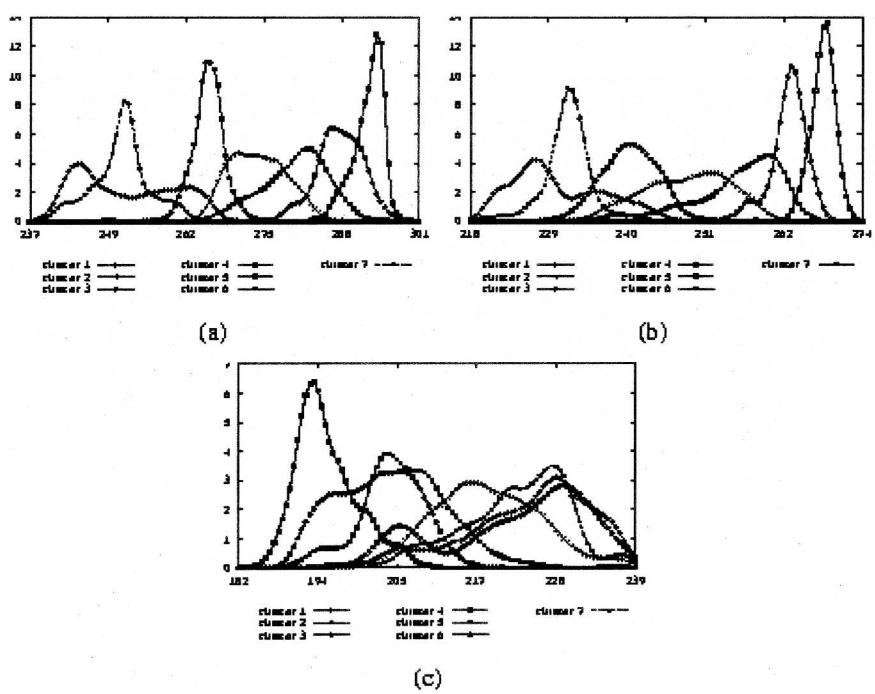

FIG. 4. Densities of temperature (K) for each cluster at several pressure levels: (a) 900 hPa (1 km); (b) 500 hPa (5.5 km); (c) 70 hPa (20 km).

TABLE 2. Parameters of the clustering in seven classes for the humidity.

Classes	β	ν_1 in q_1	ν_2 in q_1	ν_1 in q_2	ν_2 in q_2
1	0.200 01	1.772 749	36.192 062	70.0	24.484 861
2	0.016 939	6.292 977	742.659 424	30.004 604	13.469 539
3	0.619 099	0.000 001	12.617 126	16.619 267	20.368 376
4	0.445 017	0.000 001	12.617 126	29.424 589	48.242 210
5	0.100 001	0.000 001	12.617 126	6.890 862	5.887 025
6	0.020 641	0.000 001	12.617 126	38.931 557	14.840 351
7	0.017 804	2.215 375	23.266 142	70.0	18.847 424

and the resulting clusters are shown in Fig. 5a. This map is compared to the map of vertically integrated vapor content (TWVC) for each profile (Fig. 5b). Features in common between Figs. 5a,b are numerous. Incursions of humid air are precisely defined and most of the shapes are found with great precision. The previous tropical cluster is split into two clusters (3 and 4). Cluster 4 contains the most humid zones. Moreover the boundary between cluster 3 and cluster 2 (less humid air mass) is edged with cluster 5 (humid air incursions in a drier air mass). The method identifies two clusters with low humidity values (clusters 1 and 7). These later two have different parameters of copulas (0.2 and 0.018) meaning a different behavior of their distributions. Moreover, a spiral at 60°N, 60°E fixes the position of a depression (corresponding to the red disc in the clustering on temperature; Fig. 3a). The spiral seen on this map is also present at the same latitude and longitude on the TWVC (Fig. 5b).

The pdfs for each cluster and at different pressure levels give information on the discrimination between clusters. Figure 6 corresponds to densities at 900 hPa (1 km). This figure shows well-separated clusters. As expected from Fig. 5a, clusters 1 and 7 are the driest and cluster 1, with the humid right wing of its pdf, corresponds to more frequent situations than the one shown by cluster 7, even drier and associated to polar winter situations. At upper levels, the rapid decrease of humidity leads to a greater overlapping of the pdfs. Comparison with the TWVC leads to the following conclusions:

- the pdfs of clusters 1 and 7 are the driest;
- cluster 4 is the most humid with a pdf having its mean value at 0.015 kg kg^{-1}; and
- the midlatitude and transition clusters can be sorted out from the less to the more humid: clusters 3, 5, 2, and 6.

Once we have the parameters from the classification on temperature (humidity), it is easy to find out to which one of the seven clusters a new atmospheric state vector, say $\mathbf{V}_T$ ($\mathbf{V}_q$), belongs. First, the cdf, for example, F^T for temperature (F^q for humidity), associated with this vertical profile has to be calculated by (3). The decision rule is based on the values F^T (T_1) and F^T (T_2) [F^q (q_1) and F^q (q_1)] with T_1 and T_2 (q_1 and q_2) given earlier. Then, from the parameters given in Table 1 (Table 2) and the mixture ratios (not given here), it is easy to apply (12) in order to define the correct class to assign the new atmospheric state vector. It has to be noted that we do not need to keep in memory the original data or information on the 16 200 vertical profiles, to classify an independent profile. Indeed, we use parametric modeling for DFDs (beta distributions) and copulas (Frank's copulas).

c. *Clustering of coupled temperature and humidity profiles into seven clusters*

To classify combined temperature and specific humidity profiles, a mixture decomposition of multidimensional copulas can be used. Here, the coupling method developed in (Vrac 2002) is applied. From the estimated mixture parameters (i.e., parameters of copu-

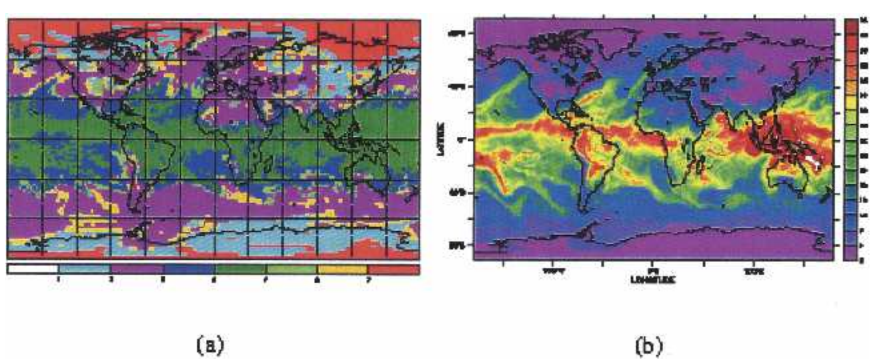

FIG. 5. (a) Clustering on specific humidity distributions and (b) total column water vapor (kg m^{-2}).

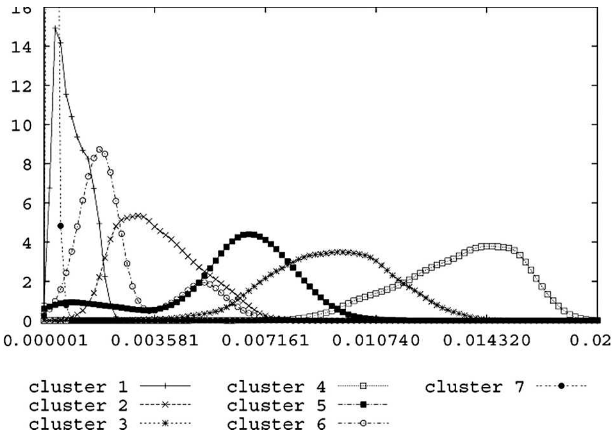

FIG. 6. Densities of humidity (kg kg^{-1}) for each cluster at 900 hPa (1 km).

las, DFDs, and mixture ratios), a value of JDFD (which is a value of cumulative distribution function) can be written for each variable (temperature and humidity) and for each profile $i = 1, \ldots, N$ with distribution F_i^T (for temperature) and F_i^q (for humidity), and described by the couples $(x_{T_1}^i, x_{T_2}^i) = [F_i^T(T_1), F_i^T(T_2)]$ (for temperature) and $(x_{q_1}^i, x_{q_2}^i) = [F_i^q(q_1), F_i^q(q_2)]$ (for humidity) by

$$H_{T_1,T_2}^T(x_{T_1}^i, x_{T_2}^i) = \sum_{k=1}^{7} p_k^T C_{\beta_k^T}[G_{T_1}^k(x_{T_1}^i, b_{T_1}^k), G_{T_2}^k(x_{T_2}^i, b_{T_2}^k)], \tag{15}$$

and

$$H_{q_1,q_2}^q(x_{q_1}^i, x_{q_2}^i) = \sum_{k=1}^{7} p_k^q C_{\beta_k^q}[G_{q_1}^k(x_{q_1}^i, b_{q_1}^k), G_{q_2}^k(x_{q_2}^i, b_{q_2}^k)], \tag{16}$$

where β_k^T (β_k^q) is the parameter of the copula in cluster k and the temperature (the humidity) and $b_{T_j}^k$ ($b_{q_j}^k$) is the parameter associated to the DFD defined in $T_j(q_j)$ in cluster k. For a profile w, a couple $[H^{\text{Temp}}(w), H^{\text{Hum}}(w)]$ of values of distributions is computed, and for the 16 200 profiles, 16 200 new couples are obtained. A mixture decomposition of copulas can be applied to this new database. The results of the partition in seven clusters are given in Table 3 and Fig. 7. The coupling gives good results with an apparent coherent mixture of the two previous partitions. In particular, the two tropical classes (cluster 4 and 3) are in much better agreement with TWVC (Fig. 5b) than were the same clusters on Fig. 5a, resulting from the clustering of the water vapor alone profiles. The other clusters (2, 5, and 6) are transitions from tropical classes (hot and humid) to polar classes (dry and cold). The spiral (60°N, 60°E) is present. In general, the details introduced by this combined clustering seem relevant: see for instance, a small colder and drier zone than its neighborhoods, south of Australia.

When comparing the densities of temperature at 900 hPa, 1 km altitude (Fig. 4a), from Fig. 3a, with those (Fig. 8a) from the partition by coupling, one can see that the last ones have a lower discriminating power. This effect was expected because the coupling accounts for the two variables considered here (temperature and humidity). However, the clusters stay relatively distinct in temperature until the tropopause is reached; densities at 900 hPa (1 km), Fig. 8a, and at 300 hPa (10 km), Fig. 8b, are somewhat less distinct than in Fig. 4a, but remain clearly separated. Above the tropopause (in the stratosphere), the pdfs tend to overlap. Figure 8c of pdfs at 70 hPa (20 km) gives an example of this effect. On this figure, cluster 4 (the warmest cluster between 900 and 300 hPa) is the coldest at 70 hPa. This is also true for cluster 3: one of the warmest clusters in the troposphere and one of the coldest in the stratosphere. This expected inversion for such tropical clusters, is perfectly highlighted. The pdfs of humidity at 900 hPa (Fig. 8d) from the coupling method are slightly less separated than the pdfs from the humidity only clustering (Fig. 6). However, most of the pdfs stay distinct and give a good description of the humidity characteristics of the clusters (cluster 4 is the lost humid, 1 and 7 are dry, etc). These seven clusters appear coherent and in agreement with the physical characteristics expected.

Once more, it is easy to figure out to which one of

TABLE 3. Parameters of the clustering into seven classes for temperature and humidity.

Clusters	β	ν_1 in Tq_1	ν_2 in Tq_1	ν_1 in Tq_2	ν_2 in Tq_2
1	0.000 001	6.712 667	2.140 064	5.703 492	5.222 391
2	0.100 001	70.0	70.0	10.424 58	14.541 202
3	0.200 001	18.965 822	88.125 916	8.056 098	145.218 979
4	0.050 867	19.533 854	112.066 284	6.489 847	357.520 905
5	0.362 295	12.315 609	31.493 969	5.033 236	18.545 059
6	0.126 157	0.864 89	7.177 879	3.316 219	7.178 005
7	0.003 896	23.222 773	4.773 149	13.366 582	3.108 013

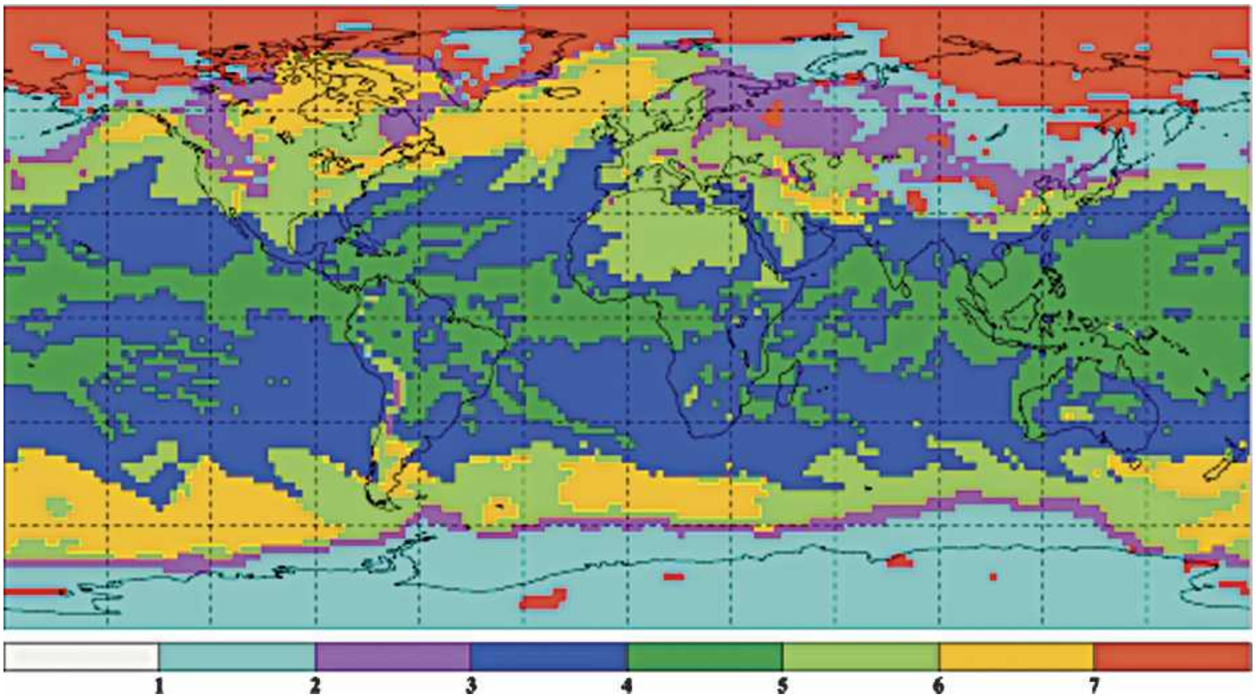

FIG. 7. Clustering in temperature and humidity cdfs (Frank's copula, DFD beta law).

these clusters a new independent vertical profile belongs. First of all, $H^T_{T_1,T_2}$ and $H^q_{q_1,q_2}$ have to be calculated by (15) and (16) from the parameters given in Tables 1 and 2 and mixture ratios (not given here). Then, from Table 3, we can apply Eq. (12) to find out the class where the new profile must be assigned, without keeping in memory information on the original profiles.

d. Clustering of coupled temperature and humidity profiles into 18 clusters

The choice of the number of clusters, even though we use a criterion, may appear quite arbitrary. For example, what happens when more than seven clusters are considered? Is every new cluster significant or are there clusters with no meaning? To try to answer these questions, we consider here a partition into 18 clusters. As previously, it is obtained by the coupling method with Frank's copulas [defined by (9)] and DFD modeled by beta law. The resulting partition is shown on Fig. 9. At first sight, the main result of the clustering into 18 clusters is to split clusters coming from the partition into seven clusters. For example: the tropical cluster of the partition in seven clusters (4/7 in the following) is split into the two clusters 10/18 and 8/18, the latter at the edges of cluster 3/7. In addition, cluster 3/7 is split into three new clusters: 5/18, 6/18, and 11/18. The pdfs of these three new clusters are well separated, covering the whole interval of the pdf of 3/7 (see Fig. 10). These pdfs show that density from 3/7 is a mixture of densities from 5/18, 6/18, and 11/18. This is true for any level of the atmosphere and for specific humidity too.

Other conclusions, similar to the seven cluster case, can be made. The pdfs of temperature are well separated below the tropopause; the distinction between pdfs of temperature is more difficult above the tropopause; the pdfs of specific humidity are distinct in the troposphere and overlap in altitude (i.e., the domain where the pdfs are not zero is about the same for each pdf when they are calculated with values over the tropopause).

Moreover, to tentatively assess the significance of the partition into 18 clusters, we have analyzed in detail cluster 13, which only includes 138 atmospheric profiles. This cluster can be seen in the North Atlantic up to Scandinavia, off the west coast of Canada, and in the Mediterranean Sea. In these three cases, the synoptic map (not presented here) shows a localized depression and strong winds. The cluster 13 actually reveals interesting information on the meteorological situations.

6. Comparisons

This original algorithm of mixture decomposition of copulas has been compared to other classical clustering methods applied to the same dataset. One of the most powerful clustering methods is the so-called mixed clus-

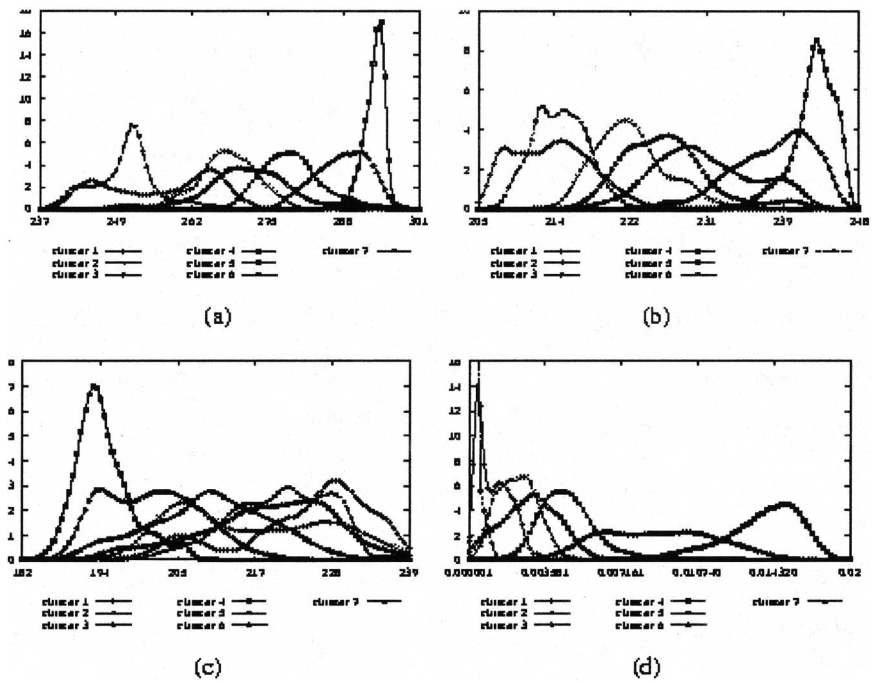

FIG. 8. Densities of (a)–(c) temperature (K) and (d) specific humidity (kg kg^{-1}) for each cluster from coupled clustering on temperature and humidity profiles. Temperature (K) at (a) 900 hPa (1 km), (b) 300 hPa (10 km), and (c) 70 hPa (20 km); and (d) densities of specific humidity (kg kg^{-1}) at 900 hPa (1 km).

tering (Molliere 1985). This strategy is well adapted to clustering huge databases and consists of a combination of a principal components analysis (PCA), a dynamical clustering method with cross-validation, a hierarchical ascending clustering method and a final stabilization by a new dynamical clustering method. For more details, see Vrac (2002). This procedure is considered to be very efficient.

The mixed clustering has been applied to our climatic database for the 0000 UTC 15 December 1998, and for the same variables (temperature and specific humidity). The 37 most reliable values of the temperature profiles (from the surface to about 30 km) and the 24 most reliable values of the specific humidity profiles (from the surface to about 13.5 km) are used for each situation. Each situation is thus described by 61 numerical variables, which are then normalized and weighted to give equal weight to both temperature variables and specific humidity variables. The algorithm is summarized in Fig. 11, and the results given in Fig. 12. Compared to Fig. 7, the depiction of the situation appears significantly looser. Tropical clusters are relatively well described, however with some marked exceptions: tropical moist air intrusions within the northern or southern midlatitudes are not properly seen. Mid-to-high latitude clusters have a pronounced zonal behavior that is less visible in Fig. 7 and in either Fig. 3b or, a fortiori, Fig. 5b. Compared to Fig. 7, many details are missing in Fig. 12. From the pdfs of each cluster (not shown), one can see that the discrimination is much weaker, compared to the partition into seven clusters with copulas. One conclusion is that the mixed clustering applied to this climatic database is less efficient than the mixture decomposition of copulas applied to our probabilistic database. Other comparisons have been carried out [EM on functional or numerical data, etc.; see Vrac (2002)] and have led to the conclusion that the mixture decomposition of copulas gives more precise and relevant clusters.

7. Conclusions

We sought to develop a statistical approach to cluster a large set of atmospheric vertical profiles. This has been performed by preprocessing the data of temperature and humidity and converting these vertical profiles into distribution functions. Then, we developed adapted models (DFD) and mixture models for such data. Moreover, in order to take into account dependencies between variables and/or inside vertical profiles, we combined these models with copula functions, linking joint and marginal distributions. The associated

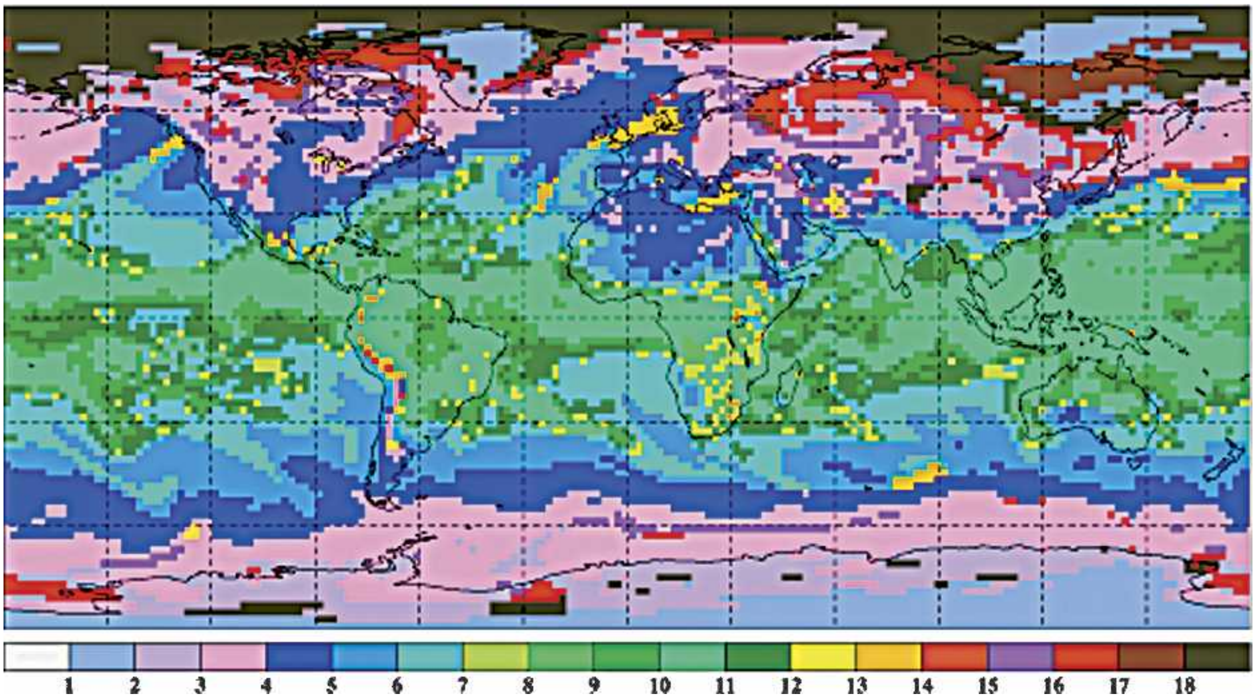

FIG. 9. Result of clustering into 18 clusters on temperature and humidity cdfs by coupling.

mixture model can be resolved by an appropriate EM-type algorithm. This original approach for mixture distributions generalizes the classical approach with the use of a higher abstraction level (i.e., dealing with cdfs instead of raw values). The results on thermodynamical cdf data seem to give realistic and relevant clusters concerning the climatology and are encouraging for the next variables. For a better evaluation of the obtained clusters, this work has to be applied to data from the radiative world, such as data given by the Infrared Atmospheric Sounding Interferometer (IASI) or by the Atmospheric Infrared Sounder (AIRS). This future

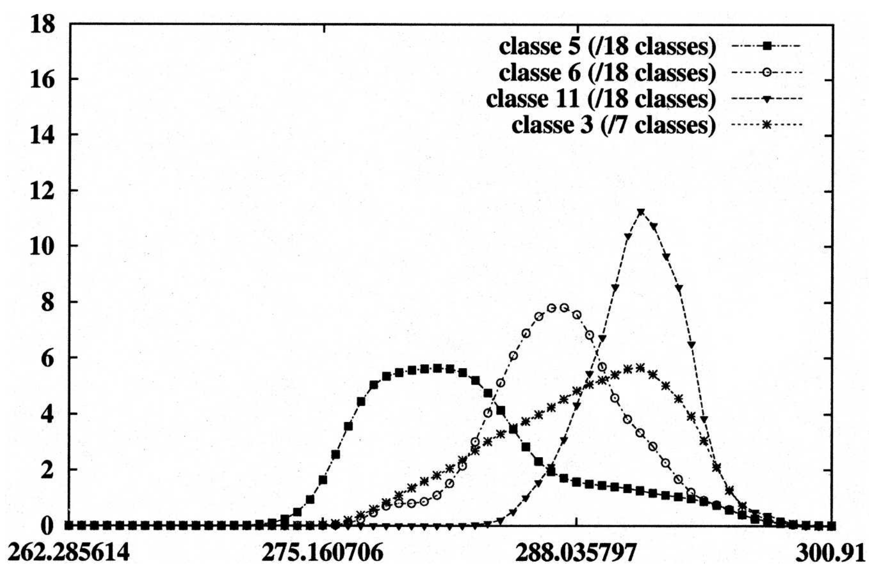

FIG. 10. Comparison of temperature densities for cluster 3 from the partition into seven clusters, and clusters 5, 6, and 11 from the partition into 18 clusters at 900 hPa.

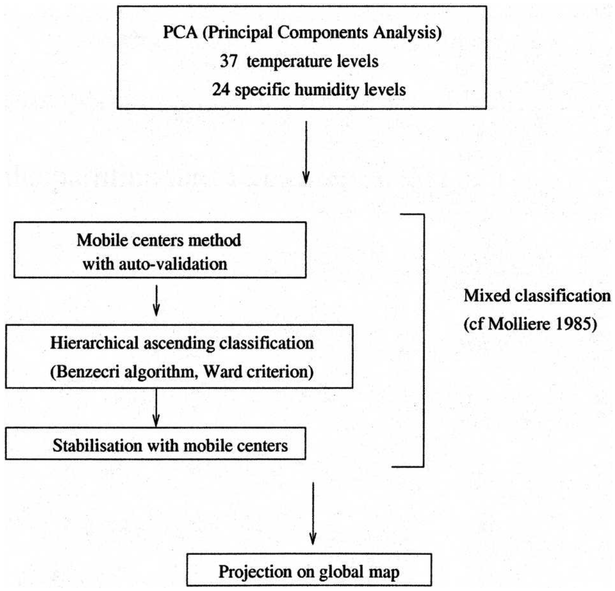

FIG. 11. Main steps of mixed clustering.

work should provide an assessment of our clusters, for instance by examining the time to find a solution by the 3I algorithm and the accuracy of the obtained solution.

From a more technical point of view, the method applied in section 5 uses a clustering approach (see section 1), but estimation approaches have also been generalized in the same sense (with copulas and for distribution functions), see Vrac (2002) or Diday and Vrac (2005) for EM, SEM, and simulated annealing (SAEM) algorithms. The parameters of copulas have been estimated by maximizing the likelihood, but Kendall's τ or Spearman's ρ could be used (Nelsen 1998). The choice of the T_i can be delicate. The efforts to automatically get an optimal T_i were not successful yet but some new approaches have been considered (Vrac 2002; Diday and Vrac 2005). Moreover, the initial cdfs have been modeled with a nonparametric distribution, and the distribution function of distributions with parametric beta distributions, but classical mixture distributions can even be used to model cdfs and DFDs themselves.

However, modeling statistical distributions by taking into account the dependencies between variables and among vertical profiles (temperature and humidity) is of high interest in climatology and meteorology. Then, the kind of results given by the proposed method should be useful in any method using a priori information on airmass types for instance. Hence, information on the distribution of the temperature and the humidity and their dependencies can be introduced as a priori information in a method for retrieving atmo-

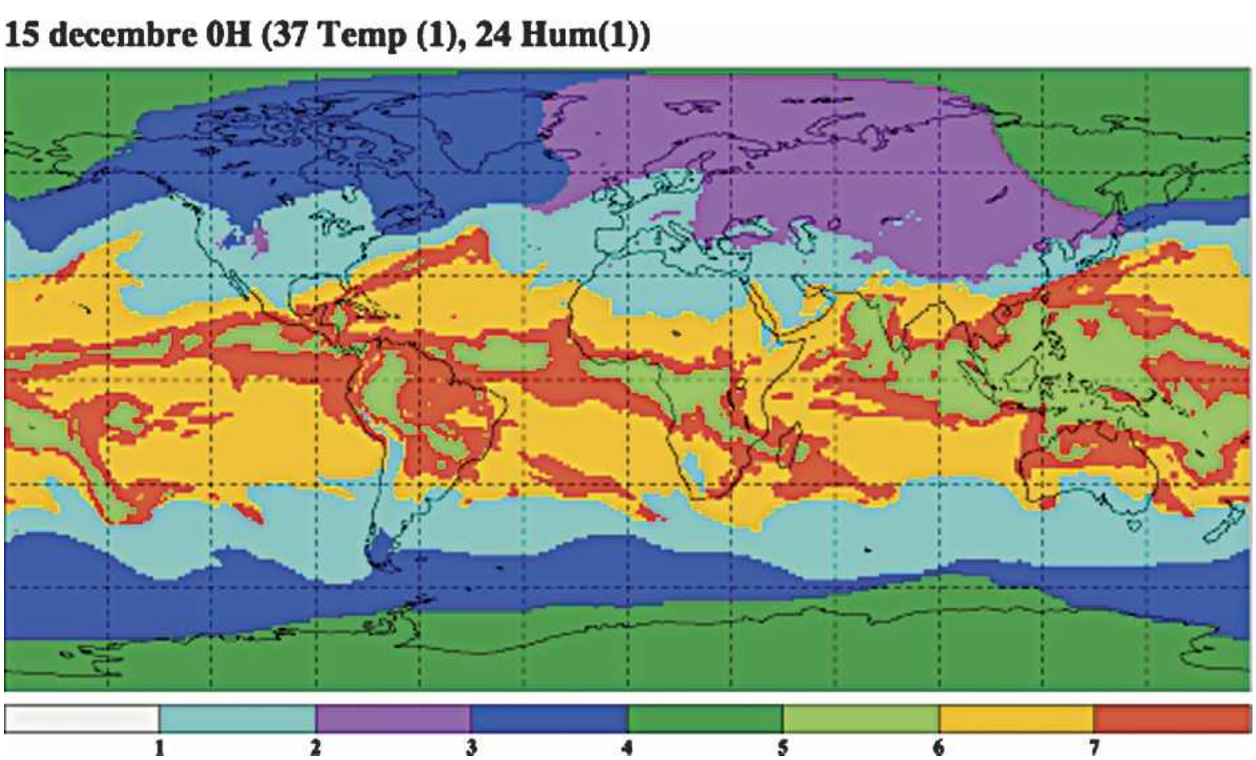

FIG. 12. Result of clustering into seven clusters on 37 temperatures at 24 specific humidities, 0000 UTC 15 Dec 1998.

spheric variables from satellite vertical sounder observations.

Acknowledgments. We thank R. Armante, N. A. Scott, and S. Serrar for their help and support through fruitful discussions. Moreover, we thank the two anonymous reviewers for their constructive comments.

REFERENCES

Achard, V., 1991: Trois problèmes clés de l'analyse 3D de la structure thermodynamique de l'atmosphère par satellite: Mesure du contenu en ozone; classification des masses d'air; modélisation hyper rapide du transfert radiatif. Ph.D. thesis, Université Paris VII.

Banfield, J., and A. Raftery, 1993: Model-based Gaussian and non-Gaussian clustering. *Biometrics,* **49,** 803–821.

Bock, H., and E. Diday, 2000: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer-Verlag, 425 pp.

Celeux, G., and G. Govaert, 1992: A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.,* **14,** 315–332.

——, and ——, 1993: Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Stat. Comput.,* **47,** 127–146.

Chédin, A., N. Scott, C. Wahiche, and P. Moulinier, 1985: The improved initialization inversion method: A high-resolution physical method for temperature retrievals from satellites of the TIROS-N series. *J. Climate Appl. Meteor.,* **24,** 128–143.

Davis, R., and D. Walker, 1992: An upper-air synoptic climatology of the western United States. *J. Climate,* **5,** 1449–1467.

Dempster, A., N. Laird, and D. Rubin, 1977: Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.,* **39,** 1–38.

Diday, E., and M. Vrac, 2005: Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Appl. Math.,* **147,** 27–41.

——, Y. Ok, and A. Schroeder, 1974: The dynamic clusters method in pattern recognition. *Proceedings of the IFIP Congress 74,* J. L. Rosenfeld, Ed., North-Holland.

Everitt, B., and D. Hand, 1981: *Finite Mixture Distributions.* Chapman and Hall, 143 pp.

Genest, C., 1987: Frank's family of bivariate distributions. *Biometrika,* **74,** 549–555.

——, and L. Rivest, 1993: Statistical inference procedures for bivariate Archimedian copulas. *J. Amer. Stat. Assoc.,* **88,** 1034–1043.

Hillali, Y., 1998: Analyse et modélisation des données probabilistes: Capacités et lois multidimensionnelles. Ph.D. thesis, Université Paris IX Dauphine.

Kalstein, L. S., J. S. Greene, M. C. Nichols, and C. D. Barthel, 1993: A new spatial synoptic climatological procedure. Preprints, *Eighth Conf. on Applied Climatology,* Anaheim, CA, Amer. Meteor. Soc., 169–173.

McLachlan, G., and D. Peel, 2000: *Finite Mixture Model.* Wiley, 419 pp.

Molliere, J., 1985: What is the real number of clusters? *Proc. Ninth Meeting of the German Classification Society*.

Nelsen, R. B., 1998: *An Introduction to Copulas.* Springer-Verlag, 216 pp.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 2002: *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1032 pp.

Ramsay, J., and B. Silverman, 1997: *Functional Data Analysis.* Springer, 434 pp.

——, and ——, 2002: *Applied Functional Data Analysis.* Springer, 190 pp.

Redner, R., and H. Walker, 1984: Mixture densities, maximum likelihood and the em algorithm. *SIAM Rev.,* **26,** 195–239.

Schweizer, B., and A. Sklar, 1983: *Probabilistic Metric Spaces.* Elsevier, 275 pp.

Scott, A., and M. Symons, 1971: Clustering methods based on likelihood ratio criteria. *Biometrics,* **27,** 387–397.

Scott, N. A., A. Chédin, R. Armante, and J. Francis, 1999: Characteristics of the TOVS pathfinder path-B dataset. *Bull. Amer. Meteor. Soc.,* **80,** 2679–2701.

Silverman, B., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Sklar, A., 1959: Fonction de répartition a n dimensions et leurs marges. *Inst. Stat. Univ. Paris Pub.,* **8,** 229–231.

Symons, M., 1981: Clustering criteria and multivariate normal mixtures. *Biometrics,* **37,** 35–43.

Vrac, M., 2002: Analyse et modélisation de données probabilistes par décomposition de mélange de copules et application à une base de données climatologiques. Ph.D. thesis, Université Paris IX Dauphine, 232 pp.