Adv. Stat. Clim. Meteorol. Oceanogr., 2, 115–136, 2016
www.adv-stat-clim-meteorol-oceanogr.net/2/115/2016/
doi:10.5194/ascmo-2-115-2016

ASCMO
Open Access

Mixture model-based atmospheric air mass classification: a probabilistic view of thermodynamic profiles

Jérôme Pernin[1], Mathieu Vrac[2], Cyril Crevoisier[1], and Alain Chédin[1]

[1]LMD, IPSL, CNRS, Ecole polytechnique, Université Paris-Saclay, 91128 Palaiseau, France
[2]LSCE, IPSL, CEA, CNRS, UVSQ, Centre d'Études de Saclay, Orme des Merisiers, Bât. 701,
91191 Gif-sur-Yvette, France

Correspondence to: Jérôme Pernin (jerome.pernin@lmd.polytechnique.fr)

Received: 15 February 2016 – Revised: 30 August 2016 – Accepted: 16 September 2016 – Published: 12 October 2016

Abstract. Air mass classification has become an important area in synoptic climatology, simplifying the complexity of the atmosphere by dividing the atmosphere into discrete similar thermodynamic patterns. However, the constant growth of atmospheric databases in both size and complexity implies the need to develop new adaptive classifications. Here, we propose a robust unsupervised and supervised classification methodology of a large thermodynamic dataset, on a global scale and over several years, into discrete air mass groups homogeneous in both temperature and humidity that also provides underlying probability laws. Temperature and humidity at different pressure levels are aggregated into a set of cumulative distribution function (CDF) values instead of classical ones. The method is based on a Gaussian mixture model and uses the expectation–maximization (EM) algorithm to estimate the parameters of the mixture. Spatially gridded thermodynamic profiles come from ECMWF reanalyses spanning the period 2000–2009. Different aspects are investigated, such as the sensitivity of the classification process to both temporal and spatial samplings of the training dataset. Comparisons of the classifications made either by the EM algorithm or by the widely used *k*-means algorithm show that the former can be viewed as a generalization of the latter. Moreover, the EM algorithm delivers, for each observation, the probabilities of belonging to each class, as well as the associated uncertainty. Finally, a decision tree is proposed as a tool for interpreting the different classes, highlighting the relative importance of temperature and humidity in the classification process.

1 Introduction

Contemporary synoptic climatology can be seen as a methodological perspective of climatology that creates and/or uses a classification of atmospheric variables at nearly any spatial or temporal scale to either simplify the climate system into a manageable set of discrete states or to gain a better understanding of how atmospheric variability impacts any climate-related outcome (Lee and Sheridan, 2015). A good overview of synoptic climatology as well as examples of studies can be found in Yarnal (1993), Yarnal et al. (2001), Barry and Perry (2001), Huth et al. (2008), Philipp et al. (2010) or Sheridan and Lee (2013).

Among the various classifications, air masses classically refer to large volumes of air which are fairly horizontally uniform with respect to temperature and humidity at any given altitude. Their thermodynamic features are related to the condition of the sea, land, or ice beneath it (Crowe, 1971). Such a definition varies somewhat from the traditional description of Bergeron (1930), which defines continental/maritime, polar/tropical (cP, cT, mP, mT) air masses according to the surface properties of their source regions (Bergeron, 1930; Willett, 1933). Therefore, air masses are characterized essentially by their thermodynamic character through various temperature and humidity variables (Kalkstein et al., 1996) defined at several lower- and mid-tropospheric pressure levels, possibly also including surface variables or even vertical profiles (Huth et al., 2008). However, additional variables, such as dynamic ones (e.g. sea level pressure, wind field), could be added to characterize missing dynamic behaviours, as in Živkovíć (1995).

This paper focuses on the classification of a large dataset of "atmospheric columns" homogeneous in both temperature and humidity on a global spatial grid and at different times. Such entities are closely related to the notion of air mass, and as such, will be mentioned by this term from now on. They are of great importance, in particular in inverse problems where climate variables as well as the three-dimensional structure of the atmosphere can be estimated from satellite data via inverse radiative transfer models (e.g. Chédin and Scott, 1985; Scott et al., 1999). Such models need to be initialized by a priori information as thermodynamic profiles and surface variables. For that purpose, real profiles and variables are usually considered, in particular those coming from radiosonde reports, as the thermodynamic Initial Guess Retrieval (TIGR) dataset (Chédin et al., 1985; Chevallier et al., 2000). With the technological advance of satellite instruments, such databases need to be enhanced in both sampling and air mass classification. Furthermore, adding prior information directly to the data leads to Bayesian statistics, in which the knowledge of the distribution of each statistic variable is essential, hence a probability point of view adopted here.

Over the past year, use of symbolic data has gained popularity (Bock and Diday, 2000; Diday, 2001; Floriana and Diday., 2003; Billard and Diday, 2012), since, as principal component analysis (PCA), they turn large databases into summarized data with manageable size while keeping useful information through a process commonly called "data aggregation" or "data compression". This process is often necessary as a step of pre-processing before any classification procedure. Symbolic data change the way in which the description of the data is viewed, since they refer to data which do not only contain values as in classical data, but also have a structure and include internal variations. This is the case with distributions, considered by Schweizer (1984) as the "the numbers of the future", such as probability density functions (PDFs) or cumulative distribution functions (CDFs).

In the present work, CDFs are used in combination with a probabilistic classification method based on a Gaussian mixture model that has been used for example in Vrac et al. (2007), Rust et al. (2010) or Carreau and Vrac (2011). Such a model relies on the assumption that observations from a given dataset come from several sub-populations and that the overall population can therefore be modelled as a Gaussian finite mixture model, or in other words a mixture of weighted PDFs, each one corresponding to a given sub-population (Fraley and Raftery, 2002). The main problem consists then in estimating the parameters of the mixture so that the model best fits the data, which can be done through two different approaches: (i) the "estimation approach" focuses on the estimation of the mixture model parameters usually using maximum likelihood estimation techniques. The most efficient algorithms rely on the iterative expectation–maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008). An optional partition can then be obtained by applying the maximum a posteriori principle; (ii) the "clustering approach" focuses directly on grouping the entities for classification into a number of clusters such that each cluster can be seen as a sub-population with a given PDF. In that case, the algorithms generally used rely on the so-called dynamical clustering algorithms (Diday et al., 1974), applicable to a multivariate Gaussian mixture (Symons, 1981; Celeux et al., 1989). These algorithms were also combined with the EM algorithm to develop a variant of the latter, the classification EM (CEM) algorithm (Celeux and Govaert, 1992). Another clustering approach is the widely used k-means algorithm (e.g. Huth, 2001). It is an iterative relocation algorithm which iteratively calculates the centre of each cluster (initially randomly chosen) and allocates the data to the cluster whose centre is closest. Other examples include hierarchical classification (e.g. Kalkstein et al., 1987; Vrac et al., 2007), neural networks and more particularly self-organizing maps (SOM) (e.g. Hewitson and Crane, 2002; Reusch et al., 2007), or the mixed clustering strategy of Molliere (1985) as in Vrac (2002), among others. An extended overview of classification can be found in Gordon (1999) or Hastie et al. (2009).

In Vrac (2002) and Vrac et al. (2005, 2011), a mixture of copulas was proposed, providing not only a partition of the atmosphere into air mass classes, but also a probabilistic model describing classes as well as the dependencies between and among temperature and humidity through the so-called copula functions (Schweizer and Sklar, 1983; Nelsen, 1999; Diday and Vrac, 2005). The thermodynamic vertical profiles characterizing each atmospheric situation were first aggregated into four CDF values (two for both temperature and specific humidity). The classification method applied to these four statistical variables was then an extension of the problem of the mixture model to these distribution-valued data, so that multidimensional copulas could be used. The algorithm chosen for this purpose, initialized by a partition based on seven zonal clusters, was a dynamical clustering method (clustering approach). As a first step, the results were validated only for a limited dataset of 1 winter day and short-range projections not exceeding 1 month.

Setting aside copula, this paper aims at consolidating the results of Vrac (2002) and Vrac et al. (2005, 2011) by proposing a robust air mass classification method based on a Gaussian mixture model and the EM algorithm (estimation approach). This method is able to deal with a much larger dataset covering a decade and providing larger-range projections without having to use arbitrary a priori information as an initialization strategy in the absence of a prior reference atmospheric column type classification. The larger spatio-temporal coverage of the dataset considered here enabled one to take into consideration more statistical variables for a bet-

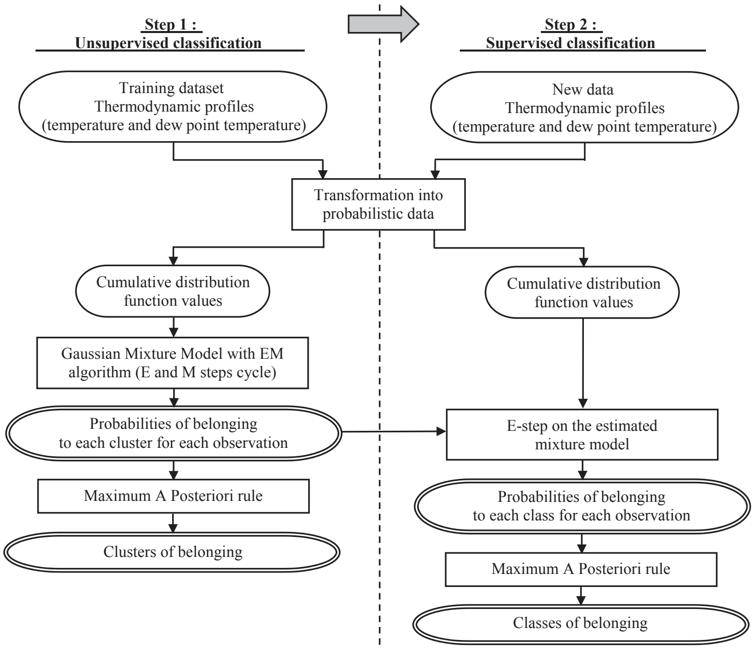
Step 1 :
Unsupervised classification
Training dataset
Thermodynamic profiles
(temperature and dew point temperature)
Transformation into probabilistic data
Cumulative distribution function values
Gaussian Mixture Model with EM algorithm (E and M steps cycle)
Probabilities of belonging to each cluster for each observation
Maximum A Posteriori rule
Clusters of belonging
Step 2 :
Supervised classification
New data
Thermodynamic profiles
(temperature and dew point temperature)
Cumulative distribution function values
E-step on the estimated mixture model
Probabilities of belonging to each class for each observation
Maximum A Posteriori rule
Classes of belonging

Figure 1. Principle of the methodology used in this paper.

ter description of the troposphere while studying more thoroughly different aspects that had not been taken into account previously, such as sensitivity to both temporal and spatial samplings of the training dataset, consistency with the number of clusters or comparison with a k-means algorithm. A decision tree is also proposed as a tool for interpreting the different classes, highlighting the relative importance of temperature and humidity in the classification process.

The article is organized as follows. Section 2 presents the data to classify, the pre-processing data compression step, the classification method and the choice of the optimal configuration of the model. Section 3 discusses different aspects regarding the quality of the classifications. Section 4 focuses on the interest of the a posteriori probabilities provided by the EM algorithm, and introduces an interpretation of the classes based on a decision tree. Section 5 concludes with a discussion.

2 Data and methods

Figure 1 outlines the methodology used in this paper. The goal is first to build air mass clusters from atmospheric situations characterized by temperature and humidity probabilistic data only, along with probabilities of the situations belonging to each cluster. The clusters have to be as coherent as possible in terms of temperature and humidity (e.g. hot and wet, or cold and dry air masses) and as different as possible from each other. To achieve this, a representative training dataset of thermodynamic profiles is first turned into a manageable set of CDF values. The latter are used as input into the EM algorithm to obtain a partition of the atmospheric situations as well as a probabilistic description of each group (cluster) of the partition, without having any prior reference classification (a process called unsupervised classification, clustering or cluster analysis). The information contained in this probabilistic description can then be used in a second step to identify to which of the existing groups (classes) any new atmospheric situation belongs, on the basis of the training dataset whose group assignment is already known (a process called supervised classification or simply classification when explicitly compared to clustering).

2.1 Thermodynamic description of the atmosphere

The atmospheric dataset used in this study is based on ERA-Interim global atmospheric reanalyses from the European Centre for Medium-Range Weather Forecasts (ECMWF) covering the period from 1 January 2000 to 31 December 2009 (Dee et al., 2011). The different daily products (e.g. surface pressure, surface temperature, temperature and specific humidity profiles) are available on a 0.75° by 0.75° latitude–longitude global grid (241 × 480 data points) every 6 h for 4 synoptic hours, i.e. 00:00, 06:00, 12:00 and 18:00 (Coordinated Universal Time). Vertical profiles are described on 60 pressure (or altitude) coordinates called "sigma levels" from the surface to 0.1 hPa (about 65 km in altitude on sea).

Unless explicitly mentioned, the set of observed data corresponding to 00:00 and 12:00 UTC of the 15th day of January, April, July and October from 2005 to 2009 will be used in this article and referred to as "training dataset". Each synoptic hour gathers $241 \times 480 = 115\,680$ atmospheric situations spatially distributed over the Earth corresponding to

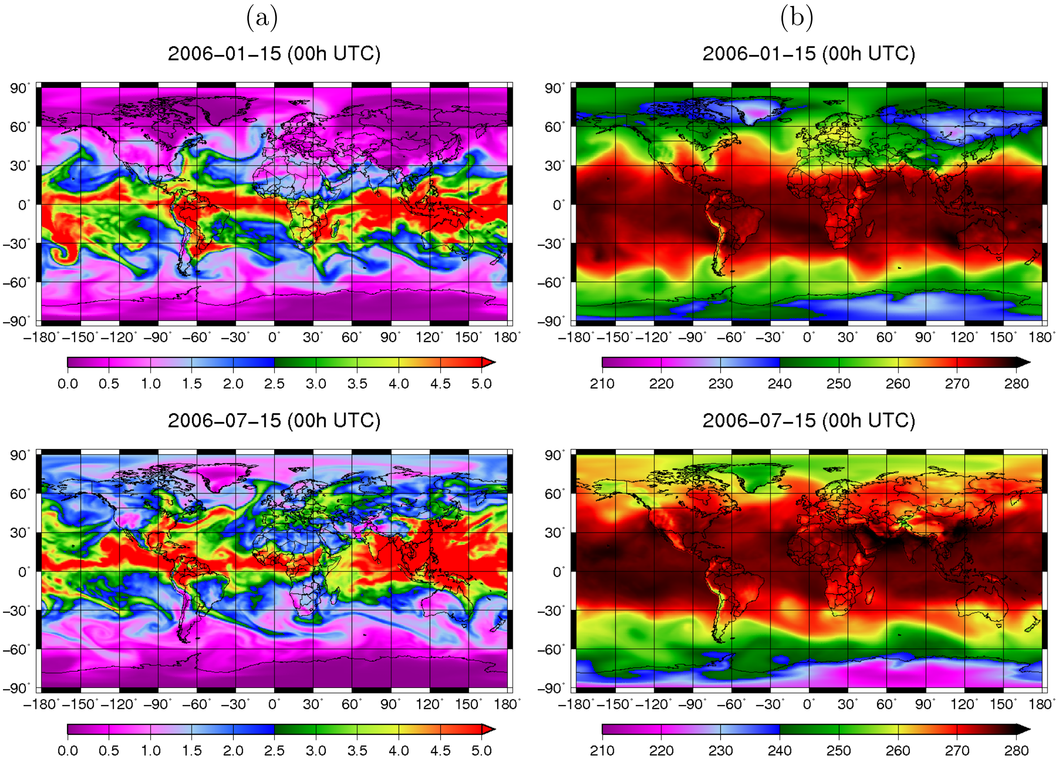

Figure 2. Total column water vapour in precipitable centimetres **(a)**, and average temperature in the 47–37 sigma-level layer (800–425 hPa on the sea) in K **(b)**.

different local hours. This implies that the previous dataset contains 115 680 × 40 = 4 627 200 atmospheric situations. However, keeping all these situations may not be necessary to get the intended atmospheric partition. The influence of the temporal and spatial sampling of the dataset on the classification will be studied in Sect. 3.2.

Humidity can be measured by various variables. Here, specific humidity, dew point temperature as well as total column water vapour will be considered. They correspond to the ratio of the mass of water vapour to the mass of dry air plus water vapour (expressed in kg kg^{-1}), the temperature at which air is saturated with water vapour (in K) and the amount of vertical integrated water vapour in the whole atmospheric column (expressed here as the height of an equivalent column of liquid water in precipitable centimetres) respectively.

In order to have data homogeneous to temperature, specific humidity from ERA-Interim reanalyses has been converted into dew point temperature, whose relation can be deduced from Buck (1981), so that each atmospheric situation is characterized by temperature and dew point temperature profiles only. The first 38 temperature values (from the surface to 67 hPa) and the first 30 dew point temperature values (from the surface to 230 hPa) from each profile are kept in order to feature the thermodynamic properties of the troposphere. Figure 2 illustrates such thermodynamic variables by representing the total column water vapour and average temperature between the 47th and 37th sigma levels (800 and 425 hPa on sea respectively). The atmospheric situations associated with elevation above sea level higher than 1 km have been discarded, which corresponds here to 13 % of the situations. This pre-filtering puts aside the question of whether the atmospheric situations corresponding to high elevations require a specific handling.

2.2 From numerical data to cumulative distribution function data

The approach proposed here consists in working with CDFs instead of using classical numerical values of temperature and dew point temperature at different pressure levels. A cumulative distribution function F of a real-valued random variable X is defined as the probability that X will take a value lower than or equal to a value x:

$$F_X(x) = P(X \leq x). \tag{1}$$

Here, X refers to either temperature or dew point temperature. Thus, for example, the previous definition means that CDF value $F_\mathrm{T}\,(290\,\mathrm{K}) = \mathrm{P}\,(T \leq 290\,\mathrm{K})$ gathers information from all the temperature values at different pressure levels which are lower than 290 K.

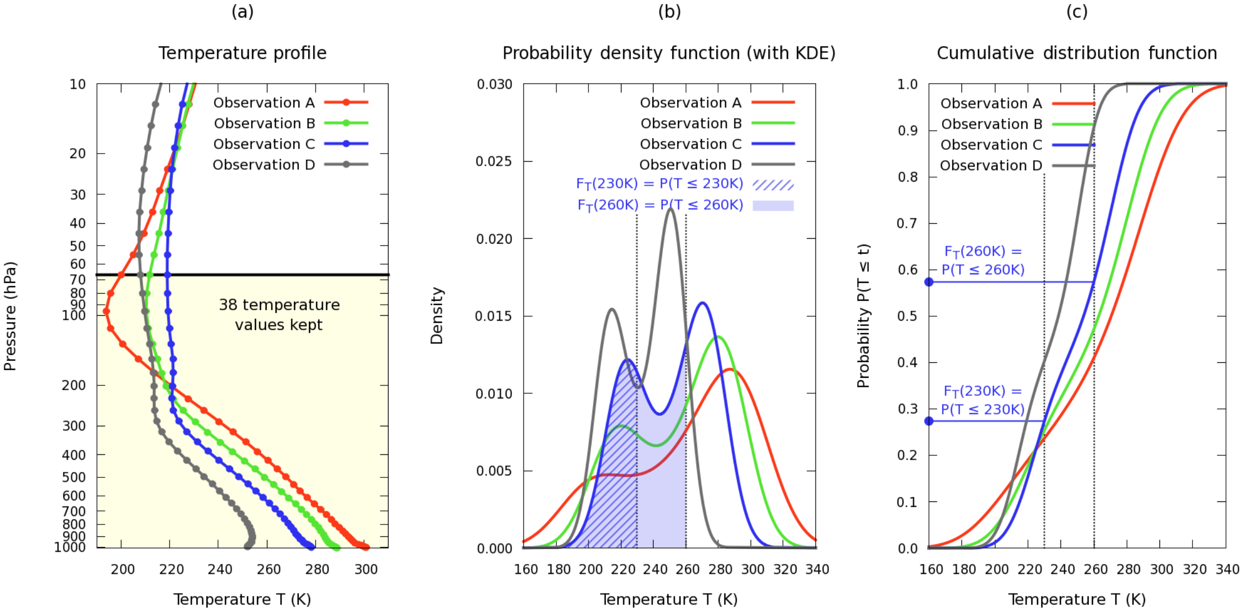

Figure 3. From numerical data to CDF values. The parts of the thermodynamic profiles (temperature ones here) of given atmospheric situations (four typical examples here, denoted by A, B, C and D), shown in **(a)**, are turned via KDE into PDFs **(b)** from which CDFs are obtained **(c)**. A discrete set of temperature and dew point temperature CDF values $F_\mathrm{T}(t) = P(T \leq t)$ and $F_\mathrm{Tdp}(t_\mathrm{dp}) = P(T_\mathrm{dp} \leq t_\mathrm{dp})$ are then chosen and used as input to the classification algorithm (like F_T (230 K) and F_T (260 K) in the plots).

Figure 3 illustrates the transformation process from thermodynamic profiles to CDFs for four example temperature profiles (the method remains the same with dew point temperature profiles). The selected part of the profiles is converted into a PDF first and then into a CDF from which a discrete set of key CDF values is selected as being the most representative information of the temperature (or dew point temperature) vertical profiles (data aggregation step). Here, since the general shape of temperature and humidity distributions is not known a priori, the conversion into PDF is performed using the non-parametric Rosenblatt–Parzen kernel density estimation (KDE) method (Rosenblatt, 1956; Parzen, 1962), which can be seen as a weighted sum of p parametric normal distributions centered at each value of the sample (e.g. the $p = 38$ values of interest of a given temperature profile here). More details can be found in Vrac et al. (2005, 2011).

A priori knowledge is often essential to choose the relevant variables (here, the CDF values) able to distinguish the intrinsic groups in a given dataset. Such a problem is known as “feature selection” or “variable selection”. Some numerical techniques exist in the literature to help make a choice (e.g. Diday and Vrac, 2005; Vrac et al., 2011; Pudil et al., 1994; Raftery and Dean, 2006), but they remain not really relevant for the datasets studied here. Looking at the diversity of the profiles leads one to select subjectively CDF values at several temperature values from 200 to 290 K, and the same for dew point temperature. A set of 10 values within this interval, every 10 K, seems a good compromise between keeping enough information and reducing as much as possible the number of dimensions.

Finally, the database used as input data into the clustering algorithms (EM or k-means) is an array of size $N \times D$, where N is the number of selected atmospheric situations (e.g. 4 627 200 observations for the training dataset presented in Sect. 2.1) and D is the dimensionality or number of variables characterizing each situation (10 CDF values for temperature and 10 CDF values for dew point temperature).

2.3 Clustering and classification using the Gaussian mixture model

2.3.1 The Gaussian mixture model and the EM algorithm

Model-based clustering relies on the assumption that a given observed dataset contains several sub-populations and that the overall population can therefore be modelled as a finite mixture model. Usually, and as done here, a Gaussian mixture model is used, that is, a mixture of weighted Gaussian PDFs, each one associated with a given cluster related to a sub-population. Let $\boldsymbol{x} = \{\ \boldsymbol{x}_i \in \mathrm{R}^D | i = 1, \ldots, N\}$ be a population of N observed data, each one represented by a D-dimensional vector $\boldsymbol{x}_i$, containing K sub-populations modelled by multivariate Gaussian PDFs parameterized by their mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k. With $\boldsymbol{\theta}_k = (\theta_k^1, \ldots, \theta_k^D) = (\ \boldsymbol{\mu}_k, \Sigma_k)$, the PDF g of $\boldsymbol{x}$ takes the following form:

$$g(\boldsymbol{x}\,|\boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k), \tag{2}$$

where $\boldsymbol{\phi} = (\pi_1,\ldots, \pi_K, \boldsymbol{\theta}_1,\ldots, \boldsymbol{\theta}_K)$ is the collection of the mixture model parameters and $\pi_1,\ldots, \pi_K$ are positive mixture proportions summing to one, corresponding to the a priori probabilities that $\boldsymbol{x}_i$ will belong to each cluster.

The main problem consists then in estimating the mixture model parameters $\boldsymbol{\phi}$, so that the density function g associated with the mixture best fits the data. Here, these parameters are estimated through maximum likelihood estimation (MLE) by using the widely applied expectation–maximization (EM) iterative algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008).

The EM algorithm alternates between the E-step and the M-step. At each iteration, the parameters $\boldsymbol{\phi}$ are updated while the log-likelihood monotonically increases (Hastie et al., 2009). The E-step computes at iteration l the a posteriori conditional probability t_{ik} that observation i belongs to cluster k:

$$t_{ik}^{(l)} = \frac{\pi_k^{(l-1)} f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^{(l-1)})}{\sum_{k'=1}^{K} \pi_{k'}^{(l-1)} f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{k'}^{(l-1)})}. \tag{3}$$

The M-step estimates the mixture model parameters $\boldsymbol{\phi}$ given the t_{ik}. Starting from some initial values of the parameters $\boldsymbol{\phi}$, the algorithm repeats the E-step and M-step cycle until it converges towards a local maximum of the log-likelihood. Then, the a posteriori probabilities t_{ik} can be used to generate a partition $P = (P_1,\ldots, P_K)$ of the $i = 1,\ldots, N$ observations by assigning each of them to the cluster k providing the highest probability t_{ik} among the K clusters. This assignment process corresponds in fact to the maximum a posteriori (MAP) estimation method, based on Bayes' theorem (Bayes and Price, 1763) and the law of total probability. From the parameters $\boldsymbol{\phi}$ obtained at the end of the unsupervised classification process, supervised classification can then be performed via a single E-step followed by the MAP principle. To implement EM, the Rmixmod (Mixmod Team, 2008; Lebret et al., 2015) S4 package has been used.

2.3.2 Initialization strategy

In order not to use arbitrary a priori information, the initialization strategy used in this paper consists in repeating r (here, $r = 50$) random draws of the component means in the dataset with short runs of EM, that is, until the number of iterations has reached a pre-defined maximum number of iterations (here, 10) if convergence has not been reached before. The resulting mixture model parameters corresponding to the random partition that maximizes the log-likelihood among the r runs are then provided as initial parameters to the normal EM procedure. This method, which was introduced by Biernacki et al. (2003), is often considered as a reference strategy since it is a trade-off between avoiding suboptimal solutions, often obtained with a single run of EM with a random initialization partition, and reducing the number of iterations to reach convergence.

2.3.3 Choice of the number of clusters and covariance matrix models

As in Banfield and Raftery (1993) and Celeux and Govaert (1995), each covariance matrix Σ_k related to the kth cluster is expressed in terms of its eigenvalue decomposition, leading to several parsimonious models (actually, 14). These models can be simply indicated by three sequential letters corresponding to three attributes characterizing the dispersion of the mixture component distributions, that is, the hypervolume, the shape and the orientation of their isocontour in the multidimensional space. Further details can be found in Appendix A and the references mentioned above.

Among the three covariance matrix models which do not lead here to poor structures in terms of air mass patterns (too much zonal structure or a preponderance of one cluster over the other ones), only two models, known as hyperspherical models, will be studied in this paper: the model denoted VII assumes isotropic dispersion which can vary between the clusters, whereas EII differs only from the previous model by constraining the dispersion to be equal between the clusters. The expressions of the Σ_k eigenvalue decomposition for these two models are respectively the diagonal matrices $\lambda_k I$ and λI, where I is the identity matrix, λ is a scalar and the presence of the subscript k implies that λ can vary between the clusters. For illustrative purposes, Fig. 4 shows the projection in a two-dimensional subspace of the result of an unsupervised classification into seven clusters with EM-VII (EM and model VII) applied to the $D = 20$ CDF values corresponding to 10 % of the observations coming from the training dataset (for 2005, 00:00 UTC only). The central figure shows the F_T (230 K) and F_T (260 K) CDF values for the N atmospheric situations as well as their corresponding cluster, where the four representative temperature profiles of Fig. 3 are also represented. Here, the projected hyperspherical isocontours are reduced to circles whose radius equals the standard deviation of the samples within each cluster. The two other plots show the two corresponding univariate densities either per cluster or for the mixture.

One of the most difficult problems in unsupervised classification is the determination of the optimal number of clusters (unless already known) which must be fixed before performing the clustering process. When the number of clusters is not apparent from prior knowledge, many methods have been established over the years to help make a suitable choice. Several criteria have been tested (e.g. Akaike, 1973; Schwarz, 1978; Raftery, 1995; Hardy, 1996, 2006; Gordon, 1999; Biernacki et al., 2000), including the approximate weight of evidence (AWE) adopted in Vrac et al. (2005,

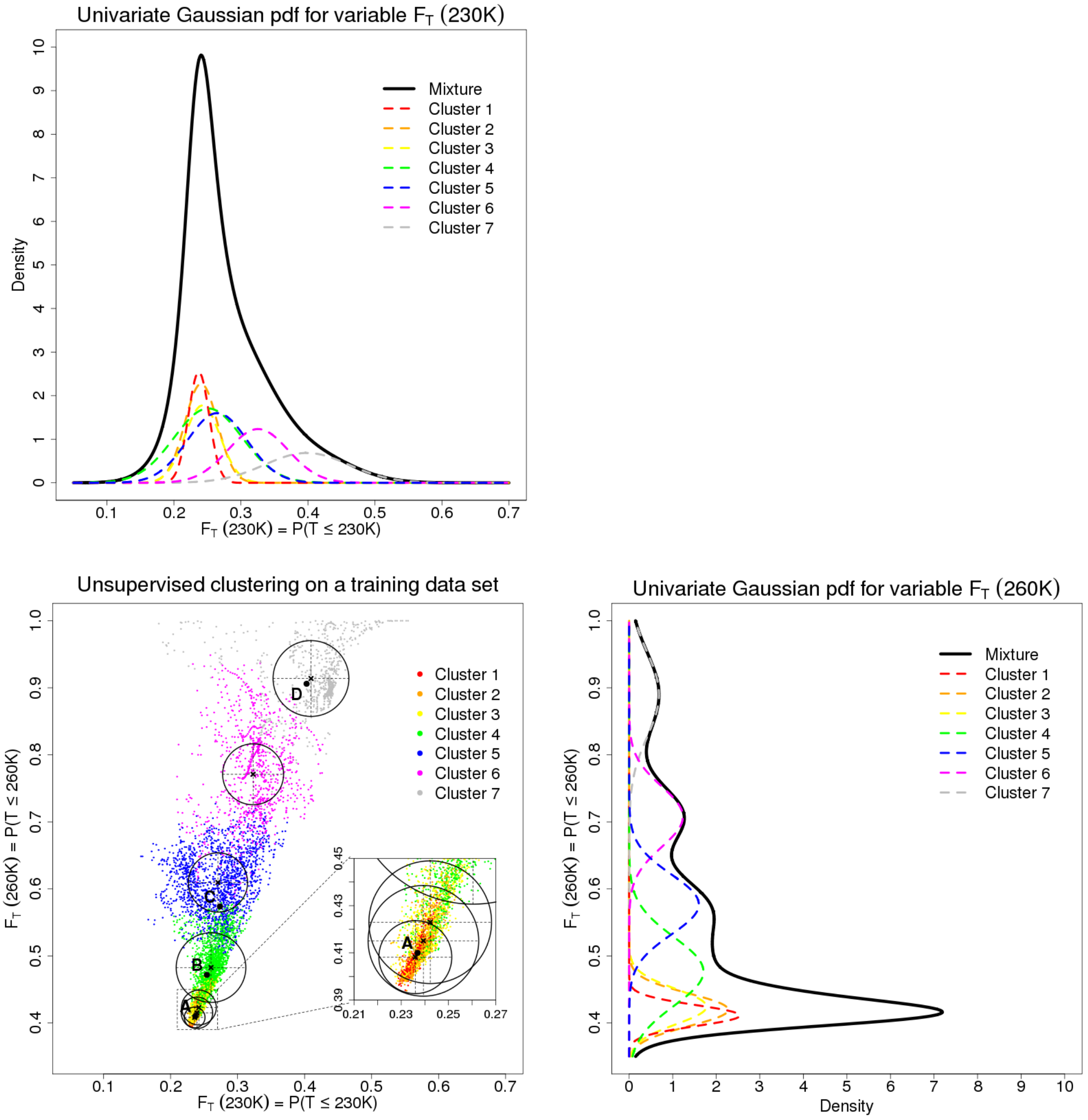

Figure 4. Result of a 20-dimensional multivariate seven-cluster unsupervised classification process via EM-VII (isotropic dispersion which can vary between the clusters) projected in a 2-dimensional subspace, that is, for the variables F_T (230 K) and F_T (260 K) here. The radius of each circle indicates the standard deviation of the cluster where each of the atmospheric situations is represented by a point (the four situations of Fig. 3 are also indicated). The adjacent plots represent, for each of the two variables chosen for the projection, the associated univariate densities per cluster (the full width at half maximum corresponds to the diameter of the associated circle) or for the mixture.

2011). Some of them are based on maximizing the log-likelihood to which a penalization term is added, depending on the number of independent parameters to estimate for the model selected (the covariance matrix model, here), but all of them lead to similar results. In the end, all the criteria are not able to distinguish which covariance matrix model or number of clusters suit the present data, so that their choice remains suggestive here.

Following Vrac et al. (2005, 2011), the number of clusters is fixed to seven. This choice was motivated by the fact that an odd number of clusters is a priori expected to take into account the natural difference in the mid-latitude and polar air masses between summer and winter (hence, at least four clusters) while favouring a kind of symmetry around the Equator with more than one air mass for the tropics (hence, at least three additional classes).

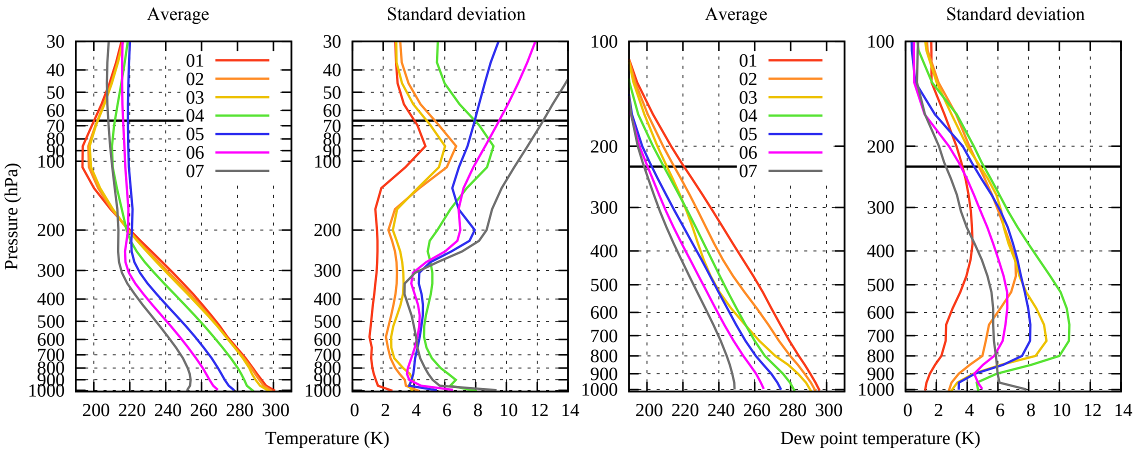

Figure 5. Average and associated standard deviation profiles for temperature (left) and dew point temperature (right) for each cluster obtained with EM-VII. The black line indicates the upper pressure level kept used for computing the cumulative distribution functions.

Table 1. Main features of the air mass clusters. The percentages have been computed after discarding the high relief atmospheric situations, so that each sum of a given row equals 100 %.

Features	1	2	3	4	5	6	7
EM-VII clustering							
Total column water vapour (cm)	5.32	3.93	2.66	1.81	1.04	0.52	0.22
800–320 hPa average temperature (K)	267.2	265.9	265.1	258.1	248.0	240.1	234.2
Average surface temperature (K)	300.7	299.2	297.9	289.3	278.1	268.8	251.0
Atmospheric situations %	9.8	13.7	11.8	22.1	17.9	13.9	10.8
EM-EII ($\approx k$-means) clustering							
Total column water vapour (cm)	4.83	2.89	1.49	1.56	0.78	0.38	0.17
800–320 hPa average temperature (K)	266.8	264.0	260.6	252.0	244.5	237.5	232.9
Average surface temperature (K)	300.1	296.9	293.3	280.7	275.1	262.5	246.4
Atmospheric situations %	17.9	20.6	11.9	15.0	16.6	11.7	6.3

3 Clustering and classification of atmospheric situations

3.1 Unsupervised classification: comparison between EM-VII and EM-EII

3.1.1 Partition and cluster features

In this section, unsupervised classification is applied to the training dataset with no spatial sampling. The situations associated with elevations above sea level higher than 1 km will be referred to as air mass 0 from now on to indicate that they have been discarded (Sect. 2.1). The seven resulting clusters are ordered from the average hottest surface temperature to the average coldest one, that is, globally from a tropical air mass (1) to a polar one (7), and can be thermodynamically correlated with the maps shown in Fig. 2. The features of each cluster can be represented for example by their mean and standard deviation profiles for temperature and dew point temperature shown in Fig. 5, and also by the total column water vapour mean and mid-tropospheric layer average temperature (here, 800–320 hPa) listed in Table 1. This table also contains the percentage of situations per cluster for the whole period on which the clustering has been performed, after discarding the high relief situations (Sect. 2.1). Figure 6 shows partitions resulting from EM-VII (a) and from EM-EII (b). These results are discussed in the following sections.

3.1.2 EM-VII clustering

The clusters shown in Fig. 6a present relevant thermodynamic homogeneous areas: three tropical/sub-tropical hot air masses which are distinguished essentially by humidity, that is, very wet (1), wet (2) and relatively wet (3) ones; one temperate air mass mixing warm to cool, relatively wet to dry atmospheric situations (4); and three sub-polar/polar air masses corresponding respectively to a relatively cold and dry air mass including northern summer situations (5), a cold and dry one (6) and finally a winter frigid, very dry one (7).

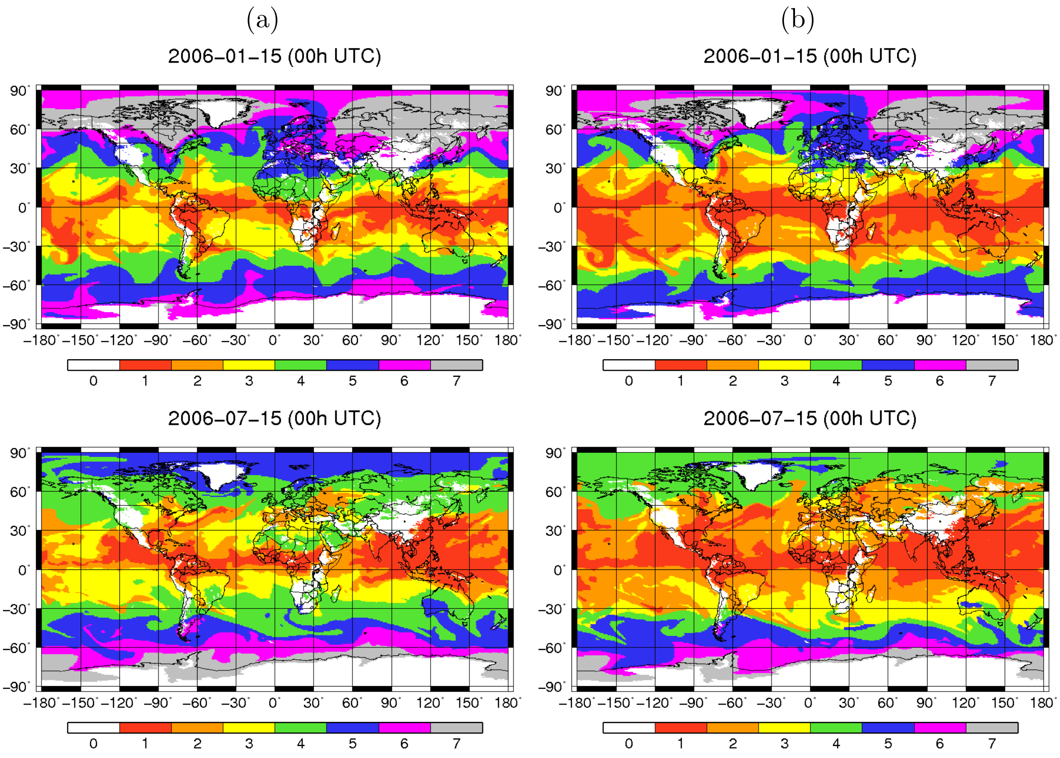

Figure 6. Seven-cluster unsupervised classification with the EM-VII model **(a)** and EM-EII (≈ k-means) **(b)**, for 2 months.

As confirmed by Fig. 5, polar air masses are characterized, as expected, by a higher variability in temperature and humidity, while air mass 4 acts as a transition cluster between tropical/sub-tropical air masses and sub-polar/polar ones. As for air mass 3, it is associated with a strong dew point vertical gradient in the middle troposphere reflecting areas of dry air subsidence in both hemispheres between 20 and 35° in latitude depending on season, which correspond to the boundaries between the Hadley and Ferrel cells of the global atmospheric circulation scheme (Peixoto and Oort, 1996; Vergados et al., 2015).

Comparing these air mass maps to Fig. 2 shows some similarities, particularly regarding the distribution of the total column water vapour regardless of the amount of humidity. Tropical situations are precisely depicted, as shown for instance by humid incursion of air masses 1 and 2 into the drier air mass 3, spiralling clockwise towards the centre of a depression in the southern Pacific Ocean between -180 and $-150°$ W on 15 January 2006, 00:00 UTC. More generally, the partition so obtained is rather well correlated with synoptic weather phenomena. However, since no dynamic variables (e.g. wind speed, speed direction, potential vorticity) have been taken into account here, air pressure and wind can vary within these air masses, as stated by Kalkstein et al. (1996). Therefore, the shapes of synoptic meteorological phenomena, as depressions, are not always depicted continuously.

The hottest and wettest air mass cluster 1, particularly, follows closely the Intertropical Convergence Zone (ITCZ). The latter consists in hot, very wet air masses meeting together due to the trade winds, and involving very hot low tropospheric temperatures as well as convective systems consisting in large-scale thunderstorms when the surface is also wet (oceans, tropical forests). The slight seasonal shift of the ITCZ location is then visible, moving annually towards the northern Tropic of Cancer in northern summer and towards the southern Tropic of Capricorn in northern winter, since the belt of maximum temperatures migrates as the Earth orbits the Sun. This is also illustrated in Fig. 7, which shows the percentages of observations per cluster, or corresponding to high relief (white colour), for the whole band of latitude (top bar) or per 10° band of latitude (lower bars) for the 15th day of January/April/July/October 2006 at 00:00 UTC. The peak in latitude of the red bars corresponding to air mass 1 moves from about -5 °C in January to 10 °C in July. This figure also highlights the transition role of air mass 4 mentioned earlier, particularly visible in summer, as well as the differences in behaviour of the sub-polar/polar air masses through the different seasons. For instance, air mass 7 is mainly located north in northern winter and south in southern winter, reflecting the role of seasonal insolation, and is thus essentially associated with extremely cold and dry winter polar atmospheric situations. This air mass corresponds closely to the traditional winter continental polar (cP) air mass of Berg-

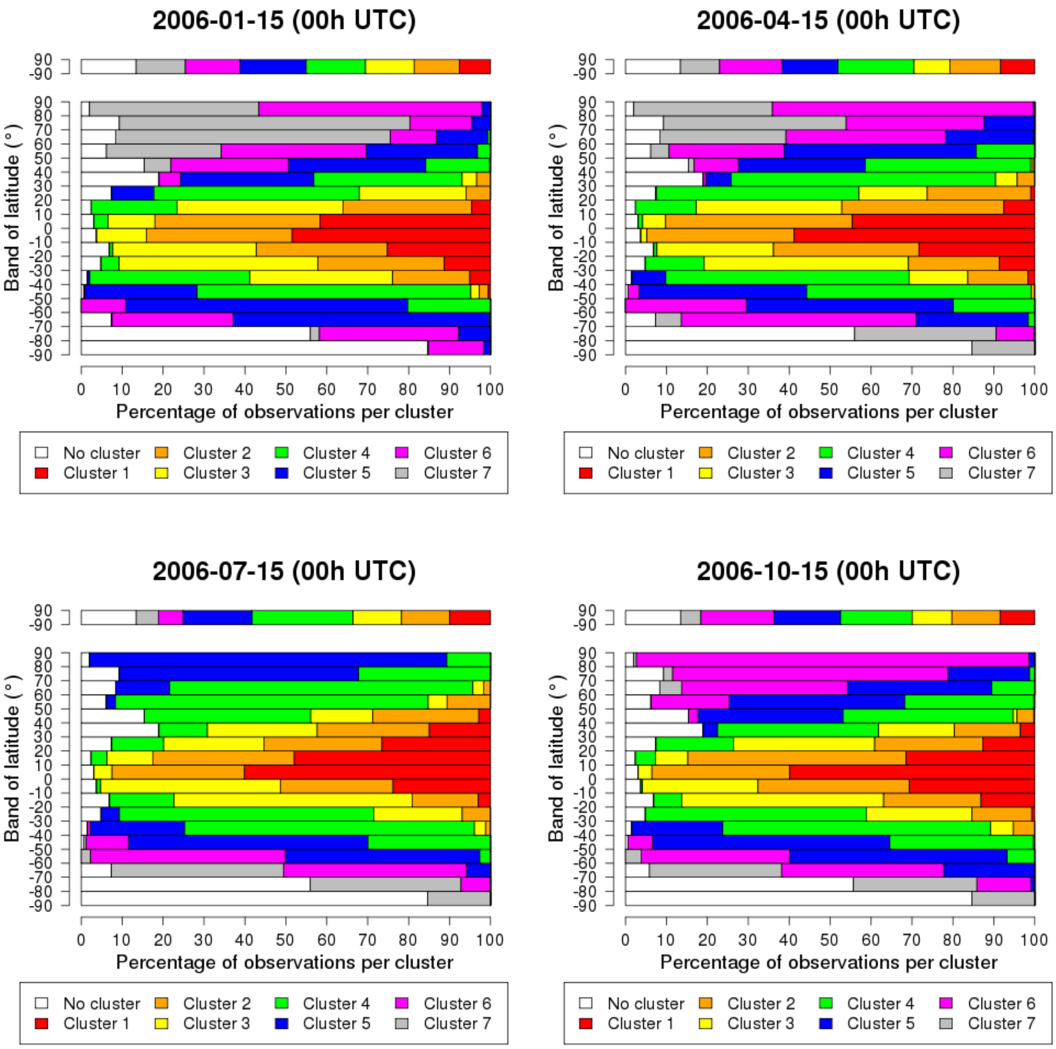

Figure 7. Percentage of observations per cluster on the whole band of latitude and per 10° band of latitude, obtained with EM-VII, for 4 months. The "No cluster" label corresponds to the high relief atmospheric situations.

eron (1930), which forms over large, high-latitude lands as source regions, like North America, Greenland or northern Asia (Siberia). These regions are usually snow-covered and characterized by short days and low solar angles, so that they reflect much of the solar radiation when it reaches the surface. The atmospheric situations are therefore extremely cold and dry, and are characterized by high lower tropospheric stability inhibiting vertical mixing, as illustrated by the temperature inversion near the surface in Fig. 5.

As in Vrac et al. (2005) the discrimination between the air masses as well as their features can also be illustrated by plotting the temperature (and dew point temperature) PDFs representing the distribution of the thermodynamic variable at a given sigma pressure level for each air mass cluster (not shown). This shows for example the behaviour of the first two tropical air masses seen from Figs. 5 and 6a (i.e. overlapping temperature PDFs corresponding to air masses 1 and 2, with well-distinguished dew point temperature PDFs) and the fact that the result of the clustering procedure is mainly due to the mid- and lower troposphere, and, to a lesser extent, to the tropopause, since discrimination between clusters decreases at higher altitudes. This explains the lower temperature variabilities in the lower and mid-troposphere and the higher temperature variabilities around the tropopause observed in Fig. 5.

3.1.3 EM-EII clustering

The second possible kind of classification leading to relevant air masses is obtained with EM-EII, whose partitions are nearly identical to those obtained with the widely used *k*-means algorithm (Lloyd, 1957; Forgy, 1965; MacQueen, 1967; Diday et al., 1974; Hartigan and Wong, 1979). This is not surprising since the classification variant of the EM algorithm (CEM) along with the EII model as well as equal mixture proportions are equivalent to the k-means algorithm: in that case, maximizing the complete log-likelihood (objective of CEM) involves minimizing the within-cluster sum of squares criterion (objective of k-means). Although CEM is substituted here by EM and the assumption about the equality of the mixture proportions is not verified here, the other assumption (equal isotropic dispersion between the clusters) is self-sufficient and proves that EM generalizes k-means. From now on, EM-EII will only be explicitly used, but it can be substituted by the k-means algorithm without changing the results.

As shown in Fig. 6, the resulting partitions coming from the two covariance matrix models (EM-EII and EM-VII) are quite different. An arrow diagram (not shown) close to those described in Huth (1996) indicates that EM-EII air mass k, except for air mass 7, approximately corresponds to a mixture of EM-VII air masses k and $k+1$, as well as $k+2$ for the second and third EM-EII air masses. At seven clusters, EM-EII brings a more balanced influence of temperature and humidity on the clustering, leading to coarser tropical air masses (Fig. 6b) associated with higher standard deviation profiles (not shown) compared to EM-VII. The latter detects more accurately the tropical air masses, but with a coarser temperate air mass acting as a transition cluster (Figs. 5 and 6a). This difference in behaviour is due to the difference in the underlying assumptions, i.e. equal isotropic dispersion against variable one, the last assumption being more physically expected.

The choice between these two models will be made after studying the sensitivity of the clustering to the choice of the spatio-temporal sampling of the dataset on which the clustering is applied.

3.2 Temporal and spatial clustering sensitivities

A good quality clustering should be relatively insensitive to changes in sample size or spatial and temporal sampling. As mentioned in Sect. 2.1, each synoptic hour (UTC) gathers different local hours spatially distributed over the Earth. Therefore, sensitivity to the temporal sampling is expected to be lower than its spatial counterpart. The former is then studied before the latter.

3.2.1 Sensitivity to the temporal sampling

Depending on the choice of the spatial sampling step on which the atmospheric situations are selected for the clustering process, air masses may be significantly different. In order to know how many years, how many months a year, and so on, should be used, a sensitivity study must be performed. Studies within the period 2000 to 2009 show that resulting partitions are similar as soon as 4 months representative of each season are used. However, there is an exception for 2003, for which the features of air masses 3 and 4 are significantly different from those corresponding to the other years (not shown). In a more general framework, in order to avoid possible singular partitions due to specific thermodynamical features over the years, the training dataset will contain 2 synoptic hours, 1 day, 4 months and 5 years, hence the choice of the training dataset mentioned in Sect. 2.1.

3.2.2 Sensitivity to the spatial sampling

The temporal sampling being adopted, the sensitivity of the clustering to the choice of the spatial sampling is now studied. The latter is characterized not only by its longitude/latitude spatial sampling step, but also by its starting grid point whose choice may also alter the resulting partition. A spatial sampling step S (in grid points unit) means that one grid point out of S consecutive ones is kept in both longitude and latitude (so one out of S^2 grid points globally). The starting grid point refers here to the first grid point to be kept from which sampling is performed and chosen among the S^2 possible ones ($-180°$ W $+x \times 0.75°$ and $-90°$ S $+x \times 0.75°$ with $x = 0, \ldots, S-1$), leading to S^2 possible spatial grids.

To evaluate the impact of decreasing the spatial sampling, misclassification rates r (from 0 to 1) are used, that is, the rates of pairs of observations which are assigned to different groups between two partitions of the same size. Here, the two partitions correspond to the training dataset with no spatial sampling and have been obtained through two different ways: first, via unsupervised classification, and second, via supervised classification from each of the S^2 possible training datasets corresponding to the S^2 possible spatial grids. The box-and-whisker plots represented in Fig. 8 show the distribution of the misclassification rates with the spatial sampling step S for both EM-EII and EM-VII. In both plots, the bottom, middle and top of the blue boxes are respectively the first quartile (25 %), the second one (the median, i.e. 50 %) and the third one (75 %), whereas the whiskers indicate the minimum and maximum of the sample consisting in the S^2 misclassification rates for a given spatial sampling step S. For EM-VII, spatial sensitivity is low until $S = 5$, since misclassification values are lower than 2 % in the worst case and lower than 1 % for at least 75 % of the possible cases. For $S = 6$ and 7, sensitivity slightly increases, but the misclassification rates are not higher than 4 %, which are still relatively low values. For $S \geq 8$, there are significant differences among the partitions. The red dashed curve shows the misclassification rate for each spatial sampling step S when random starting grid points are selected, which is what will be used in practice. This curve is generally below the curve which would link the median values, meaning that randomly drawing the starting grid point for each synoptic hour selected appears to reduce even more the misclassification rate. However, EM-EII is much more sensitive to the spatial grid than its counterpart. In the following, S is set to 5 (thus 3.75° in longitude and latitude) for building the training dataset since it is a good trade-off between reducing the size of the training database and avoiding differences in terms of clustering.

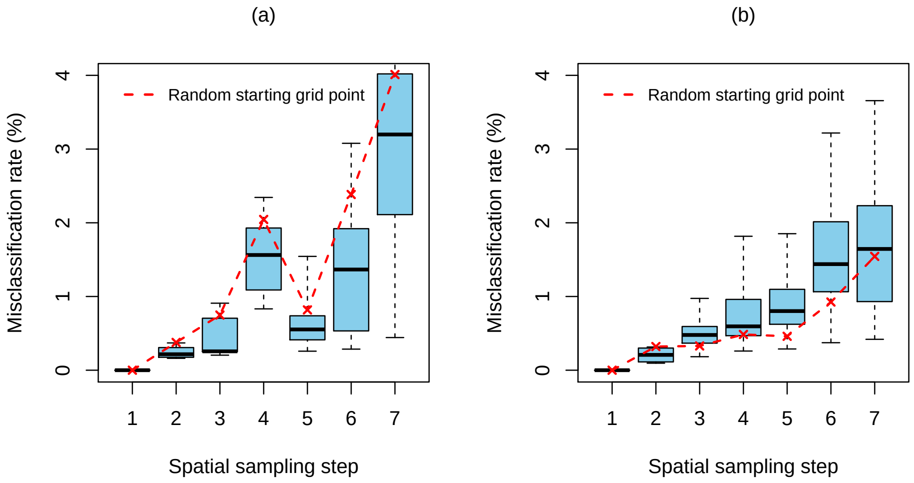

Figure 8. Clustering sensitivity to spatial sampling step and starting grid point through misclassification rates (%) for EM-EII ($\approx k$-means) **(a)** and EM-VII **(b)**.

It should be noted that the choice of the sampling in both time and space based on the results found in the present Sect. 3.2 does not change the results presented in Sect. 2.3.3 relating to the choice of the covariance matrix model and of the number of clusters.

3.3 Consistency with the number of clusters

From now on, EM-VII will be used since it relies on better physical assumptions (Sect. 2.3.3), depicts more accurately tropical air masses (Sect. 3.1) and has lower sensitivity to the choice of the spatial sampling (Sect. 3.2.2) compared to EM-EII.

Since no criterion was able to help us select the optimal number of air mass clusters, it has been subjectively fixed to seven in this paper (Sect. 2.3.3). However, such a choice may seem rather arbitrary, especially as air mass 4 acts in fact as a coarse transition cluster between tropical/sub-tropical and sub-polar/polar (Sect. 3.1.2). We now focus on the evolution of the classification with the number of clusters.

Dealing with eight clusters involves the separation of the previous air mass 4 associated with the seven-cluster partition, denoted $4_{(7)}$, into two air masses as shown in Fig. 9a, with an interesting distinction between the Northern and Southern hemispheres in northern summer: a warm, dry air mass (new air mass $4_{(8)}$), and a cool, relatively wet one (new air mass $5_{(8)}$). The first air mass can be found over both traditional continental tropical (cT) source regions (deserts) and maritime tropical (mT) ones, whereas the second air mass is located near polar source regions. At 13 clusters (Fig. 9b), we globally find back the previous 8 clusters, except for sub-polar new air masses $7_{(13)}$ and $8_{(13)}$ for which significant modifications can be noticed in northern summer, due to their high variability.

For an easier comparison of partitions for two different numbers of clusters, an arrow diagram can be used such as the one shown in Fig. 10. This figure reflects the fact that the classification is rather consistent with the number of clusters, meaning that a successive increase in the number of clusters (caused by a change in this pre-set parameter) leads to the division of a rather small set of clusters while not changing the other ones, alternating tropical/temperate air masses and polar ones at each successive increase.

In particular, the classification is stable from 7 to 8 and from 12 to 13 clusters, which would indicate at least two suitable range numbers of clusters to consider as priority. Moving from 7/8 to 12/13 is straightforward. For example, transition air masses $4_{(8)}$ and $5_{(8)}$, whose union matches $4_{(7)}$, correspond respectively to $5_{(13)}$ plus $7_{(13)}$ and $6_{(13)}$ plus $8_{(13)}$. It can be noted that polar air masses $11_{(13)}$ and $12_{(13)}$ are quite similar in the lower and mid-troposphere in both temperature and humidity, whereas their temperature profiles diverge drastically above 300 hPa.

As expected, if the clustering is found to be rather consistent with the number of clusters, the mixture model can hardly reach the perfect consistency provided by hierarchical clustering by definition (Huth, 1996; Huth et al., 2008). Even if some numbers of clusters can be chosen as priority based on the previous figure, confirming our initial choice (seven clusters), the quality of the classification ensures that the choice of the number of clusters mainly depends on the intended objective.

3.4 Supervised classification

In the following sections, the supervised classification of the atmospheric situations corresponding to a given synoptic hour is obtained by using the mixture model parameters estimated via unsupervised classification of the training dataset

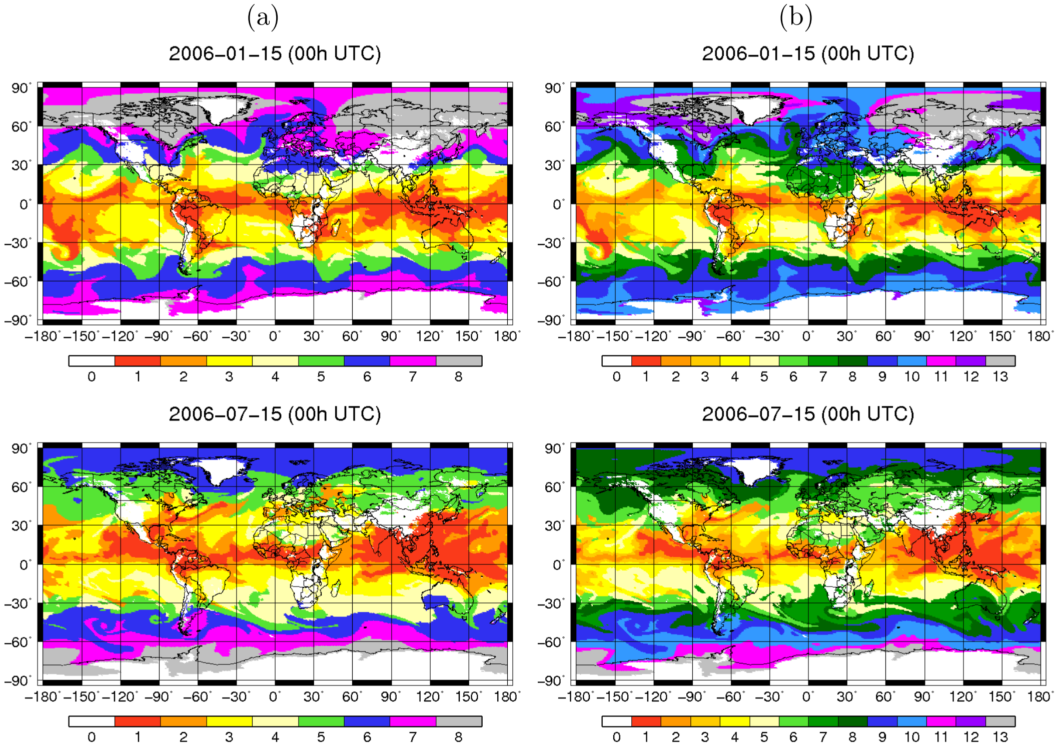

Figure 9. Unsupervised classification with EM-VII into 8 **(a)** and 13 **(b)** classes.

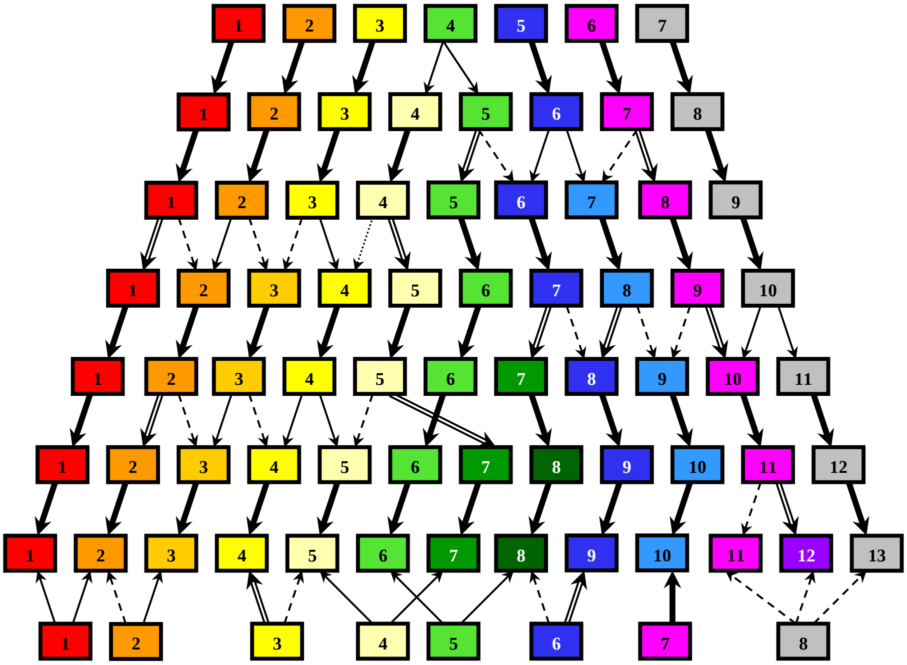

Figure 10. Arrow diagram illustrating the correspondence between the K classes of the unsupervised classification results obtained with EM-VII from $K = 7$ to $K = 13$. The classes are indicated by their number (from 1 to K). The evolution from one classification to the next one is indicated by the arrows. The style of the arrows characterizes the percentage of atmospheric situations shared relative to the size of the original class: bold (higher than 80%), doubled (between 80 and 60 %), single (between 60 and 40 %), dashed (between 40 and 20 %), dotted (lower than 20 %).

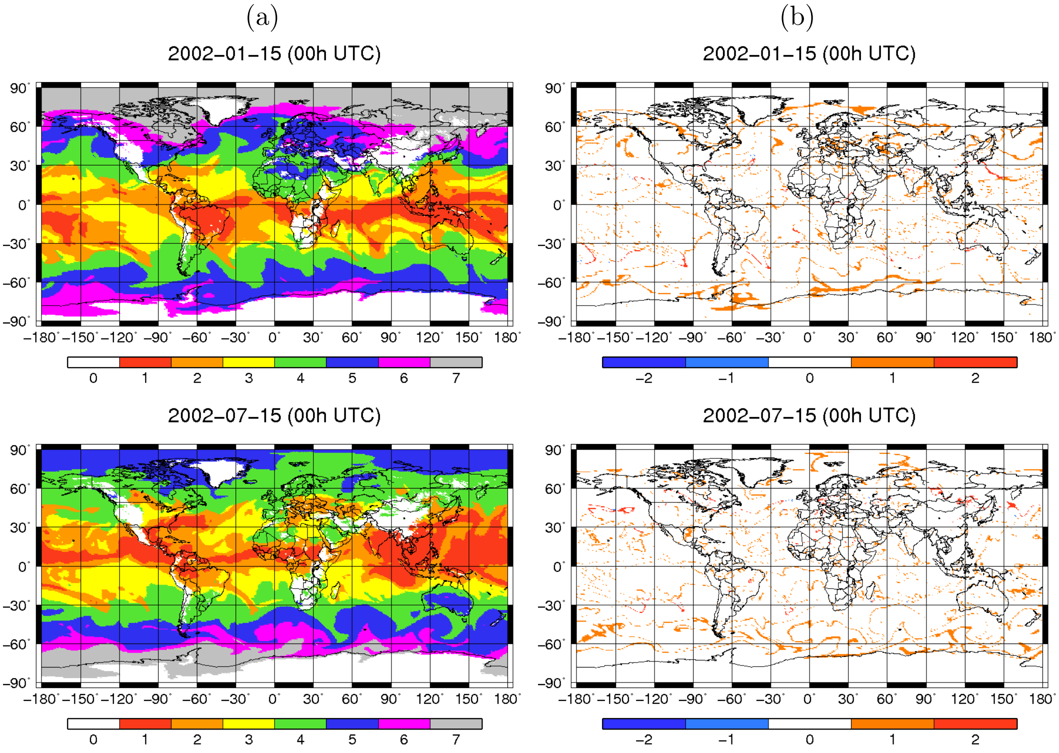

Figure 11. **(a)** Seven-cluster supervised classification with EM-VII for 2002; **(b)** cluster index difference between supervised and unsupervised classification.

with a spatial sampling step of 3.75° in both longitude and latitude and random starting grid points.

Figure 11a shows examples of supervised classification maps for the 15th day of January and July at 00:00 UTC for 1 year outside the 5-year time period of the training dataset, that is, 2002, for which the subtraction between supervised classification and unsupervised classification (using in that case the period 2000–2004 instead of 2005–2009 for the training dataset) is also shown. It is important to notice that air mass patterns are similar to the ones resulting from unsupervised classification at the same day, which is expected since air masses retain their essential features when they are not sensitive to the choice of the spatial and temporal richness of the training dataset. According to Fig. 11b, misclassified situations are mainly located in the narrow regions between the air mass clusters. If this property is verified at the temporal scales studied in this paper, that may not be the case for studies over longer periods spanning several decades.

4 Cluster analysis

4.1 A posteriori probabilities of belonging to each cluster and uncertainty

In contrast to the k-means algorithm, EM gives access to a posteriori probabilities of belonging to a given class for each of the atmospheric situations studied, and thus to the error probability, that is, the probability that an atmospheric situation will be assigned to a group which is not associated with the highest a posteriori probability. Figure 12 focuses on the geographic area from 60 to 30° S and from 105 to 30° W on 15 July 2006, 00:00 UTC, which is characterized by several types of surface, from mountain to sea, and by a depression located west of Chile. In this figure are represented, from top to bottom and from left to right, (a) the corresponding supervised classification map; (b) the error probability for each of the atmospheric situations, that is, one minus the highest occurrence probability among the posterior probabilities t_{ik} that observation i will belong to class k; (c) to (f), the posterior probabilities t_{i4}, t_{i5}, t_{i6} and t_{i7} of belonging to classes 4, 5, 6 and 7 respectively.

Non-zero error probabilities are located at the transition between different air masses where they take the highest values, but only a few situations are involved, meaning that air mass classes are rather well separated. Besides, a plot of the competitiveness index p_1/p_2 between the two highest occurrence probabilities per class p_1 and p_2 where p_x is the posterior probability and x the probability rank from 1 (the highest probability) to K (the lowest probability), as in Levavasseur et al. (2012), shows that this index is highly correlated with the error probability; the probabilities $p_3,\ldots, p_7$ are thus almost always negligible (not shown).

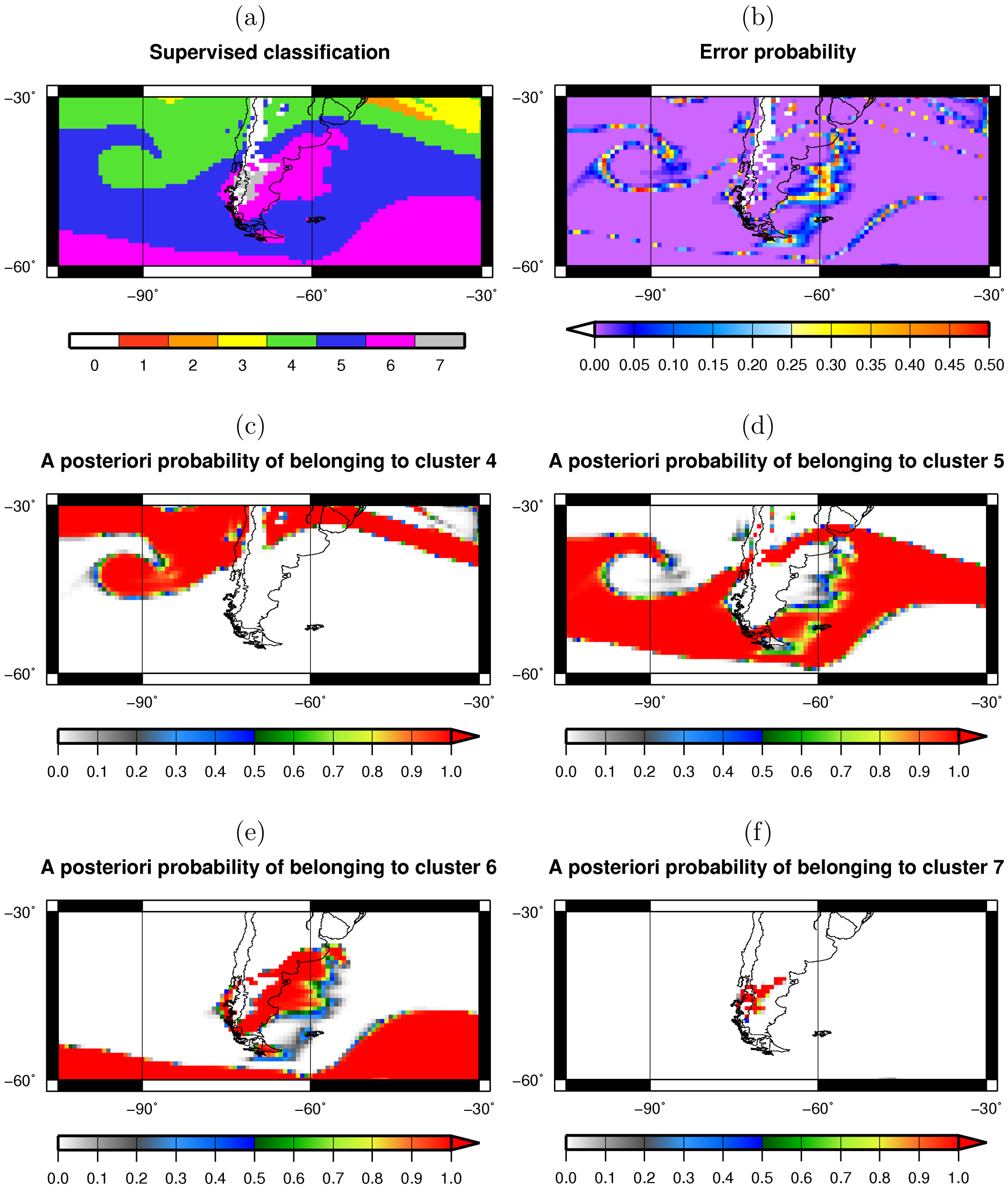
(a)
Supervised classification
(b)
Error probability
(c)
A posteriori probability of belonging to cluster 4
(d)
A posteriori probability of belonging to cluster 5
(e)
A posteriori probability of belonging to cluster 6
(f)
A posteriori probability of belonging to cluster 7
−30°
−60°
−90°
−60°
−30°
0 1 2 3 4 5 6 7
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Figure 12. Geographical zoom on 15 July 2006 (00:00 UTC). Supervised classification map **(a)** with the associated error probability **(b)** and the probabilities of belonging to clusters 4 to 7 (**c** to **f**).

The boundary regions separating the different air masses, across which the thermodynamic conditions change rapidly, correspond to meteorological fronts whose 1 to 3° narrow extents in longitude and latitude are visible in plots (c) to (f). The depression located west of Chile is highlighted by the separation between air mass 4 and the drier, colder air mass 5. Besides, the latter rises to higher altitudes when approaching the Andes Mountains, becoming colder and drier such that it is converted progressively to air masses 6 and 7, before descending to lower altitudes beyond the high reliefs by finding back warmer and wetter features.

A plot representing the percentage of atmospheric situations against the number of classes per error probability step (every 0.1 for example) confirms that not only is the number of situations associated with a high error probability low, but also indicates that it slightly increases with the number of classes, since there are more boundary regions between air masses. Furthermore, it shows that the range 7/8 and 12/13 classes mentioned in Sect. 3.3 (as well as 16/17) are associated with a slight decrease in error probability, although the corresponding value is too low to indicate any optimal number of classes (not shown).

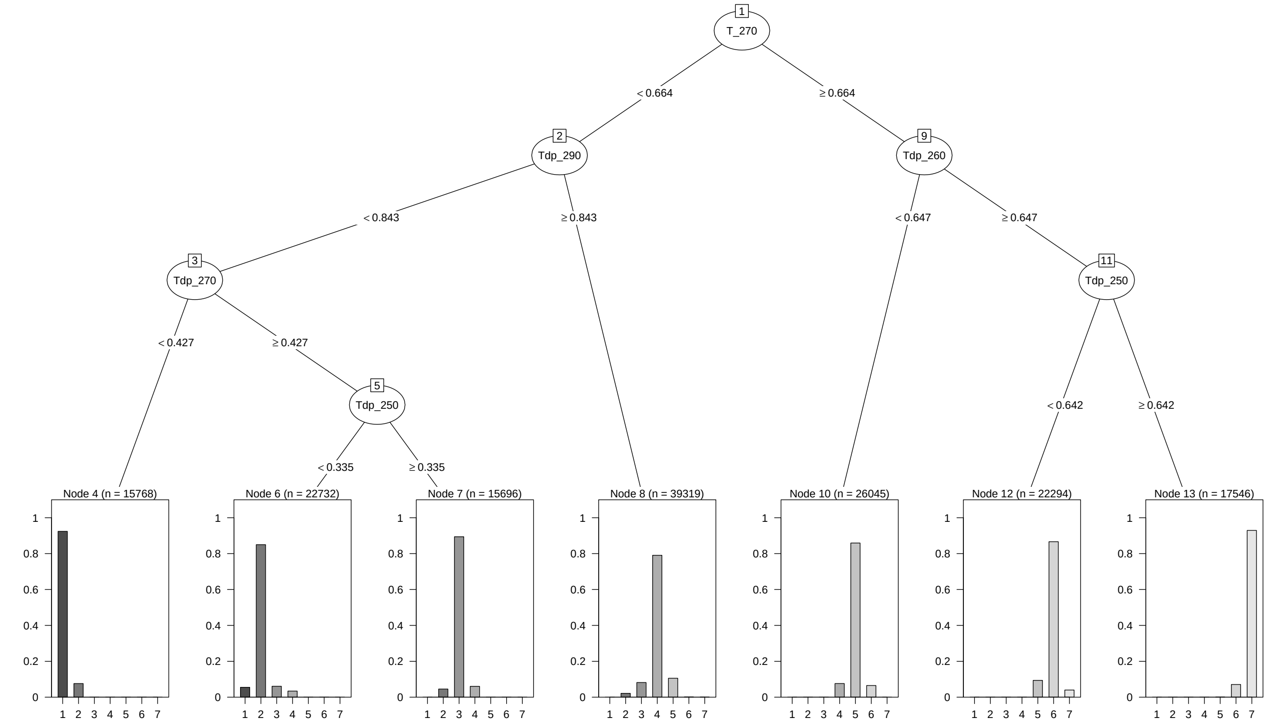

Figure 13. Decision tree obtained from the seven-cluster partition built with the training database and EM-VII.

Error probabilities can be used for adding one or several transition classes associated with a low level of confidence or for keeping observations whose classification is considered sufficiently representative of the corresponding class features to be associated with some meteorological phenomena. Probability distributions can also be used to enhance a priori information in remote sensing applications, for example.

4.2 Cluster interpretation with a decision tree

Another interesting way to interpret a partition is to build a decision tree, that is, a supervised classification method in the form of a tree structure separating a dataset into smaller and smaller subsets through some decision rules given a partition known a priori. The goal of such a tree is to predict the value of the variable to be explained (here, the class to which a given observation belongs) given a subset of input explanatory variables (here, the $D = 20$ CDF values) corresponding to observations whose partition is already known. For this purpose, R package "rpart" (Therneau and Atkinson, 2015) has been used, based mainly on Breiman et al. (1984) and on binary trees: each node has at most two children. In this section, the terms "cluster" and "class" will refer respectively to the unsupervised classification and supervised classification results.

Separations between the nodes have been performed via maximal impurity reduction, with the use of the Gini index as an impurity function. That means that the tree tries to build nodes containing as few clusters from the reference partition as possible. In order to compare the classes obtained from the decision tree to the reference partition, the tree has been pruned to seven terminal nodes. It is done by setting a complexity parameter measuring the "cost" of adding another explanatory variable among the 20 possible ones in the model underlying the decision tree. For more technical details, see Therneau and Atkinson (2015). Pruning the tree implies that we may not have the same set of explanatory variables as used previously in the EM algorithm.

Figure 13 shows the classification tree obtained by considering as a priori probabilities of belonging to each cluster the mixture proportions of the reference partition obtained via unsupervised classification with EM-VII on the training dataset. These mixture proportions are similar to those listed in the fourth line of Table 1 for EM-VII. In Fig. 13, CDF values $F_{\mathrm{T}}\ (x) = P\ (T \leq x)$ for temperature and $F_{\mathrm{Tdp}}\ (x) = P\ (\mathrm{Tdp} \leq x)$ for dew point temperature are denoted respectively by "T_x" and "Tdp_x". The terminal nodes could be considered as air mass classes by following the decision rules. For instance, class 7 on the bottom right would be associated with the decision rule $F_{\mathrm{T}}\ (270\,\mathrm{K}) \geq 0.664$ AND $F_{\mathrm{Tdp}}\ (260\,\mathrm{K}) \geq 0.647$ AND $F_{\mathrm{Tdp}}\ (250\,\mathrm{K}) \geq 0.642$), and thus can be interpreted as an air mass whose atmospheric situations satisfy $P\ (T \leq 270\,\mathrm{K}) \geq 0.664$ and $P\ (\mathrm{Tdp} \leq 250\,\mathrm{K}) \geq 0.642$. Each of the bar charts related to these terminal nodes indicates the proportions of observations (from 0 to 1) belonging to each cluster of the reference partition.

The most striking feature appearing in Fig. 13 is the fact that temperature is used first to separate the atmospheric situations into two main groups, i.e. polar and sub-polar air masses on the one hand, and temperate, sub-tropical and tropical ones on the other hand. Then, humidity is used to make the remaining separations in order to obtain a seven-class partition. This confirms the findings described in Sect. 3 (high correlation between air mass clustering and humidity). From the decision tree, temperature variables could be considered less important compared to humidity ones, in terms of the respective number of variables used to build the tree. However, removing the temperature CDF value used as the first variable, i.e. F_T (270 K) here, used to separate the root node of the tree, only leads to replacing it by F_T (260 K) with a different threshold value for the first decision rule, the rest of the tree remaining similar to the one displayed in Fig. 13. And this process can be continued twice with the following successive variables: F_T (280 K) and F_T (290 K). That proves that the tree is here not only very sensitive to the first variable used to make the first separation, but also that temperature is necessary in the clustering process. That is confirmed by the partition obtained with EM-VII by taking into account only the five variables which have been used in the model underlying the decision tree: one CDF value in temperature against four in dew point temperature is not enough to counterbalance the fact that air masses are too closely correlated with humidity (not shown).

Misclassification rates, i.e. one minus the highest proportion of observations belonging to each cluster, are particularly low. For instance, considering the bottom right group as an air mass class implies that the latter would contain here 17 546 observations coming at 93 % from cluster 7 of the reference partition and at 7 % from cluster 6, and would be associated with a misclassification rate of 7 %. These low values indicate that the clustering process used to create the reference partition is efficient and robust and that resulting partitions make sense.

The same study with EM-EII shows that the resulting decision tree alternates temperature and humidity decision rules (not shown): temperature seems to have a higher importance in the partitioning with EM-EII than with EM-VII, which explains the difference between both types of classification.

5 Summary and conclusions

In this paper, a methodology for unsupervised and supervised classifications of various and large atmospheric databases into distinct air masses has been proposed and applied to thermodynamic profiles (temperature and dew point temperature) from ECMWF reanalyses. These three-dimensional data are gridded in latitude, longitude and vertical layers, homogeneously distributed over the Earth, and span the period 2000–2009. This methodology follows a similar probabilistic point of view considered by Vrac et al. (2005, 2011) through a different approach to the problem of mixture models (estimation approach against clustering one previously). It relies (i) on a probabilistic classification method based on a multivariate Gaussian mixture model whose parameters are estimated via maximum likelihood estimation by the expectation–maximization (EM) algorithm; and (ii) on the use of probabilistic data: classical thermodynamic values at different pressure levels are converted into a set of cumulative distribution function (CDF) values whose number represents the number of statistical variables needed to characterize each atmospheric situation. This data compression step implies a description of the data different from the common ones, giving information on the vertical distribution of the temperature and dew point temperature values regardless of the successive pressure levels.

In Vrac et al. (2005, 2011), (i) a limited set of observations consisting of only 1 winter day was used as a training dataset for classifying new data through projections not exceeding 1 month; (ii) only four statistical variables were used to characterize each atmospheric situation; (iii) an initial partition based on seven subjective zonal clusters homogeneous in temperature and humidity was used. Such choices were not enough to steadily characterize air masses at any time and any location on large temporal scales. To overcome this problem, several updates have been implemented. First, a much larger set of observations has been selected as a training dataset in order to take into account their high variability, that is, 2 synoptic hours of the central day of 4 months representative of each season for a period covering 5 years. Second, each atmospheric situation has been characterized by a substantially higher number of statistical variables for a better thermodynamical description of the profiles: 10 CDF values for temperature, and 10 for dew point temperature. And third, an initialization strategy for EM based on the use of a suitable random initial partition has been adopted to avoid the use of arbitrarily chosen prior information.

Furthermore, 14 models adding different constraints (or not) to the structure of the covariance matrices and thus to the dispersion of the observations have been studied, since dealing with the unconstrained model does not provide representative partitions. Several criteria have been tested as a selection criterion for both the covariance matrix model and the number of clusters. However, no optimal number of clusters emerges from their evolution. Hence, following Vrac et al. (2005, 2011), seven clusters have been subjectively chosen.

If most of the covariance matrix models imply either too much zonal structure or a preponderance of one air mass class over the other ones, three of them lead to relevant air mass spatial regions. These three models are distinguished by a different relative influence of temperature and humidity on the classification process, as shown by the use of a decision tree for helping in the interpretation of the resulting clusters. For instance, the two models EII and VII assume either equal isotropic dispersion between the clusters (equivalent in fact to the widely used k-means algorithm) or variable one. The latter is more physically expected and leads to a classification which depicts more accurate tropical air masses due to a stronger influence of humidity on the classification. These results show that EM generalizes k-means by providing, in addition, the probabilities of belonging to each class for each atmospheric situation and thus the corresponding uncertainties as well.

The proposed method shows low temporal and spatial sensitivity to the choice of the training dataset. Within the period 2000–2009, we have not only shown that the partitions are similar as soon as the training dataset contains 1 day of 4 months representative of each season and spans several years, but also that a 3.75° by 3.75° spatial sampling delivers partitions similar to those obtained with no spatial sampling. This low sensitivity provides the ability to classify new data through long-range projections (more than 3 years).

Our method complies with the five properties introduced by Huth (1996) and Huth et al. (2008) to assess the quality of a classification: (i) the method reproduces expected patterns known to exist in the data, as low-pressure systems or the traditional winter continental polar (cP) air mass; (ii) it shows little sensitivity in time and space to the choice of the training dataset, both in terms of observation selection and size; (iii) it shows neither high equability (clusters tending to be equal in size) nor low equability (a huge cluster accompanied by small ones, called the snowballing effect); (iv) it makes a good distinction between clusters since the boundary regions separating the air masses, associated with high uncertainty, present narrow extents not exceeding 3° in longitude and latitude, despite the difficulties induced by the continuous nature and the high variability of the atmosphere; and (v) it is in fact quite consistent with the number of clusters, meaning adding successively one cluster does not drastically change most of the clusters.

Based on temperature and dew point temperature variables, the proposed classification method is applicable to most atmospheric datasets used by the atmospheric science community, such as radiosonde measurements, meteorological reanalyses or satellite data. Depending on the intended objective, other variables could also be considered, especially dynamic variables to help monitor air mass movement, such as potential vorticity, which is commonly used for weather analysis (e.g. Emanuel, 2008). An important feature of this method consists in providing probabilistic information, which can be used to provide the uncertainties associated with the classes or improving a priori information in many atmospheric applications such as in remote sensing. Finally, through the evolution of the classes and their associated probabilities along several decades, the method could be easily adapted to evaluate general circulation models and study climate variability and potential changes at different spatial and temporal scales.

6 Data availability

The temperature and specific humidity profiles as well as the surface temperatures and pressures used in this study come from ERA-Interim global atmospheric reanalyses (ECMWF, 2016) and can be downloaded for example from http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/. The elevations above sea level are from the 1 / 6° global elevation dataset compiled by the U.S. Navy Fleet Numerical Oceanography Center (2016) (which can be downloaded from http://rda.ucar.edu/datasets/ds754.0/). The EM algorithm has been implemented using the Rmixmod S4 package (Mixmod Team, 2008; Lebret et al., 2015), which is available at https://cran.r-project.org/web/packages/Rmixmod/index.html (CRAN, 2016a). As for the decision tree, it has been implemented using R package rpart, which can be downloaded from https://cran.r-project.org/web/packages/rpart/index.html (CRAN, 2016b).

Appendix A: Eigenvalue decomposition of the covariance matrices

Table A1. The 14 covariance matrix models.

Model	Distribution category	Identifier
λI	Hyper- spherical	EII
$\lambda_k I$		VII
λE	Hyper- diagonal	EEI
$\lambda_k E$		VEI
λE_k		EVI
$\lambda_k E_k$		VVI
$\lambda V E V'$	Hyper- ellipsoidal	EEE
$\lambda_k V E V'$		VEE
$\lambda V E_k V'$		EVE
$\lambda_k V E_k V'$		VVE
$\lambda V_k E V_k'$		EEV
$\lambda_k V_k E V_k'$		VEV
$\lambda V_k E_k V_k'$		EVV
$\lambda_k V_k E_k V_k'$		VVV

Here we give more details on the eigenvalue decomposition of the covariance matrices described in Banfield and Raftery (1993) and Celeux and Govaert (1995). Each covariance matrix Σ_k related to the kth cluster is expressed in terms of their eigenvalue decomposition, whose general form is

$$\Sigma_k = \lambda_k V_k E_k V_k', \tag{A1}$$

where λ_k, E_k and V_k determine respectively the hypervolume, the shape and the orientation of the isocontour of mixture component distributions associated with the kth cluster. We denote by V_k the matrix of eigenvectors, by V_k' its transpose and by E_k the diagonal matrix of eigenvalues. The latter is scaled so that $|E_k| = 1$, with the normalized eigenvalues of Σ_k in decreasing order. Then, $\lambda_k = |\Sigma_k|^{1/D}$, where D is the number of dimensions. The presence of the subscript k implies that λ, E or V can vary between the clusters or are equal otherwise.

Such decomposition leads to 14 parsimonious models (column 1 in the table below) depending on whether some assumptions about the structure of the covariance matrices are added or not.

These models can be classified into three families (column 2): the hyperspherical models (isotropic dispersion), the hyperdiagonal models (coordinate axis-aligned orientation) and the hyperellipsoidal models (free orientation).

These models can be simply indicated by three sequential letters (column 3) corresponding to the three attributes characterizing the dispersion of the mixture component distributions, that is, the hypervolume, the shape and the orientation of their isocontour in the multidimensional space, providing an easy geometric interpretation of the models. Each letter indicates whether the corresponding attribute is equal (E) or variable (V) between the clusters, or does not make sense (I), so that the Σ_k are in that case assumed to be either identity matrices in the case of the hyperspherical models or diagonal matrices in the case of the hyperdiagonal models (V_k are then permutation matrices).

For illustrative purposes, the typical isocontours of the mixture component distributions are commonly drawn in a two-dimensional subspace, where the hypervolume, shape and orientation features then correspond respectively to the surface, the major and minor axis ratios, and the orientation of the major axis of the elliptic isocontours. In the case of the two hyperspherical models, elliptic isocontours are reduced to circles.

Edited by: W. Kleiber
Reviewed by: two anonymous referees

References

Akaike, A.: Information theory and an extension of the maximum likelihood principle, in: Second International Symposium on Information Theory, edited by: Petrov, B. N. and Csaki, F., Akadémiai Kiado, Budapest, Hungary, 267–281, 1973.

Banfield, J. D. and Raftery, A. E.: Model-based Gaussian and non-Gaussian clustering, Biometrics, 49, 803–821, doi:10.2307/2532201, 1993.

Barry, R. G. and Perry, A. H.: Synoptic Climatology and Its Applications, in: Synoptic and Dynamic Climatology, edited by: Barry, R. G. and Carleton, A. M., Routledge, London, UK, 547–603, 2001.

Bayes, T. and Price, M.: An Essay towards solving a Problem in the Doctrine of Chances, Philos. T. R. Soc. Lond., 53, 370–418, doi:10.1098/rstl.1763.0053, 1763.

Bergeron, T.: Richtlinien einer dynamischen klimatologie, Meteorol. Z., 47, 246–262, 1930.

Biernacki, C., Celeux, G., and Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood, IEEE T. Pattern Anal., 22, 719–725, doi:10.1109/34.865189, 2000.

Biernacki, C., Celeux, G., and Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models, Comput. Stat. Data An., 41, 561–575, doi:10.1016/S0167-9473(02)00163-9, 2003.

Billard, L. and Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley series in computational statistics, John Wiley & Sons Ltd, Chichester, UK, 330 pp., doi:10.1002/9780470090183.ch1, 2012.

Bock, H. H. and Diday, E.: Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Berlin and Heidelberg, Germany, 443 pp., doi:10.1007/978-3-642-57155-8, 2000.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone C. J.: Classification and Regression Trees, The Wadsworth and Brooks-Cole statistics-probability series, The Wadsworth statistics/probability series, Wadsworth and Brooks, Monterey, CA, USA, 368 pp. 1984.

Buck, A. L.: New equations for computing vapor pressure and enhancement factor, J. Appl. Meteorol., 20, 1527–1532, doi:10.1175/1520-0450(1981)020<1527:NEFCVP>2.0.CO;2, 1981.

Carreau, J. and Vrac, M.: Stochastic downscaling of precipitation with neural network conditional mixture models, Water. Resour. Res., 47, W10502, doi:10.1029/2010WR010128, 2011.

Celeux, G. and Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data An., 14, 315–332, doi:10.1016/0167-9473(92)90042-E, 1992.

Celeux, G. and Govaert, G.: Gaussian parsimonious clustering models, Pattern Recogn., 28, 781–793, doi:10.1016/0031-3203(94)00125-6, 1995.

Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., and Ralambondrainy, H.: Classification automatique des données, Dunod Informatique, Paris, France, 1989.

Chédin, A. and Scott, N. A.: Initialization of the radiative transfer equation inversion problem from a pattern recognition type approach. Applications to the satellites of the Tiros-N Series, in: Advances in Remote Sensing Retrieval Methods, Deepak A., Academic Press, New York, USA, 495–515, 1985.

Chevallier, F., Chédin, A., Cheruy, F., and Morcrette, J. J.: TIGR-like atmospheric-profile databases for accurate radiative-flux computation, Q. J. Roy. Meteor. Soc., 126, 777–785, doi:10.1002/qj.49712656319, 2000.

CRAN: EM algorithm, available at: https://cran.r-project.org/web/packages/Rmixmod/index.html, last access: 10 October 2016a.

CRAN: Decision tree, available at: https://cran.r-project.org/web/packages/rpart/index.html, last access: 10 October 2016b.

Crowe, P. R.: Concepts in Climatology, Longman, London, UK, 612 pp., 1971.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Normann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimverger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

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

Diday, E.: A generalization of the mixture decomposition problem in the symbolic data analysis framework, Research Report, CEREMADE, Paris, France, 2001.

Diday, E. and Vrac, M.: Mixture decomposition of distributions by copulas in the symbolic data analysis framework, Discrete Appl. Math., 147, 27–41, doi:10.1016/j.dam.2004.06.018, 2005.

Diday, E., Schroeder, A., and Ok, Y.: The dynamic clusters method in pattern recognition, in: Proceedings of International Federation for Information Processing congress 74, Stockholm, Sweden, 5–10 August 1974, 691–697, 1974.

ECMWF: ERA Interim, Daily, available at: http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/, last access: 10 October 2016.

Emanuel, K.: Back to Norway: An essay, Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting, Meteor. Mon., 33, 87–96, doi:10.1007/978-0-933876-68-2, 2008.

Floriana, E. and Diday, E.: An introduction to symbolic data analysis and the SODAS software, Intell. Data Anal., 7, 583–601, 2003.

Forgy, E. W.: Cluster analysis of multivariate data: efficiency vs interpretability of classifications, Biometrics, 21, 768–769, 1965.

Fraley, C. and Raftery, A. E.: Model-based clustering, discriminant analysis and density estimation, J. Am. Stat. Assoc., 97, 611–631, doi:10.1198/016214502760047131, 2002.

Gordon, A. D.: Classification (2nd Edition), Chapman and Hall/CRC Press, London, UK, 256 pp., 1999.

Hardy, A.: On the number of clusters, Comput. Stat. Data An., 23, 83–96, doi:10.1016/S0167-9473(96)00022-9, 1996.

Hardy, A.: NBCLUST, A module for the determination of the number of clusters in the SODAS 2 software, IFCS 2006 Conference, Ljubjiana, Slovenia, 2006.

Hartigan, J. A. and Wong, M. A.: Algorithm AS136, A *k*-means clustering algorithm, J. Roy. Stat. Soc. C-App., 28, 100–108, doi:10.2307/2346830, 1979.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition), Springer, New York, USA, 745 pp., doi:10.1007/978-0-387-84858-7, 2009.

Hewitson, B. C. and Crane, R. G.: Self-organizing maps: applications to synoptic climatology, Clim. Res., 22, 13–26, doi:10.3354/cr022013, 2002.

Huth, R.: An intercomparison of computer-assisted circulation classification methods, Int. J. Climatol., 16, 893–922, doi:10.1002/(SICI)1097-0088(199608)16:8<893::AID-JOC51>3.0.CO;2-Q, 1996.

Huth, R.: Disaggregating climatic trends by classification of circulation patterns, Int. J. Climatol., 21, 135–153, doi:10.1002/joc.605, 2001.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J., and Tveito, O. E.: Classifications of atmospheric circulation patterns. Recent advances and applications, in: Trends and Directions in Climate Research, Ann. NY Acad. Sci., 1146, 105–152, doi:10.1196/annals.1446.019, 2008.

Kalkstein, L. S., Tan, G., and Skindlov, J. A.: An evaluation of three clustering procedures for use in synoptic climatological classification, J. Clim. Appl. Meteorol., 26, 717–730, doi:10.1175/1520-0450(1987)026<0717:AEOTCP>2.0.CO;2, 1987.

Kalkstein, L. S., Barthel, C. D., Nichols, M. C., and Greene, J. S.: A New Spatial Synoptic Classification: Application to Air Mass Analysis, Int. J. Climatol., 16, 983–1004, doi:10.1002/(SICI)1097-0088(199609)16:9<983::AID-JOC61>3.0.CO;2-N, 1996.

Levavasseur, G., Vrac, M., Roche, D. M., and Paillard, D.: Statistical modelling of a new global potential vegetation distribution, Environ. Res. Lett., 7, 044019, doi:10.1088/1748-9326/7/4/044019, 2012.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G.: Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library, J. Stat. Softw., 67, 241–270, doi:10.18637/jss.v067.i06, 2015.

Lee, C. C. and Sheridan, S. C.: Synoptic Climatology: An Overview, Reference Module in Earth Systems and Environmental Sciences, doi:10.1016/B978-0-12-409548-9.09421-5, 2015.

Lloyd, S.: Least squares quantization in PCM, Technical Note, Bell Telephone Laboratories Paper, published in journal much later in 1982 in IEEE T. Inform. Theory, 28, 128–137, doi:10.1109/TIT.1982.1056489, 1957.

MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, USA, 1, 281–297, 1967.

McLachlan, G. J. and Krishnan, T.: The EM algorithm and Extensions (2nd edition), Wiley, New York, USA, 400 pp., 2008.

Mixmod Team: Mixmod Statistical Documentation, Technical report, CNRS, Université de Besançon, Université de Franche-Comté, Besançon, France, 2008.

Molliere, J. L.: What is the real number of clusters?, 9th meeting of the German Classification Society, 26–28 June 1985, University of Karlsruhe, Karlsruhe, Germany, 1985.

Nelsen, R. B.: An Introduction to Copulas, Springer-Verlag, New York, USA, 1999.

Parzen, E.: On estimation of a probability density function and mode, Ann. Math. Stat., 33, 1065–1076, doi:10.1214/aoms/1177704472, 1962.

Peixoto, J. P. and Oort, A. H.: The climatology of relative humidity in the atmosphere, J. Climate, 9, 3443–3463, doi:10.1175/JCLI3956.1, 1996.

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., Huth, R., James, P., Jourdain, S., Kreienkamp, F., Krennert, T., Lykoudis, S., Michalides, S. C., Pianko-Kluczynska, K., Post, P., Alvarez, D. R., Scheimann, R., Spekat, A., and Tymvios, F. S.: Cost733cat – a database of weather and circulation type classifications, Phys. Chem. Earth, 35, 360–373, 2010.

Pudil, P., Novovičová, J., and Kittler, J.: Floating Search methods in Feature Selection, Pattern Recogn. Lett., 15, 1119–1125, doi:10.1016/0167-8655(94)90127-9, 1994.

Raftery, A. E.: Bayesian Model Selection in Social Research, Sociol. Methodol., 25, 111–164, doi:10.2307/271063, 1995.

Raftery, A. E. and Dean, N.: Variable Selection for Model-Based Clustering, J. Am. Stat. Assoc., 101, 168–178, doi:10.1198/016214506000000113, 2006.

Reusch, D. B., Alley, R. B., and Hewitson, B. C.: North Atlantic climate variability from a self-organizing map perspective, J. Geophys. Res., 112, 2104, doi:10.1029/2006JD007460, 2007.

Rosenblatt, W.: Remarks on some nonparametric estimates of a density function, Ann. Math. Stat., 27, 832–837, doi:10.1214/aoms/1177728190, 1956.

Rust, H., Vrac, M., Lengaigne, B., and Sultan, B.: Quantifying differences in circulation patterns based on probabilistic models: IPCC-AR4 multi-model comparison for the North Atlantic, J. Climate, 23, 6573–6589, doi:10.1175/2010JCLI3432.1, 2010.

Schwarz, G.: Estimating the Dimension of a Model, Ann. Stat., 6, 461–64, doi:10.1214/aos/1176344136, 1978.

Schweizer, B.: Distributions are the numbers of the future, in: Proceedings of the Mathematics of Fuzzy Systems Meeting, edited by: diNola, A. and Ventre, A., University of Naples, Naples, Italy, pp. 137–149, 1984.

Schweizer, B. and Sklar, A.: Probabilistic Metric Spaces, North-Holland Publishing Company, New York, USA, 1983.

Scott, N. A., Chédin, A., Armante, R., Francis, J., Stubenrauch, C., Chaboureau, J. P., Chevallier, F., Claud, C., and Chéruy, F.: Characteristics of the TOVS Pathfinder Path-B data set, B. Am. Meteorol. Soc., 80, 2679–2701, doi:10.1175/1520-0477(1999)080<2679:COTTPP>2.0.CO;2, 1999.

Sheridan, S. C. and Lee, C. C.: Synoptic Climatology, in: Oxford Bibliographies in Geography, edited by: Warf, B., Oxford University Press, New York City, USA, 2013.

Symons, M. J.: Clustering criteria and multivariate normal mixtures, Biometrics, 37, 35–43, doi:10.2307/2530520, 1981.

Therneau, T. M. and Atkinson, E. J.: An Introduction to Recursive Partitioning Using the RPART Routines, Technical Report Series No. 61, MN: Section of Biostatistics, Mayo Clinic, Rochester, USA, 2015.

U.S. Navy Fleet Numerical Oceanography Center: U.S. Navy 10-Minute Global Elevation and Geographic Characteristics, available at: http://rda.ucar.edu/datasets/ds754.0/, last access: 10 October 2016.

Vergados, P., Mannucci, A. J., Ao, C. O., Jiang, J. H., and Su, H.: On the comparisons of tropical relative humidity in the lower and middle troposphere among COSMIC radio occultations and MERRA and ECMWF data sets, Atmos. Meas. Tech., 8, 1789–1797, doi:10.5194/amt-8-1789-2015, 2015.

Vrac, M.: Analyse et Modélisation de Données Probabilistes par Decomposition de Mélange de Copules et Application à une Base de Données Climatologiques, PhD, Dissertation, University of Paris, France, 2002.

Vrac, M., Chédin, A., and Diday, E.: Clustering a Global Field of Atmospheric Profiles by Mixture Decomposition of Copulas, J. Atmos. Ocean. Tech., 22, 1445–1459, doi:10.1175/JTECH1795.1, 2005.

Vrac, M., Hayhoe, K., and Stein, M.: Identification and intermodal comparison of seasonal circulation patterns over North America, Int. J. Climatol., 27, 603–620, doi:10.1002/joc.1422, 2007.

Vrac, M., Billard, L., Diday, E., and Chédin, A.: Copula Analysis of Mixture Models, Computational Stat., 27, 427–457, doi:10.1007/s00180-011-0266-0, 2011.

Willett, H. C.: American air mass properties, Papers Physical Oceanography and Meteorology, Massachusetts Institute of Technology, Cambridge, MA, USA, 2, 1–116, 1933.

Yarnal, B.: Synoptic Climatology in Environmental Analysis: A Primer, Belhaven Press, London, UK, 256 pp., doi:10.1002/joc.3370140116, 1993.

Yarnal, B., Comrie, A. C., and Frakes, B., and Brown, D. P.: Developments and prospects in synoptic climatology: A Primer, Int. J. Climatol., 21, 1923–1950, doi:10.1002/joc.675, 2001.

Živkovíć, M.: Hierarchical clustering of atmospheric soundings, Int. J. Climatol., 15, 1099–1124, doi:10.1002/joc.3370151004, 1995.