Weather regimes designed for local precipitation modelling: Application to the Mediterranean basin

Mathieu Vrac, Pascal Yiou

▶ **To cite this version:**

Mathieu Vrac, Pascal Yiou. Weather regimes designed for local precipitation modelling: Application to the Mediterranean basin. Journal of Geophysical Research: Atmospheres, 2010, 115 (D12), pp.D12103. 10.1029/2009JD012871 . hal-00531248

HAL Id: hal-00531248
https://hal.science/hal-00531248

Submitted on 17 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 115, D12103, doi:10.1029/2009JD012871, 2010

Click Here for Full Article

Weather regimes designed for local precipitation modeling: Application to the Mediterranean basin

Mathieu Vrac[1] and Pascal Yiou[1]

Received 22 July 2009; revised 26 January 2010; accepted 4 February 2010; published 16 June 2010.

[1] Although weather regimes are often used as a primary step in many statistical downscaling processes, they are usually defined solely in terms of atmospheric variables and seldom to maximize their correlation to observed local meteorological phenomena. This paper compares different clustering methods to perform such a task. The correlation clustering model is introduced to define regimes that are well correlated to local-scale precipitation observed on seven French Mediterranean rain gauges. This clustering method is compared to other approaches such as the *k*-means and "expectation-maximization" (EM) algorithms. The two latter are applied either to the main principal components of large-scale reanalysis data (geopotential height at 500 mbar and sea level pressure) covering the Mediterranean basin or to the canonical variates associated with large scale and resulting from a canonical correlation analysis performed on reanalyses and local precipitation. The weather regimes obtained by the different approaches are compared, with a focus on the "extreme content" captured within the regimes. Then, cost functions are developed to quantify the errors due to misclassification, in terms of local precipitation. The different clustering approaches show different misclassification and costs. EM applied to canonical variates appears as a good compromise between the other approaches, with high discrimination, overall for extreme precipitation, while the precipitation costs due to bad classification are acceptable. This paper provides tools to help the users choose the clustering method to be used according to the expected goal and the use of the weather regimes.

Citation: Vrac, M., and P. Yiou (2010), Weather regimes designed for local precipitation modeling: Application to the Mediterranean basin, *J. Geophys. Res.*, *115*, D12103, doi:10.1029/2009JD012871.

1. Introduction and State of the Art

[2] Weather regimes (WRs) of the atmospheric circulation provide a simple approach to characterize the main atmospheric variability over a given region. A weather regime can be defined as a recurrent large-scale spatial atmospheric structure (a deformation radius of at least several hundreds of kilometers), usually described in terms of circulation variables (geopotential height, pressure, etc.). It is generally assumed that the spatial specificity of a WR induces recurrent local-scale meteorological conditions at some correlated locations. Many different methods can be employed to define WRs. Those methods can be divided into subjective and objective approaches.

[3] In the subjective definition, a meteorological expert decides what the most recurrent WRs are, and how a large-scale field (e.g., characterizing a day) must be associated (or attributed) to one of the regimes. The most famous subjective WRs are the Lamb weather types [*Lamb*, 1972] for Great Britain, or Hess and Brezovsky regimes [*Hess and Brezowski*, 1977] for central Europe. It is clear that this approach requires a strong meteorological knowledge of the atmospheric conditions over the region of interest. Moreover, it presents the advantage of allowing the expert to subjectively adapt the WRs according to their use. For example, if the WRs have to be related to local-scale precipitation or to wind, even though they are defined in terms of the same atmospheric variable (e.g., geopotential height), the WRs may be slightly different to take into account local and environmental specificities of the two variables.

[4] The objective approach is based on mathematical clustering methods to automatically group together large-scale atmospheric fields that are close to each other, and to create different clusters for fields that are very different from each other. Many methods have been developed and applied to perform such a task. The most employed is certainly the *k*-means algorithm [*Diday et al.*, 1974], iteratively calculating the center of each cluster (initially randomly chosen) and allocating the data to the cluster whose center is the closest. The *k*-means algorithm has been used, for example, to describe recurrent or quasi-stationary North Atlantic weather regimes [*Michelangeli et al.*, 1995], to characterize climatic trends through circulation types over Europe [*Huth*, 2001], as

[1]Laboratoire des Sciences du Climat et de l'Environnement, Centre d'Étude de Saclay, Gif-sur-Yvette, France.

Copyright 2010 by the American Geophysical Union.
0148-0227/10/2009JD012871

a basis for a heat watch warning system [*Sheridan and Kalkstein*, 1998], to relate extreme temperature and precipitation events to North Atlantic WRs [*Yiou and Nogaj*, 2004], or to study teleconnection patterns [*Cassou*, 2008].

[5] Hierarchical agglomerative clustering (HAC) [*Ward*, 1963] methods have also been quite popular the last decades. From all elements considered as separated clusters, HAC generates a tree, successively grouping two clusters into one until the chosen number of clusters is reached. In climate studies, HAC has been employed, for example, to derive a climatology of severe storms in Virginia [*Davis et al.*, 1993]; to define climate regions in the northern plains [*Bunkers et al.*, 1996]; to short- and medium-range predictability of weather regimes [*Vannitsem*, 2001]; or to identify winter weather regimes for the Pacific–North American sector [*Casola and Wallace*, 2007].

[6] The "expectation-maximization" (EM) method [*Dempster et al.*, 1977] has also shown useful climate applications. EM determines weather regimes through a mixture of Gaussian distribution, where each distribution is statistically associated to a WR. *Smyth et al.* [1999] used EM to define regimes in Northern Hemisphere height fields, while *Gaffney et al.* [2007] took advantage of EM to cluster wintertime extratropical cyclones tracks. EM showed some significant differences in WRs over Eastern US in comparison to *k*-means [*Vrac et al.*, 2007a], and also provided meaningful WRs for precipitation downscaling [*Vrac et al.*, 2007b]. The EM approach has also been extended through a mixture of copula functions applied to determine weather types from vertical atmospheric profiles of humidity and temperature [*Vrac et al.*, 2005].

[7] Artificial neural networks can also be designed to perform clustering and define WRs through the so-called self organizing map (SOM) approach. For example, *Hewitson and Crane* [2002] defined WRs through SOM and employed them as conditional bases to downscale daily precipitation, and *Leloup et al.* [2008] used SOM to compare how different general circulation models (GCM) simulate the spatial characteristics of the twentieth century El Niño-Southern Oscillation (ENSO).

[8] Fuzzy rules are issued from artificial intelligence methods and can be used for clustering by optimizing chosen objective functions. For example, this type of rules was successfully employed by *Pongracz et al.* [2001] to determine monthly patterns of precipitation over Hungary, and by *Bárdossy et al.* [2002] to identify large-scale WRs to condition temperature and precipitation downscaling models.

[9] Some intercomparison frameworks have been built to better understand the differences between WRs obtained from different clustering methods, depending on the region of interest, the atmospheric variables used, etc. [e.g., *Huth*, 1996; *Vrac et al.*, 2007a](or the COST 733 project on "harmonisation and applications of weather type classifications for European regions," http://www.cost733.org/). The main conclusions generally indicate that there is no "best" clustering method, and that the choice largely depends on the region and its specificity, the variable of interest and the goal that has to be reached. However, those frameworks were not intended to compare whether or not the insertion of local information into the clustering method improves the "applicability" of the obtained regimes.

[10] The questions we treat here are: Does the inclusion of observed local information into the clustering of large-scale atmospheric fields help to define weather regimes that are well discriminated in terms of local-scale precipitation characteristics? Does this inclusion perturb/complicate the attribution process when a new day (i.e., a new field) arises and has to be classified? What is the cost of misclassification, for example, when simulating precipitation based on the characteristics conditional on the WRs? This study aims at answering those questions by comparing five different clustering methods corresponding to three objective approaches.

[11] In section 2, the data used in this study are presented, as well as the different methodologies employed to define weather regimes. Those methodologies involve various amounts of local-scale information about precipitation and, therefore, do not generate the same WRs. The clustering methods are applied on the Mediterranean basin data in section 3, and a common number of WRs is determined and fixed to compare the methods. To understand the implication of misclassification of the obtained WRs when used to simulate precipitation (say for new, potentially future, days), pattern attribution performances are performed in section 4 and precipitation cost functions are developed and applied. A summary and a discussion of the results are provided in section 5, while conclusions and some perspectives are presented in section 6.

2. Data and Clustering Methodologies

[12] This study is performed on the French southern Mediterranean region. Indeed, this region presents some geographical specificities accentuating the natural variability of precipitation. The Mediterranean sea with its chiseled coast, the proximity of three chains of mountains (the Alps, the Pyrénées, and the Massif Central), and the very present urbanization strongly enhance the spatial and temporal variability of precipitation in this region. As a consequence, extreme events of precipitation are relatively frequent. This makes French Mediterranean climate a difficult but interesting region to model.

2.1. Local- and Large-Scale Data

[13] Local-scale data correspond to daily times series of precipitation ranging from 1 January 1959 to 31 December 2004 for seven French rain gauges extracted from the "European Climate Assessment & Dataset" (ECA&D) [*Klein Tank et al.*, 2002]. Those rain gauges are located at (1) Marseille, (2) Perpignan, (3) Mont-Aigoual, (4) Le Massegros, (5) Nimes, (6) Orange, and (7) Sète. Large-scale predictors are NCEP/NCAR (National Centers for Environmental Prediction-National Center for Atmospheric Research) reanalysis data [*Kalnay et al.*, 1996] sea level pressure (SLP) and geopotential heights at 500 mbar (Z500), with 2.5° of horizontal resolution. Both variables cover the region [−15°E; 42.5°E] × [27.5°N; 50°N] encircling the Mediterranean Sea, and corresponding to 240 grid cells. This domain is an attempt to visualize the main weather regimes that are typical of the Mediterranean region. Indeed, previous studies [e.g., *Plaut and Simonnet*, 2001] have found relative correlations between South of France and the North Atlantic region, which is much more classically

Table 1. Methods Summarizing the Analysis Procedure[a]

Clustering Methods	PCA	CCA	WRs Construction Scale
CCM	-	WR metric	large and local
k-means	information reduction	-	large
EM	information reduction	-	large
k-means(*w*)	-	WR metric on *w* CVs	large and local
EM(*w*)	-	WR metric on *w* CVs	large and local

[a]Clustering Method indicates the name of the clustering method. PCA and CCA give the type of use of PCA and CCA, respectively (a dash indicates no PCA or no CCA). WRs Construction Scale indicates the scale information used to construct the WR.

studied. Here, the domain is arbitrarily chosen to test if the region loosely encompassing the Mediterranean sea is able to provide valuable large-scale information linked to observed local-scale (potentially extreme) precipitation in the south of France. Note that this region is chosen sufficiently large to partially capture some features of the North Atlantic region, hence keeping a part of the well-known associated variability. We used Z500 to define WRs, like, for example, in work by *Michelangeli et al.* [1995] or *Yiou and Nogaj* [2004]. However, information at much lower altitude has shown useful [e.g., *Vrac et al.*, 2007b] to relate large- and local-scale features, and SLP is generally considered as valuable information for precipitation [e.g., *Busuioc et al.*, 2008]. Indeed, Z500 is a spatially smooth variable and SLP is more sensitive to horography and land properties, and thus, may allow to capture more spatial variability. In the following, only winter months (November–March) are considered, and we removed seasonal cycles and a linear trend to those large-scale predictors.

[14] A few studies suggested that large-scale WRs may be related to local-scale precipitation in the Mediterranean or the Alps [e.g., *Plaut and Simonnet*, 2001] and may also influence heavy precipitation [e.g., *Plaut et al.*, 2001; *Sanchez-Gomez and Terray*, 2005; *Sanchez-Gomez et al.*, 2008]. *Plaut et al.* [2001] looked at circulation regimes conditionally on intense local precipitation. Our study is more general, since, here, all days are considered (i.e., not only those with high precipitation) to investigate the links between WRs obtained from different clustering methods, and their informative content in terms of local precipitation, and particularly extremes. To do so, three approaches are used, comprising five methods: (1) the *k*-means and EM algorithms applied to the main principal components from a principal component analysis (PCA) performed on the NCEP/NCAR reanalyses; (2) the *k*-means and EM algorithms applied to the main canonical variates (associated to reanalyses) from a canonical correlation analysis (CCA) performed between reanalyses and local precipitation (CCA is summarized in section 2.3); and (3) the correlation clustering model (CCM) developed by *Fern et al.* [2005] (detailed in section 2.4). Table 1 summarizes the use of those methods, which are detailed in the following.

2.2. Expectation-Maximization and *k*-Means Algorithms

[15] In those two methods, a PCA is first applied to the large-scale variables (Z500 and SLP), and the first 10 principal components (PCs), corresponding to 95% of variance, are kept. Then, those PCs are used as inputs of the EM and *k*-means algorithms.

[16] In the EM approach, we estimate *f*, the multivariate probability density function (PDF) of the PCs, as a weighted sum (or mixture) of *K* parametric PDFs f_k ($k = 1,\ldots, K$) [*Pearson*, 1894] with parameters $\boldsymbol{\alpha}_k$:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\alpha}_k), \tag{1}$$

where π_k are called the "mixture ratios" and correspond to the prior probability of belonging to component *k*, or the *k*th WR. In this formulation, the *k*th WR, say W_k, is associated with and is actually defined by the *k*th PDF f_k. In this work, we consider that the f_k are Gaussian PDFs. Hence, we deal with a mixture of Gaussians where $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\mu}_k$ a vector of means, and $\boldsymbol{\Sigma}_k$ the variance-covariance matrix of f_k. For a chosen number *K*, the estimation-maximization (EM) algorithm [*Dempster et al.*, 1977; *McLachlan and Peel*, 2000] is used to estimate the parameters of the mixture model. It consists of two successive and iterative steps of expectation and maximization of the so-called complete log likelihood. Various constraints are tested on the variance matrix (it can be spherical, diagonal, or ellipsoidal and with equal or varying volumes) to avoid overfitting (i.e., reduce the number of parameters to be estimated). For the chosen variance model and *K*, WRs are obtained by applying the principle of posterior maximum. For regime W_k,

$$W_k = \left\{ \boldsymbol{x}; \pi_k f_k(\boldsymbol{x}, \boldsymbol{\alpha}_k) \geq \pi_j f_j(\boldsymbol{x}, \boldsymbol{\alpha}_j), \forall j = 1, \ldots, K \right\}. \tag{2}$$

Hence, each $\boldsymbol{x}$ (i.e., each day characterized by its PCs $\boldsymbol{x}$) is attributed to the regime for which the associated model maximizes the posterior probability that *x* belongs to this regime. This approach has already been successfully applied to determine WRs over eastern United States [*Vrac et al.*, 2007a], which are well related to local precipitation [*Vrac et al.*, 2007b].

[17] The *k*-means algorithm is an iterative clustering process. First, a random clustering is performed; that is, every day (and their associated PCs) is randomly assigned to *K* clusters. Then, the iterative process is as follows.

[18] 1. Once all the days have been clustered in the previous step, the center of each cluster (also called "centroid") is calculated. The center is the average of all the days within the cluster; that is, its coordinates are the arithmetic mean for each dimension separately over all the days in the cluster.

[19] 2. Then, the *k*-means algorithm assigns each day to the cluster whose center is the nearest according to the Euclidean distance.

[20] 3. If the obtained clusters are the same as (or, according to a given criterion, not too different from) the clusters at the

previous iteration, then stop: the clusters are the weather regimes. Otherwise, go back to step 1.

[21] This algorithm strongly depends on the initialization step (i.e., initial random clustering) and generally has to be performed several times to retain the resulting regimes maximizing a criterion such as the intraclass variance [e.g., *Michelangeli et al.*, 1995].

[22] Those two methods are applied to PCs associated to large-scale data. Hence, they do not use any local-scale information on precipitation. To incorporate such information, EM and *k*-means are also applied to canonical variates (CVs) associated to large-scale data, obtained from a canonical correlation analysis (CCA) performed between reanalysis data and observed precipitation time series. In the following, such an use of *k*-means and EM will be denoted *k*-means(w) and EM(w). As a reminder, the basics of CCA are provided in section 2.3.

2.3. Basics of Canonical Correlation Analysis

[23] Our goal is to define large-scale recurrent structures that are correlated to local-scale precipitation characteristics. Linear correlations between two data sets can be detected by canonical correlation analysis (CCA) [*Hotelling*, 1936; *Barnett and Preisendorfer*, 1987]. CCA is closely related to the principal component analysis (PCA), which determines linear combinations of the variables in the initial data set to obtain new variables (i.e., principal components denoted PCs) maximizing the variance. In the same way, a CCA computes linear combinations of the variables of two initial data sets to obtain two new sets of variables (i.e., canonical variates denoted CVs) maximizing the correlation between the CVs.

[24] Let $\mathbf{X} = (X_1,\ldots, X_n)$ and $\mathbf{Y} = (Y_1,\ldots, Y_m)$ be two vectors of random variables to be related. In this study, $\mathbf{X}$ has seven dimensions and is associated to local precipitation observed at the seven rain gauges in French Mediterranean region. $\mathbf{Y}$ has 480 dimensions (corresponding to twice the number of NCEP grid cells) and is associated to the large-scale reanalysis data. CCA transforms pairs of original centered data $\mathbf{X}'$ and $\mathbf{Y}'$ into sets of new variables, the canonical variates, v_i and w_i, defined by

$$v_i = \boldsymbol{A}_i^T X' = \sum_{j=1}^{n} a_{ij} X_j', \quad i = 1,\ldots,\min(n,m) \tag{3}$$

$$w_i = \boldsymbol{B}_i^T Y' = \sum_{j=1}^{m} b_{ij} Y_j', \quad i = 1,\ldots,\min(n,m), \tag{4}$$

with the constraint that v_{i+1} and w_{i+1} are not correlated with any previous pair of CVs, and where $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices whose lines $\boldsymbol{A}_i = (a_{i1},\ldots, a_{in})$ and $\boldsymbol{B}_i = (b_{i1},\ldots, b_{im})$ are called "canonical vectors." It is not necessary for the spatial domains associated to X and Y to be the same, and indeed in the applications of CCA appeared in the literature, they are usually different. The number M of pairs of canonical variates that can be extracted from the two data sets is equal to the smaller of the dimension of $\mathbf{X}$ and $\mathbf{Y}$, that is, $M = \min(n, m)$. Moreover, the construction of the CVs ensures that

$$\mathrm{corr}(v_1, w_1) \geq \mathrm{corr}(v_2, w_2) \geq \cdots \geq \mathrm{corr}(v_M, w_M) \tag{5}$$

and that $\mathrm{corr}(v_p, w_q) = r_p$ if $p = q$, and 0 if $p \neq q$. In practice, CCA can be performed according to different approaches (e.g., singular value decomposition, eigendecomposition) that can be retrieved with much more details, for example, from *Wilks* [2006].

2.4. Correlation Clustering Model via Mixture of CCAs

[25] Another clustering approach is now suggested to take even more advantage of the CCA. This approach is based on the method initially developed by *Fern et al.* [2005] to relate vegetation and precipitation. It consists in a mixture model of CCAs. Indeed, while one single CCA allows to detect linear correlation between two data sets, this model introduces nonlinearities through piecewise linear correlations. The main idea of the "correlation clustering mixture" (CCM) is to gather and separate data (here days) in groups (or clusters) that have the "best" CCA models, i.e., with the highest correlation between CVs inside each group (i.e., each WR). Hence, each obtained cluster is characterized by its own CCA model. Note that the CCA models are not obtained a posteriori of the clustering: the clusters are designed to optimize the CCA models.

[26] CCM presented here to obtain K clusters is a reformulation of the method given by *Fern et al.* [2005], adapted to our data. (1) In initialization, the clusters of days are first randomly chosen, i.e., each day is randomly assigned to one of the K groups. (2) For modeling, for each cluster $k = 1,\ldots, K$, a CCA is performed to construct the kth CCA model $CM_k = \{(v_j^k, w_j^k), r_j^k, (\boldsymbol{A}_j^k, \boldsymbol{B}_j^k); j = 1,\ldots, M\}$, corresponding to the M pairs of CVs, the correlation r_j between the jth pair, and the M pairs of canonical vectors (i.e., projection vectors). (3) Next, for assignment, each day is reassigned to a cluster based on its local-scale precipitation ($\mathbf{X}$) and large-scale atmospheric ($\mathbf{Y}$) features and on the K CCA models ($CM_{1,\ldots,K}$). Details about the reassignment process are given in equations (6)–(8). (4) If the assignment has changed (i.e., if the clusters are different) from the previous iteration, go to step 2. Otherwise, stop, and return the K current clusters (which are the WRs) and the associated CCA models. The assignment step is performed as follows: For each cluster k and its CCA model CM_k, a linear regression is modeled for each pair of CVs:

$$\hat{v}_j^k = a_j^k \times w_j^k + b_j^k, \quad j = 1,\ldots, M. \tag{6}$$

Then, for each specific day (characterized by X and Y) and each cluster k, we compute the CVs under CM_k,

$$v_j^k = \boldsymbol{A}_j^k \mathbf{X}'^{\{k\}} \quad \text{and} \quad w_j^k = \boldsymbol{B}_j^k \mathbf{Y}'^{\{k\}}, \quad j = 1,\ldots, M, \tag{7}$$

where $\mathbf{X}'^{\{k\}}$ and $\mathbf{Y}'^{\{k\}}$ are the centered data of the kth cluster, and then compute $\hat{v}_j^k$ the estimate of v_j^k through equation (6), and the weighted error err^k

$$err^k = \sum_{j=1}^{M} \frac{r_j^k}{r_1^k} \times (v_j^k - \hat{v}_j^k)^2, \tag{8}$$

where r_j^k / r_1^k is the weight of the jth error. The day is then assigned to the cluster minimizing err^k. Note that r_j^k (i.e., the correlation between v_j^k and w_j^k) is decreasing while j is increasing. Hence, the weight of the first error is one, and the weights for the others are smaller depending on the correla-

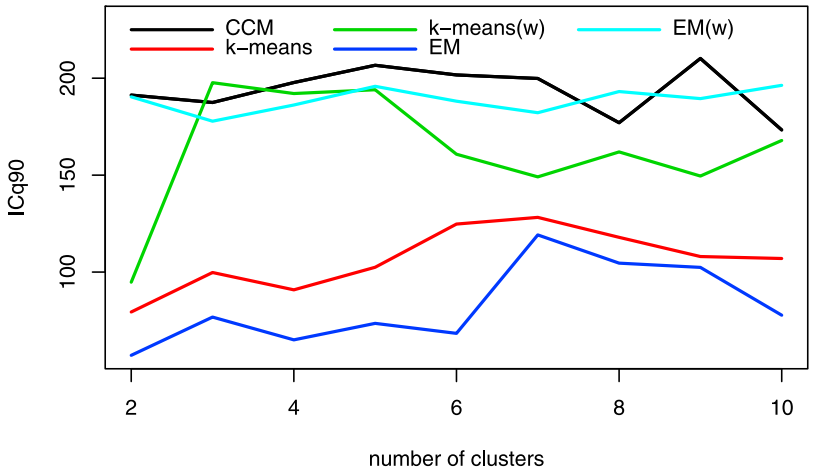

Figure 1. Median ICq_{90} values (computed from the seven stations) for the five tested clustering methods (see Table 1 for labels) and for a number of clusters ranging from 2 to 10.

tions between the other CVs. This allows the weighted error to focus more on the strongly correlated canonical variates.

[27] As a first evaluation, this procedure has been tested on artificial data simulated conditionally on different clusters defined with specified correlations. As concluded by *Fern et al.* [2005], the results (not presented here) have shown that the clustering procedure was able to retrieve the right clusters, with the correct correlation parameters.

2.5. Selection of the Number of Clusters

[28] The correlation clustering detailed in section 2.4 requires fixing of K, the number of clusters. This choice is usually not trivial even though some diagnostic tools may be available depending on the clustering method used (e.g., the Bayesian information criterion for EM developed and used by *Schwarz* [1978] and *Fraley and Raftery* [2002]; the "elbow" criterion for HAC used by [*Vrac et al.*, 2007b]; and the v-fold cross-validation algorithm for k-means as provided by *Breiman et al.* [1984]). However, no method appears as the "best."

[29] Moreover, because CCM is a k-means-like algorithm in the sense that it is based on an iterative process of modeling and assignment, for a given K, the final clusters are sensitive to the initialization step (i.e., the initial clusters). Hence, to ensure "optimal" final clusters, a multistart technique is applied to all clustering methods in the following; that is, CCM, k-means, and EM (applied to PCs or CVs) are run several times, and the clustering maximizing a given criterion is kept. Depending on the goal, different criteria could be used but our goal here is to focus on large-scale atmospheric patterns correlated to local-scale extreme events of precipitation. In that sense, for each given K between 2 and 10, each clustering method is performed 20 times (with 20 random different initializations), and the clustering maximizing the information criterion ICq_{90} is kept. The information criterion IC_a is defined as in work by *Moron et al.* [2008] by

$$IC_a = \sum_{k=1}^{K} |n_{k,a} - (p_a \times n_k)|, \tag{9}$$

where $n_{k,a}$ is the number of days in cluster k that receive a rainfall amount greater than a, p_a is the probability of such rainy days in the whole population, and n_k is the number of days in cluster k. As we focus on intense precipitation events, the rainfall amount a is taken as the 90th quantile (q_{90}) of the strictly positive rainfall amounts for all times series of the seven weather stations altogether. This criterion allows us to quantify the discrimination of the clusters in terms of intense rainfall (>90th quantile). The higher the value of ICq_{90}, the more discriminant the clusters are.

[30] The selection of the final number of clusters is done according to two criteria: the mean correlation (c_m) and the weighted correlation (c_w), both computed from the "optimal" clusters for each number $K = 2,\ldots, 10$ of clusters. For each given K, those correlation criteria are computed by

$$c_m(K) = \frac{1}{K}\sum_{k=1}^{K} r_1^k \tag{10}$$

$$c_w(K) = \frac{1}{N}\sum_{k=1}^{K} (r_1^k \times N_k), \tag{11}$$

where r_1^k is the correlation of the first pair of canonical variates (v_1,w_1) computed from cluster k, N is the total number of days in the whole population, and N_k is the size (i.e., number of days) of cluster k. The final selected number of clusters must yield acceptably high c_m and c_w criteria. While c_m provides a value of correlation averaged over all the clusters, c_w weights each intracluster correlation by the number of days allocated to each cluster. Hence, a small cluster will have a smaller weight in c_w than a large cluster. This can help to choose a number of clusters adapted to the goal to be reached. For example, if the phenomena to be captured by the regimes are not a priori related to the size of the clusters (i.e., if two regimes with different sizes must have an equal importance), c_m will provide useful information. However, if the size of the clusters matters, weights have to be inserted, and c_w is an option to do so. In our case, both criteria are used to enlarge the comparisons of the different clustering results.

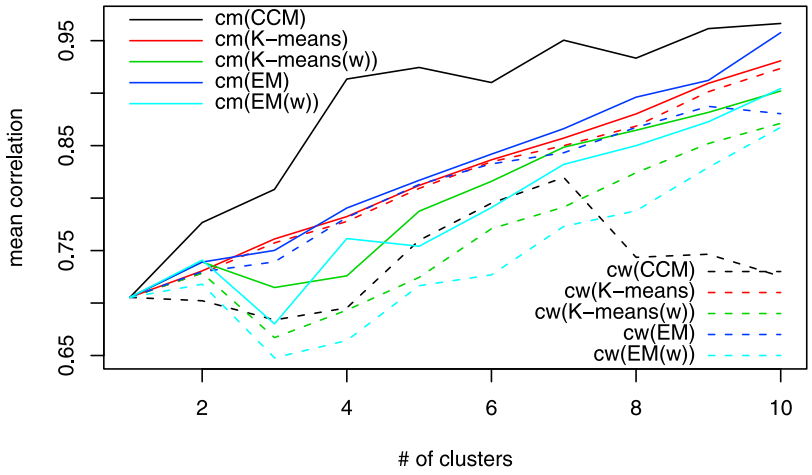

Figure 2. The c_m and c_w correlation criteria for the five tested clustering methods.

[31] The five methods, corresponding to the three approaches, are now applied and compared on real data for the French Mediterranean basin.

3. Results on the Mediterranean Basin

3.1. Cluster Selection

[32] As explained in section 2, for each number of clusters K = 2,…, 10, the five clustering methods are performed 20 times (with 20 different initializations), and only the clustering maximizing ICq_{90} is kept for each K. According to the size of the data set and the number of regimes, the CCM, EM and EM(w) methods can be computationally expensive. However, it has to be noted that a larger number of "runs" does not provide results significantly different (not shown). In other words, 20 runs are enough to capture and estimate the main weather regimes (for each of the five methods). The median ICq_{90} values (computed from the seven stations) from the five clustering methods are shown in Figure 1. We see that, although the curves are relatively flat, the different methods indicate different optimal ICq_{90} values (i.e., different highest values), with some local optima. However, CCM and EM(w) provide higher ICq_{90} values than those given by the other methods. Hence, in general, the discrimination of the intense precipitation events is more efficient with CCM and EM(w).

[33] The selection of the number of clusters is done through the c_m and c_w criteria. Figure 2 presents the results for both criteria and the five clustering methods. The classical k-means applied to PCs does not provide much differences between c_m and c_w. This is due to the fact that this approach tends to provide clusters of equivalent size (i.e., with similar number of days). While the variability of the number of days per pattern is higher, we observe a similar behavior for the correlation curves from EM applied to PCs. For those two methods, c_m and c_w increase almost linearly with the number of clusters. Indeed, for those PCs-based methods, the two indices will have a tendency to increase until the number of clusters is equal to the number of days.

[34] Although k-means(w) and EM(w) present higher ICq_{90} values than classical k-means and EM (Figure 1), they show c_m and c_w values that are lower than those from k-means and EM applied to PCs. Indeed, here, one has to keep in mind that, for each method tested, the retained clustering result is the one maximizing ICq_{90} (for each number K of clusters). Hence, our goal is to obtain clusters with a good "discrimination" of the intense precipitation events. In this exercise, by including local-scale information through w-CVs, EM(w) and k-means(w) are more efficient than classical EM and k-means. However, a consequence is that those two CVs-based methods show a high correlation between large- and local-scale data for clusters related to intense precipitation (and thus containing a small number of days), and a relatively low correlation for the other clusters (with more days). Consequently, the correlation criteria c_m and c_w are lower for EM(w) and k-means(w) than for their classical versions. Note that, according to the use of the resulting clusters, there is a trade-off between the information (IC) and the correlation criteria for the selection of the clustering method.

[35] CCM shows c_m values that are clearly higher than from the other methods. This was expected since CCM looks for clusters optimizing the CCA models and therefore maximizing correlation between v and w CVs inside each cluster. Moreover, CCM has a tendency to provide one relatively large cluster (i.e., with a high number of days, ~3/5 of the data set), and the others usually much smaller (the last 2/5 about equally distributed). Indeed, CCM looks for large-scale structures that are best correlated to intense local-scale observations. For a given number of WRs, once the main extreme events are associated to WRs, the remaining data are grouped together and associated to one large-scale structure. Hence, this cluster gathering about 3/5 of the data set is not totally surprising. The number of days per cluster from seven patterns is given in Table 2. This explains the relatively low c_w values for CCM, since the large cluster is associated to relatively small correlation r between CVs, while the other clusters (that are smaller in population) have much higher r values. For CCM, K = 7 clusters provide a clear optimum for c_w and a local optimum for c_m, which is close to the global optimum between 2 and 10 clusters. For those reasons, and for illustration purposes and understanding of the mechanisms and performances of CCM, the following comparisons of the clustering methods are based on the results obtained for K = 7 clusters from each method.

Table 2. Number of Days per Pattern for Each Method and Corresponding Percentage With Respect to the Whole Population[a]

	Patterns						
	1	2	3	4	5	6	7
CCM	650(9)	385(6)	483(7)	463(7)	4000(58)	413(6)	460(7)
k-means	1050(15)	1084(16)	647(9)	1053(15)	746(11)	1194(17)	1080(16)
EM	1452(21)	930(14)	259(4)	761(11)	1253(18)	1162(17)	1037(15)
k-means(*w*)	1525(22)	1482(22)	466(7)	584(9)	429(6)	1093(16)	1275(19)
EM(*w*)	952(14)	1656(24)	557(8)	561(8)	1366(20)	430(6)	1332(19)

[a]Corresponding percentage is given in brackets.

3.2. Analyses of the Seven-Cluster Results

[36] In the following analyses, the numbering of the obtained patterns is random and depends only on each method. Hence, two clusters with the same numbering but from two different clustering methods do not necessarily represent the same physical pattern.

[37] Table 2 presents the number of days per pattern for each method and the corresponding percentages with respect to the whole population, for the results in seven clusters. Figure 3 characterizes the mean positive precipitation in mm (Figure 3, top), and the probability of rainfall occurrence (Figure 3, bottom), averaged for the seven rain gauges, for each pattern from each clustering method. Figure 3 (top and bottom) allows one to visualize the degree of discrimination of the WRs in terms of mean occurrence and intensity of precipitation.

[38] The daily and monthly seasonal cycles (on November–March) of the frequency of occurrence of each pattern have been calculated for each clustering method. Those (not shown) do not present any signal different from one month to another.

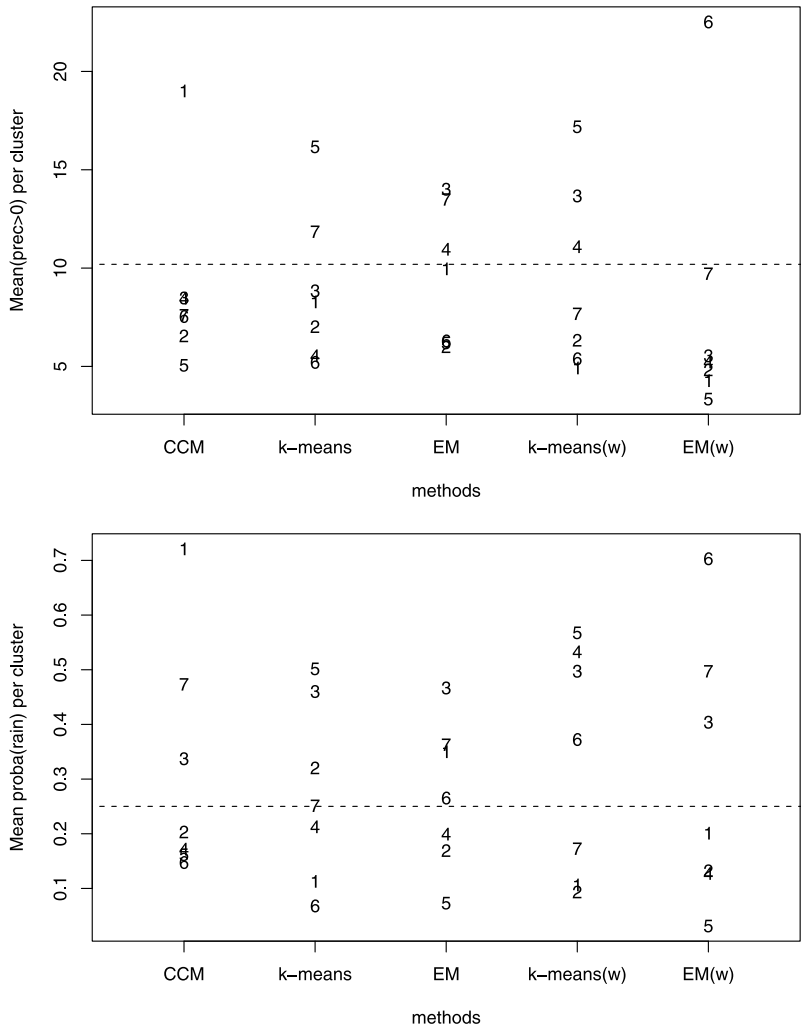

Figure 3. Characterization of the mean (top) positive precipitation in millimeters and (bottom) probability of rainfall occurrence, averaged from the seven rain gauges, for each pattern from each clustering method. Climatological values are indicated by the dashed lines.

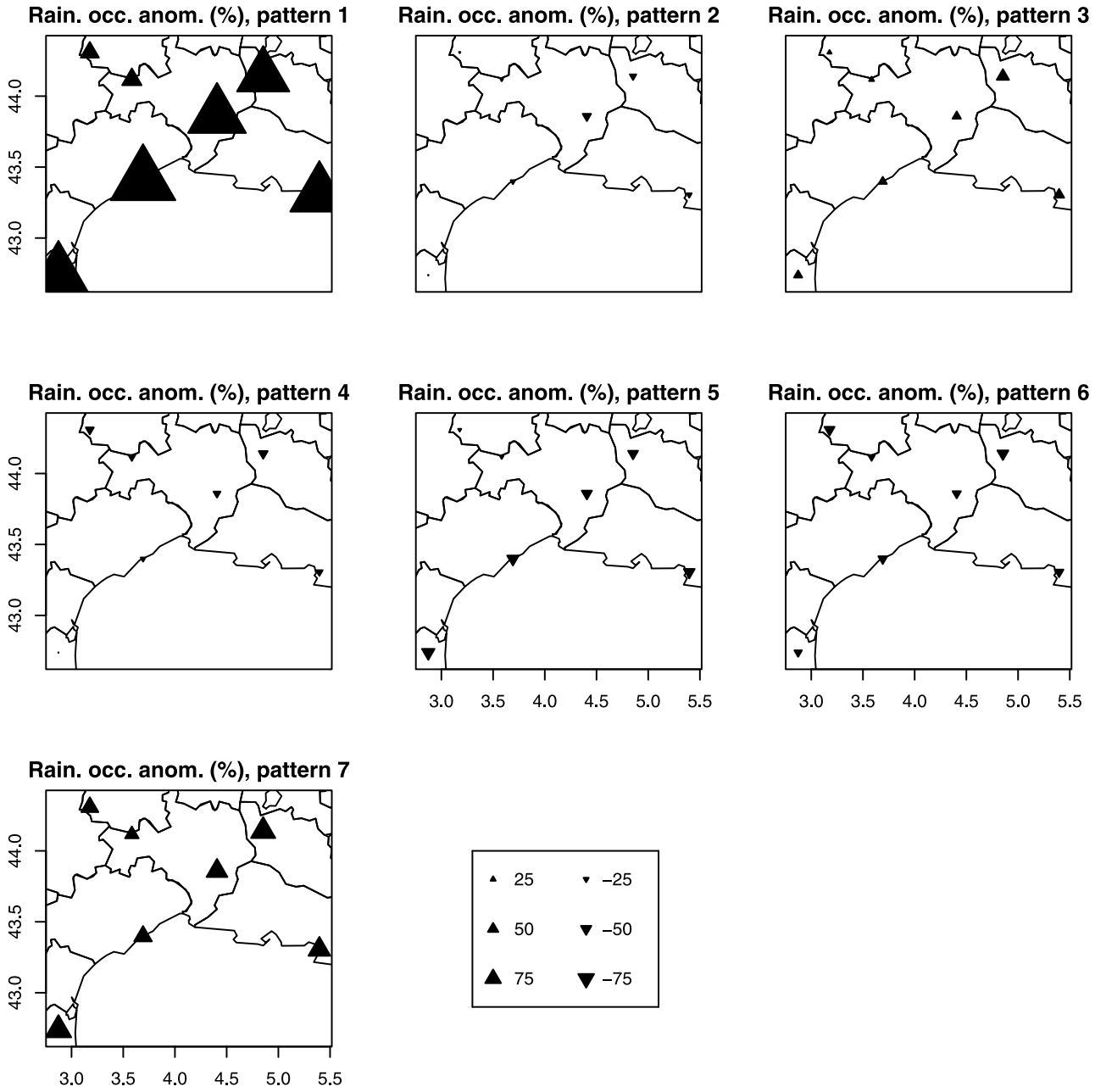

Figure 4. From CCM, spatial distribution of daily rainfall occurrence anomalies (relative to the long-term mean).

All the frequencies of occurrence are relatively flat and mostly represent the proportion of each cluster in the whole population.

[39] As a first evaluation of the correlation between large-scale atmospheric variables (Z500 and SLP) and local-scale precipitation present in each pattern, the correlation values between v_1 and w_1 are summarized in Table 3 (note that the scatterplots of v_1 versus w_1 CVs for each pattern from each clustering method are provided in the auxiliary material).[1] Moreover, to identify the main patterns characterizing local-scale (potentially intense) precipitation events, Figures 4, 5, 6, 7, and 8 show the spatial distribution of daily rainfall occurrence as anomalies relative to the long-term mean for each station and pattern from the same five tested clustering methods. Intensity anomalies (not shown) present similar characteristics, albeit less pronounced.

[40] Although differences of occurrence anomalies exist between patterns, *k*-means and EM applied to PCs (Figures 5 and 6) do not really identify patterns with as strong occurrence anomalies as from the three CVs-based methods. For example, *k*-means(w) defines three patterns (3, 4, and 5) with relatively high occurrence anomalies, and very good correlations between v_1 and w_1 CVs (see Table 3, *k*-means row). Nevertheless, those three patterns capture different spatial distributions. This phenomenon is visible when plotting the

[1] Auxiliary materials are available in the HTML. doi:10.1029/2009JD012871.

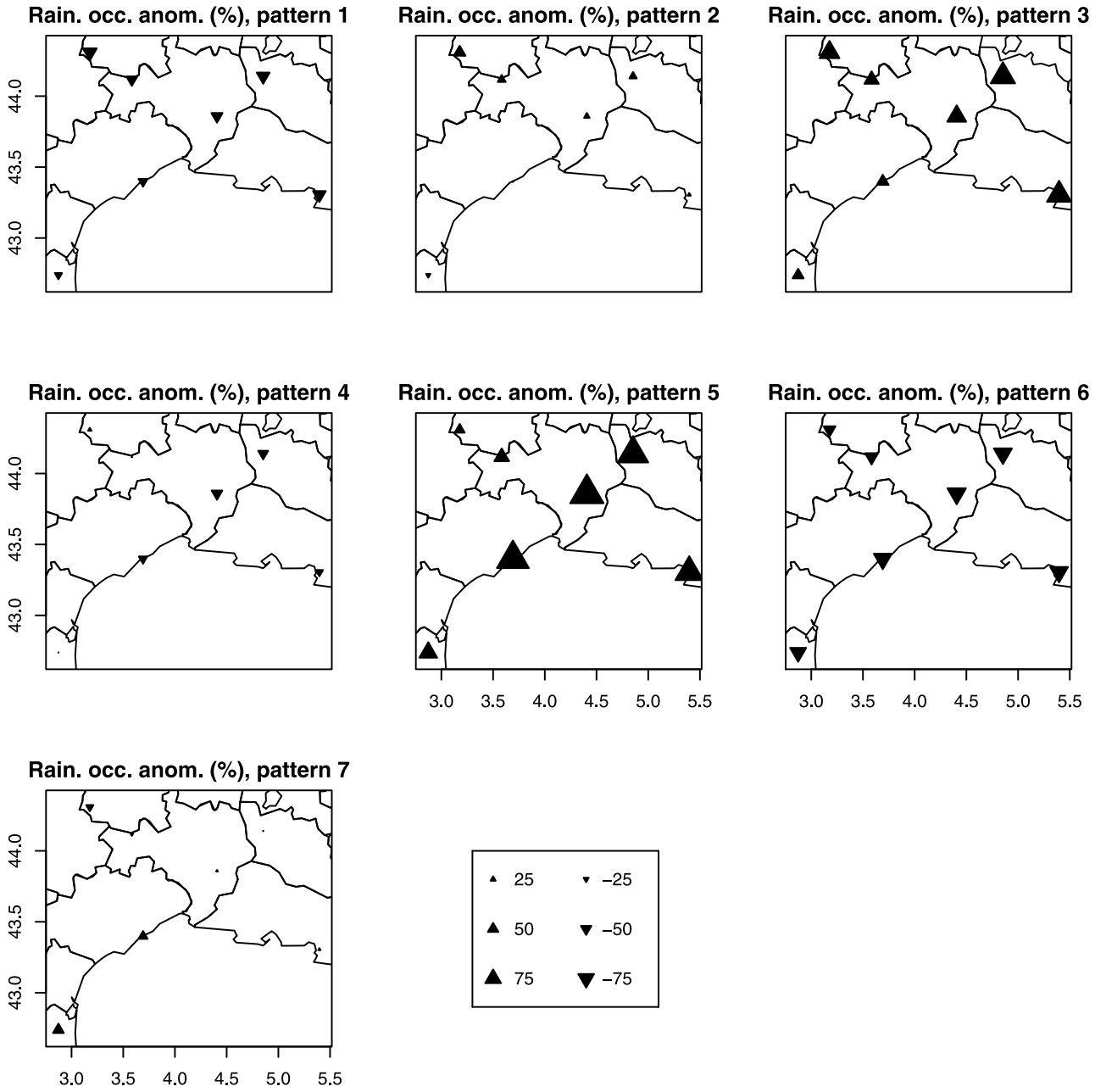

Figure 5. Same as Figure 4 but for *k*-means.

box-and-whisker plots, representing the statistical distribution of the data per cluster and per station through the 25th, 50th, and 75th percentiles, as well as the H and L values, where H = min[75th percentile + 1.5 × interquartile(25th, 75th), max] and L = max[25th percentile - 1.5 × interquartile (25th, 75th), min] (not shown). Moreover, the strongest occurrence anomalies appear in the CCM pattern 1, and in the EM(*w*) pattern 6. Those patterns are very similar to each other in terms of intensity and spatial distribution among the seven stations (compare pattern 1 in Figure 4 and pattern 6 in Figure 8b), as well as in terms of Z500 anomaly patterns over the Mediterranean region presented for those two methods (CCM and EM(*w*)) in Figures 9 and 10, respectively: we can compare pattern 1 in Figure 9 and pattern 6 in Figure 10. A common structure with strong negative Z500 anomalies over Spain characterizes the two patterns. This cyclonic structure is close to the one found by *Plaut and Simonnet* [2001] with an atmospheric circulation clustering only, defined conditionally on intense local precipitation. This type of pattern tends to bring moist air to western Europe (including Spain and France), while Italy, Greece and Romania are relatively dry. SLP maps are not presented but provide equivalent information. The other methods (*k*-means(*w*) and overall *k*-means and EM) do not capture exactly this particular anomaly feature over Spain: it is either less pronounced, larger, or even shifted (not shown). This can explain why only CCM and EM (*w*) defined high precipitation anomalies patterns. Those two methods also present very similar (high) correlations between

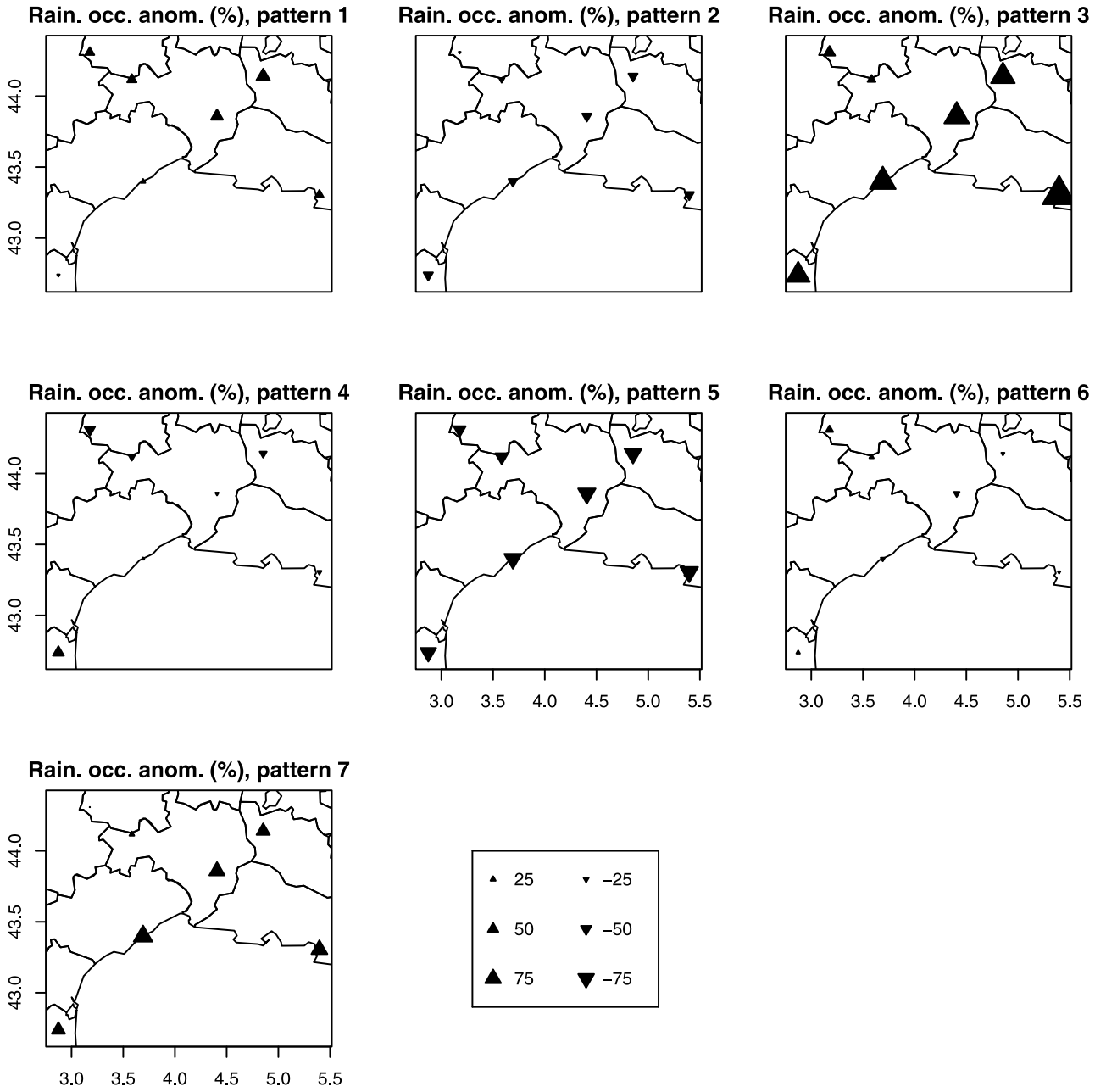

Figure 6. Same as Figure 4 but for EM.

v_1 and w_1: 0.95 for CCM pattern 1 and (very close to) 1 for EM(w) pattern 6. This difference is due to the slightly higher number of days in CCM1 than in EM(w)6 (see Table 2). Hence, in general, it appears that CCM and CVs-based methods (i.e., k-means(w), and EM(w)) are more discriminant, at least in terms of strong (occurrence and intensity) precipitation anomalies, than the classical PCA-based methods. For their extreme patterns identified, those three CCA-based methods also show very high correlation values between v_1 and w_1, ranging from 0.94 to 1. "Classical" k-means and EM applied to PCs do not seem to be as efficient in such a context.

[41] The atmospheric structures (Figures 9 and 10) corresponding to the clusters from CCM and EM(w) are quite different, when seven clusters are imposed for both techniques. The similar features include the atmospheric structure that carries the heaviest precipitation (CCM1 and EM(w)6). The two other common atmospheric structures convey similar rainfall anomalies for both methods (CCM5 and EM(w)1; CCM6 and EM(w)5). This suggests a robust link between precipitation and those three atmospheric circulation patterns. Given the geometry of the domain, the atmospheric patterns are rather different from those obtained from a classification of North Atlantic geopotential data [*Michelangeli et al.*, 1995; *Yiou and Nogaj*, 2004]. In both methods, the blocking regimes (CCM6 and EM(w)5) consistently indicate drier conditions over the South of France, which is consistent with more global

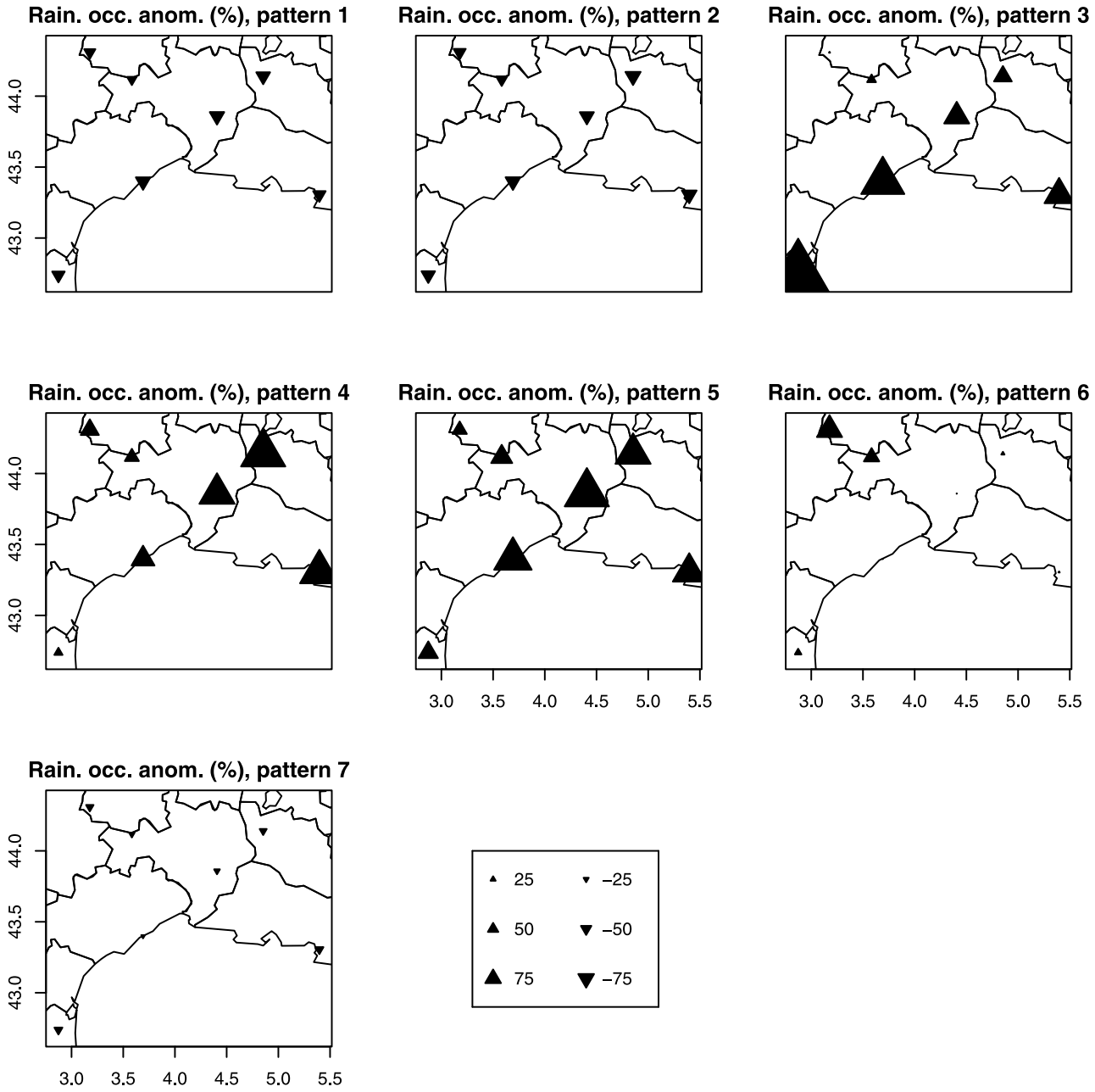

Figure 7. Same as Figure 4 but for *k*-means(*w*).

analyses [*Yiou and Nogaj*, 2004; *Plaut and Simonnet*, 2001; *Sanchez-Gomez and Terray*, 2005].

[42] As a more precise indicator of the "extreme content" captured by the methods, the ratios of the number of days with precipitation exceeding the 97.5th percentile, or exceeding the 99th percentile, with respect to the total number of days exceeding this percentile, are computed for each pattern and station. These two ratios are denoted Rq97.5 and Rq99. For each station and method, Table 4 shows Rq97.5 and Rq99 for the pattern with the highest ratio. For each method, we see some variability from one station to another. A better stability is observed for CCM, in the sense that one single pattern (CCM1) is enough to characterize local extreme precipitation events while (most of) the other methods need more patterns. More generally, except for the large cluster gathering about 3/5 of the data, CCM provides clusters that, by construction, are strongly correlated to intense precipitation.

[43] Moreover, although EM(*w*) seems more unstable than CCM (e.g., the highest ratios belong to more than one pattern, and EM(*w*)6 Rq97.5 and Rq99 are clearly lower than for CCM1), it also seems to behave better than the more classical *k*-means and EM applied to *w* CVs. Hence, it seems that clustering *w* CVs improves the links between large- and local-scale variables in the defined clusters.

[44] Table 5 contains the sum of the ratios from the two best patterns in terms of Rq97.5 and Rq99. CCM remains generally better than *k*-means, EM, and *k*-means(*w*) to capture extremes precipitation for the seven stations, but EM(*w*)

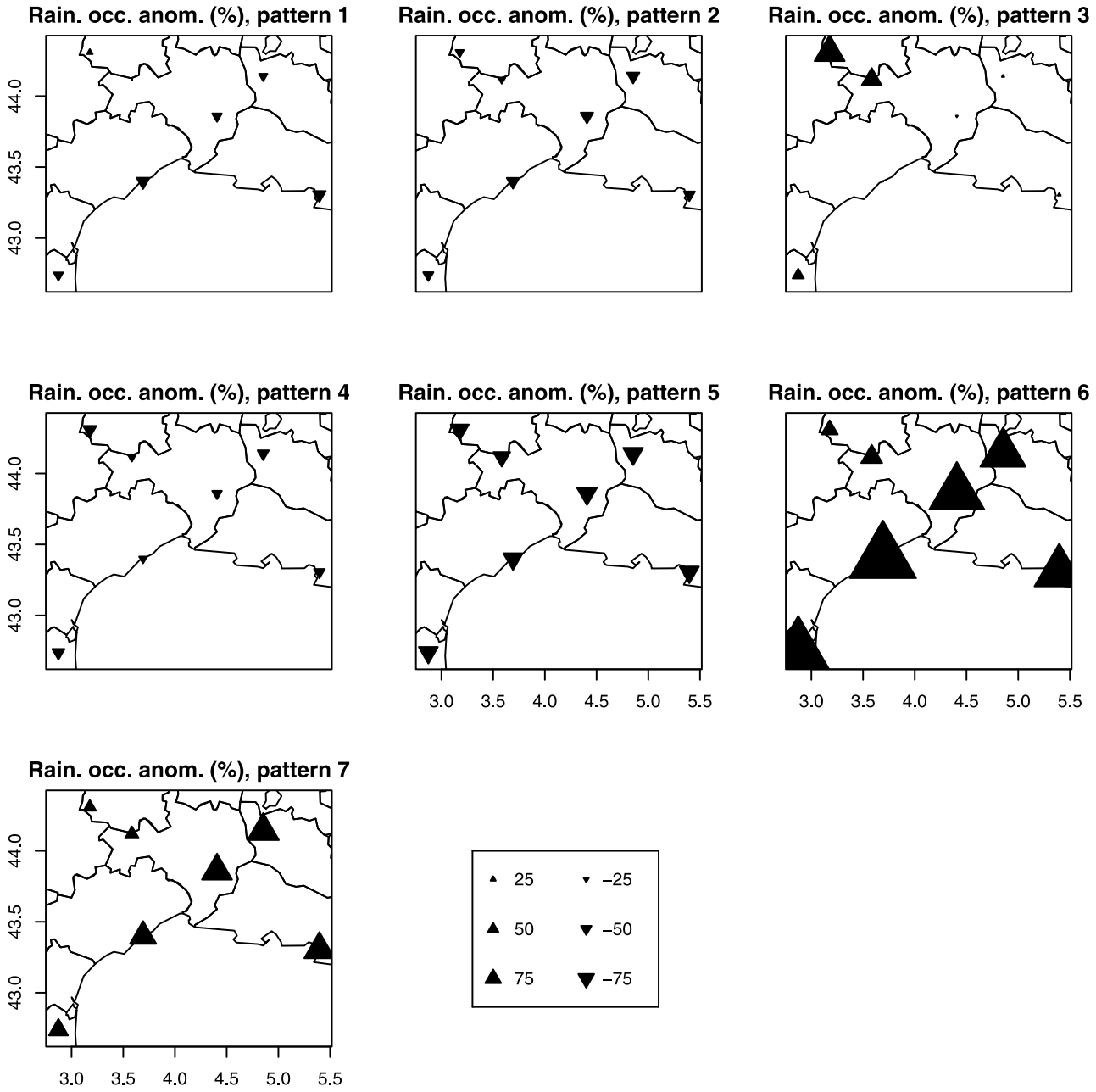

Figure 8. Same as Figure 4 but for EM(w).

becomes this time globally more efficient than any other method. This approach even allows to gather 100% of the extremes for two stations (Mont-Aigoual and Orange) with only two patterns (6 and 7). The k-means(w) results are globally not too far from CCM and EM(w) results, strengthening the idea that clustering w CVs brings more local-scale information to the large-scale patterns.

3.3. Sensitivity to the Size of the Domain: North Atlantic WRs

[45] In order to test the sensitivity of the intercomparison results to the size of the large-scale domain, the five clustering methods have also been performed using the more classical North Atlantic (NA) region defined here as (77.5°W–42.5°E, 22.5°N–70°N), which is relatively similar to the NA region used by *Yiou and Nogaj* [2004]. Indeed, many weather regimes studies over western Europe are based on this type of patterns [e.g., *Yiou and Nogaj*, 2004; *Plaut and Simonnet*, 2001; *Sanchez-Gomez and Terray*, 2005]. For comparison purpose, the NA regimes have been defined based on Z500

Table 3. Correlation Between v_1 and w_1 CVs for Each Clustering Method and Obtained Weather Regime

	WR 1	WR 2	WR 3	WR 4	WR 5	WR 6	WR 7
CCM	0.95	1	1	1	0.7	1	1
k-means	0.83	0.82	0.92	0.85	0.91	0.8	0.86
EM	0.81	0.84	1	0.91	0.78	0.85	0.87
k-means(w)	0.69	0.7	1	0.96	1	0.82	0.77
EM(w)	0.8	0.66	0.97	0.96	0.68	1	0.75

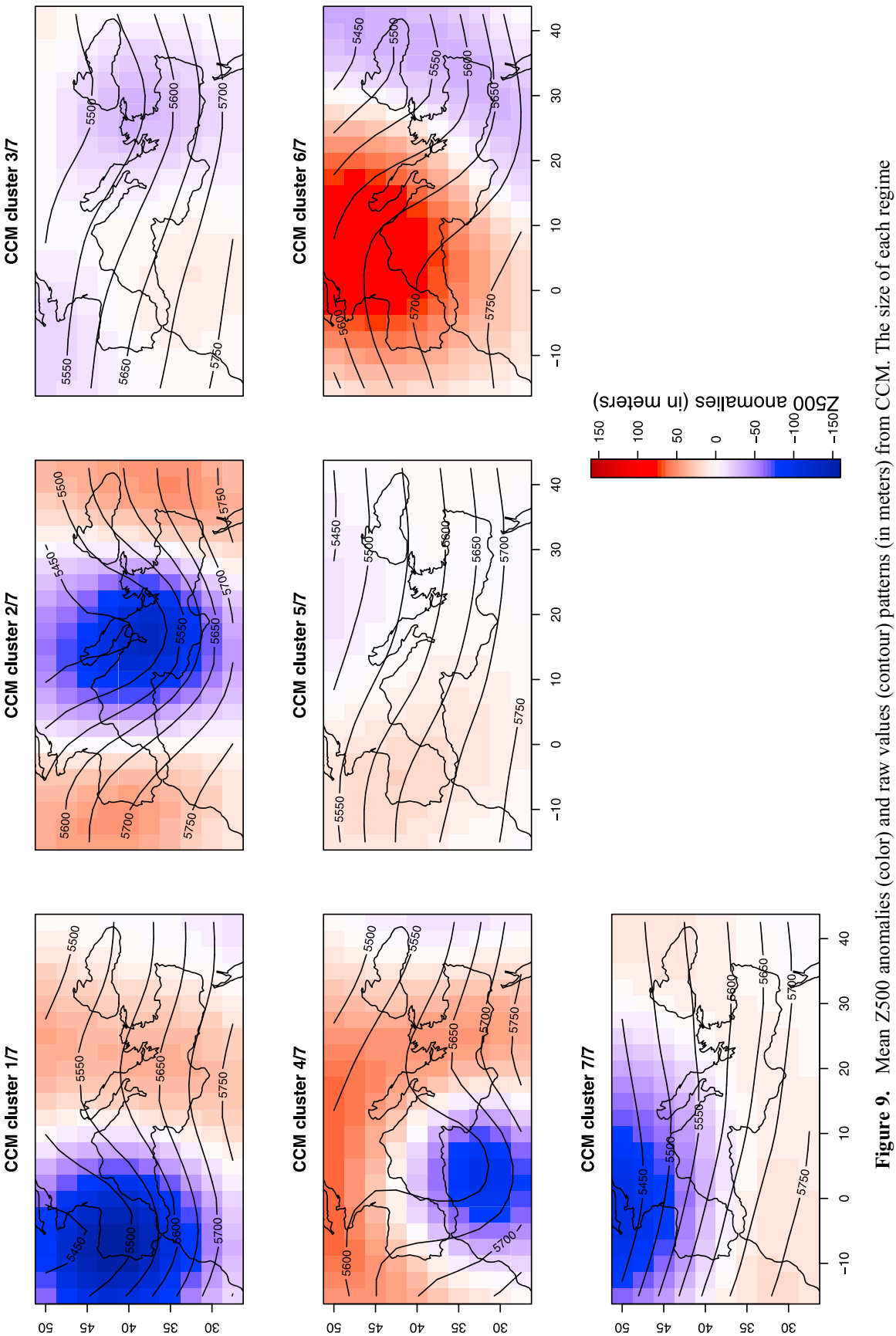

Figure 9. Mean Z500 anomalies (color) and raw values (contour) patterns (in meters) from CCM. The size of each regime is indicated in Table 2.

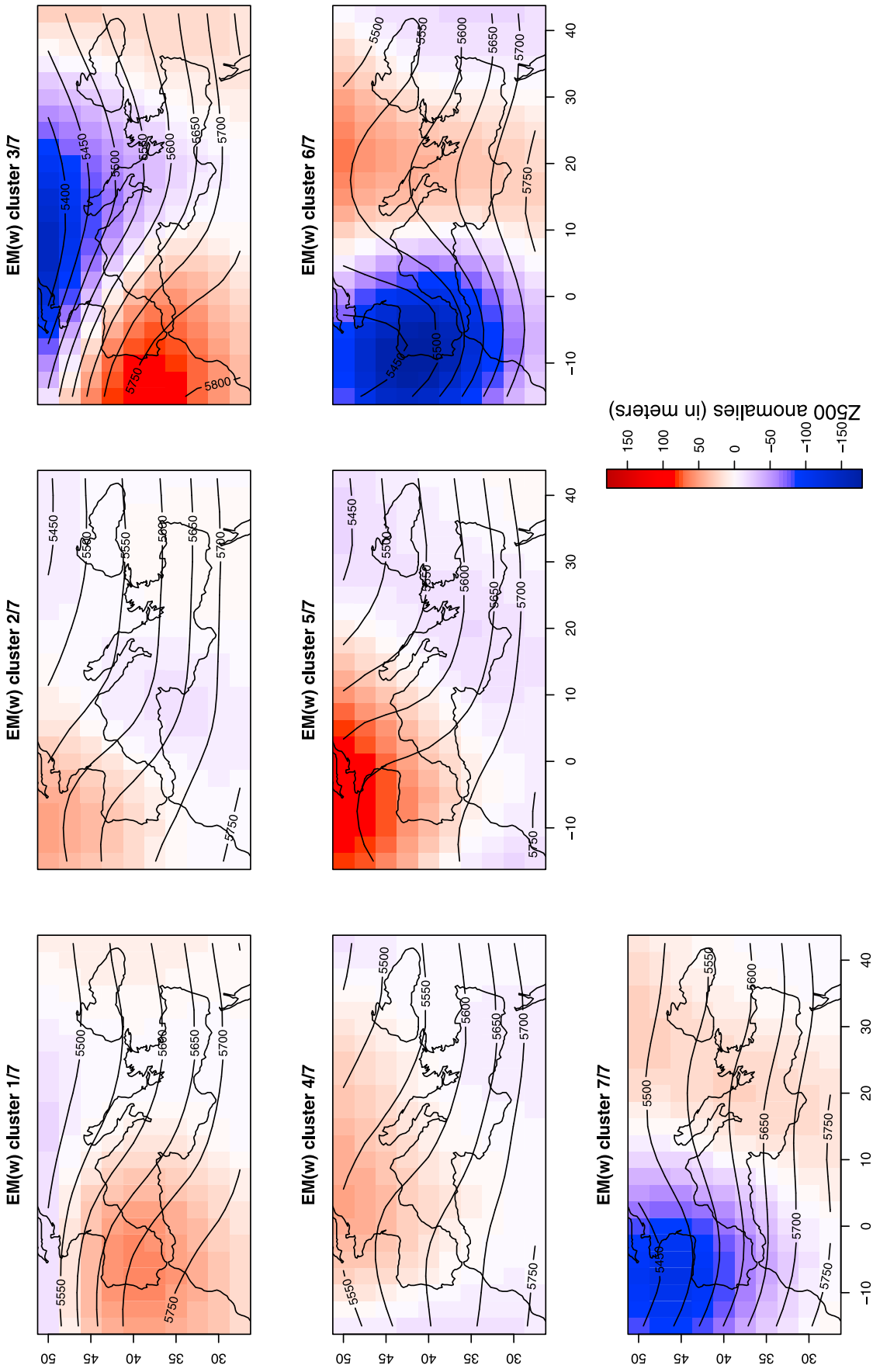

Figure 10. Same as Figure 9 but from EM(w).

Table 4. For Each Pattern and Station, Ratio of Number of Days With Precipitation Exceeding the 97.5th Percentile, or Exceeding the 99th Percentile, Calculated With Respect to the Total Number of Days Exceeding This Percentile[a]

	Pattern						
Stations	1	2	3	4	5	6	7
				CCM			
Rq97.5	**80.2(1)**	**84.9(1)**	72.1(1)	**42.7(1)**	**80.9(1)**	**77.5(1)**	**77(1)**
Rq99	**85.5(1)**	**92.8(1)**	79.7(1)	**56.5(1)**	**85.5(1)**	**81.2(1)**	**84.1(1)**
				k-Means(PCs)			
Rq97.5	39(5)	45.3(7)	72.7(5)	32.6(5)	49.7(5)	42.2(5)	40.2(5)
Rq99	40.6(5)	47.8(7)	81.2(5)	44.9(5)	55.1(5)	47.8(5)	39.1(5))
				EM(PCs)			
Rq97.5	36.6(7)	36.6(7)	47.7(7)	37.6(1)	37.6(7)	37(1)	41.4(7)
Rq99	39.1(7)	40.6(4)	49.3(7)	24.6(7)	30.4(7)	37.7(1)	39.1(7)
				k-Means(w)			
Rq97.5	45.9(4)	73.3(3)	65.7(5)	40.4(6)	33.5(5)	57.2(4)	40.2(3)
Rq99	49.3(4)	85.5(3)	82.6(5)	36.2(5)	37.7(4)	69.6(4)	46.4(3)
				EM(w)			
Rq97.5	47.7(6)	67.4(6)	**74.4(6)**	33.1(3)	57.2(6)	54.3(7)	60.3(6)
Rq99	62.3(6)	87(6)	**85.5(6)**	44.9(6)	69.6(6)	59.4(6)	69.6(6)

[a]For each method and station, only the results from the pattern with the highest ratio are shown. The associated pattern numbering is indicated in brackets. Bold values correspond to maximum value (i.e., "best" method) per station. Rq97.5, ratio of number of days with precipitation exceeding the 97.5th percentile; Rq99, ratio of number of days with precipitation exceeding the 99th percentile.

alone, which is the classical atmospheric variable usually employed, and on SLP and Z500 together, as in the present study. For both approaches, the *k*-means and EM clustering methods have been applied to generate four NA regimes (the usual number) and seven NA regimes for comparison with the previous results of this study. Note that the Mediterranean WRs are not directly compared to those obtained over the North Atlantic as in work by *Michelangeli et al.* [1995]. The five different clustering methods are applied to the North Atlantic region as, for example, in work by *Michelangeli et al.* [1995], before making some intercomparisons of the different results.

[46] We computed the correlation between canonical variates (i.e., corr(v_1, $w1$)) conditionally on each patterns on SLP and Z500 fields. The *k*-means and EM NA results in seven patterns are relatively equivalent to those on the Mediterranean basin, with correlations ranging from 0.8 to 0.94 depending on the regime. For four clusters, the correlations are slightly lower (from 0.75 to 0.88). However, the occurrence and intensity anomalies, as well as the number of days with precipitation exceeding the 90, 97.5, or 99% quantiles, do not show any signal on the NA regimes. Those do not seem to be related in any way to local-scale precipitation observed in South of France at seven rain gauges we considered.

[47] For the NA region, working on Z500 only instead of SLP and Z500 together does not change much the results in terms of relationship between the obtained patterns and the observed precipitation. The atmospheric variability captured from this big NA region does not allow to define distinct local-scale precipitation regimes in South of France.

[48] Moreover, the number of days with precipitation exceeding the 97.5% or the 99% quantile is better when working on the Mediterranean basin. For example, the highest Rq97.5 and Rq99 values per station, for NA patterns, belong to [20%, 33%] and [20%, 36%], respectively (depending on the station), for seven clusters from SLP and Z500) and to [28%,59%] and [30%,63%], respectively, for four clusters. Once more, for the NA patterns, obtained from Z500 rather than SLP and Z500 does not bring different results. Those Rq values (from Z500 only or SLP and Z500) are in the range of (or sometimes lower than) those obtained from *k*-means and EM applied to PCs over the Mediterranean region, which were already surpassed by the CCM patterns (see Table 4). The results are the same in terms of the sum of the best two Rqs (per station) as in Table 5, where CCM and CCA-based methods (EM(*w*) and *k*-means(*w*)) are still clearly better.

Table 5. Same as Table 4 but for the Sum of the Ratios From the Best Two Patterns[a]

	Pattern						
Stations	1	2	3	4	5	6	7
				CCM			
Rq97.5	86(1,5)	89.5(1,4)	89(1,5)	62.4(1,7)	87.3(1,3)	84.4(1,7)	84.5(1,3)
Rq99	89.9(1,5)	97.1(1,6)	92.8(1,5)	71(1,5)	91.3(1,4)	87(1,3)	92.8(1,3)
				k-Means			
Rq97.5	59.9(5.3)	76.7(7,5)	90.1(5,7)	52.2(5,2)	68.8(5,3)	62.4(5,3)	71.3(5,7)
Rq99	62.3(5,3)	78.3(7,5)	95.7(5,7)	60.9(5,2)	72.5(5,3)	68.1(5,3)	76.8(5,7)
				EM			
Rq97.5	66.3(7,1)	70.9(7,4)	83.1(7,1)	57.9(1,6)	66.5(7,1)	65.9(1,7)	63.2(7,1)
Rq99	65.2(7,1)	72.5(4,7)	87(7,1)	56.5(1,7)	56.5(7,1)	62.3(1,7)	59.4(7,4)
				k-Means(w)			
Rq97.5	72.7(4,5)	88.4(3,5)	83.1(5,3)	**65.7(6,5)**	71.1(4,5)	82.1(4,5)	62.1(3,7)
Rq99	78.3(4,5)	97.1(3,5)	92.8(5,3)	50.7(5,3)	71(4,5)	92.8(4,5)	68.1(3,7)
				EM(w)			
Rq97.5	**93.6(6,7)**	**94.2(6,7)**	**97.7(6,7)**	65.2(3,6)	**94.8(6,7)**	**97.1(6,7)**	**93.7(6,7)**
Rq99	**97.1(6,7)**	**98.6(6,7)**	**100(6,7)**	**75.4(6,3)**	**97.1(6,7)**	**100(6,7)**	**97.1(6,7)**

[a]Best two patterns are indicated in brackets.

[49] This brief analysis of sensitivity of the (intercomparison) results to the size of the large-scale domain show that working on this North Atlantic region does not bring improvement compared to the Mediterranean basin in terms of description of local-scale extreme precipitation events for our rain gauges.

3.4. Mean Duration and Persistence Analyses

[50] Mean duration of each cluster has been computed for each method. Note that the five clustering methods compared in this study do not employ any information on temporal sequence of the data. In the present study related to precipitation, this sequence may have importance (e.g., for impact studies). It is found that the computed duration is very different from a method to another, and clearly smaller from the CCA-based methods than from the classical PCA-based ones. Indeed, by averaging the mean durations per method, we obtain: 2.7 days for *k*-means; 2.16 for EM; and 1.55 for CCM, *k*-means(*w*), and EM(*w*) over the Mediterranean region. The latter three methods are in agreement with the mean duration of the events exceeding the 97.5% or the 99% percentile, which are 1.2 and 1.1 days, respectively, from all stations averaged. For the North Atlantic region, in general, the mean duration of each pattern varies between 3 and 6 days (depending on the variable and the number of clusters), which is much higher than the mean durations over the Mediterranean basin.

[51] Moreover, a simple persistence analysis has been performed, by redoing the same analyses as previously and redoing Figures 1–10 (e.g., correlation(v_1, w_1), occurrence anomalies, composite maps) only for days in cluster spells persisting at least three days over the Mediterranean basin. When computing the *v* and *w* CVs per cluster from these days, the correlation between v_1 and w_1 improves for all method and pattern (with respect to values given in Table 3), with different intensities according to the WR and method (not shown). For the occurrence anomaly maps, the (at least) 3 day persistence analysis per method shows the exact same structures as on unconditional Figures 4–8 but with slightly higher values (not shown). However, the composite maps (not shown) of clusters persisting more than three days do not show any difference with the unconditional maps presented for example on Figures 9 and 10.

[52] Those complementary remarks about persistence strengthen the main results brought by the previous unconditional analyses. We note that persistence is certainly specific to the region and variables studied. The present results concern cluster persistence of at least three days but can be different for longer persistence (e.g., 4 or 5 days). However, those further analyses are out of the range of the present study and will be carried out in a future study.

4. Pattern Attribution Performances and Precipitation Implications

[53] The next question that we are trying to answer here is the following: If the large-scale characteristics of a new day are now available, can we use them to retrieve the pattern where this new day should belong to? This is an important question when working on a downscaling (weather typing) context or more generally in a modeling context conditionally on patterns, such as in projection of climate change. However, in this study, this question is difficult to answer because some of our patterns (CCM, *k*-means(*w*), and EM(*w*)) are not only defined in terms of large-scale variables but also based on local-scale observed precipitation. Yet, the latter is generally not available: for example, in downscaling of climate scenarios for the end of the century, observations will only be available at the end of the century by definition. Thus, only the large-scale information (that can be provided by General circulation models (GCMs) for instance) can be used to associate a new day to one given pattern.

[54] Moreover, knowing the error of pattern attribution for each method, what is the cost in terms of rainfall occurrence probability and rainfall intensity? Indeed, even though the new day is associated to a "wrong" pattern, does this error make a strong difference for local-scale characteristics of precipitation? How much?

[55] To answer those questions, the capability of retrieving the right patterns from each clustering method is studied here, and some associated precipitation cost functions are developed. For each clustering method, the clusters are the same as in sections 2 and 3; that is, they were defined for the whole time period 1959–2004. Then, two time periods are defined: 1959–1989 (learning period) from which different attribution methods (i.e., supervised classification methods) will be calibrated based on the daily sequence of patterns and on the large-scale information; and 1990–2004 (projection period) whose the large-scale data are used in the different attribution methods to associate every day to one pattern. Hence, the new sequence of patterns, obtained from each couple (clustering method, attribution method), can be compared to the already known sequence for 1990–2004. Note that although the patterns are initially defined for the whole 1959–2004 period, neither local-scale observations nor 1990–2004 (large- or local-scale) data are used in the learning step; moreover, we emphasize that no local-scale observations are used in the projection step. In this evaluation of the pattern attribution performances, nine different attribution methods are tested: (1) the Euclidean distance on raw atmospheric data (EA), where each day is allocated to the pattern whose centroid is the closest with respect to the Euclidean distance; (2) the Euclidean distance on the *w* canonical variates defined from the whole (i.e., disregarding the patterns) 1959–1989 time period (Ew1), which is same as the EA method but with centroids defined on those *w*-CVs; (3) the Euclidean distance on the conditional *w* canonical variates (Ew2), which is the same as Ew1 but with the *w* CVs, used to compute the centroids, defined separately from the different 1959–1989 patterns; (4) the "Classification And Regression Trees" (CART) method [*Breiman et al.*, 1984] applied to the raw atmospheric data (CART.A), a method hat iteratively splits the initial data set (i.e., Z500 and SLP) into two data subsets (and so on) in order to maximize the so-called "impurity" criterion (also called Gini index) at each step, and explain the clustering provided as input (see *Breiman et al.* [1984] for more technical details); (5) the CART method applied to the *w*-CVs defined from the 1959–1989 time period (CART.w), which is the same as CART.A (above) but where the variables to be split to explain the clustering are now the *w*-CVs; (6) the CART method applied to both the atmospheric data and the *w*-CVs (CART.A.and.w), which is same as CART.A but where the variables to be split are both the raw atmospheric

Table 6. Results of Bad Classification per Clustering and Attribution Method[a]

	EA	Ew1	Ew2	CART.A	CART.w	CART.A.and.w	knnA10	MM	MMw
CCM	88.1	40.5	79.3	42.3	40.8	38.6	37.5	**37.4**	38.2
k-means	78	76.5	66.9	18.9	69.3	19.4	16.2	**15.8**	64.7
EM	74.5	72.9	66.8	39.1	68.7	38.4	26.7	**22.6**	63.9
k-means(*w*)	84.2	73.9	**30.6**	63.6	37.9	41.4	62.3	56.7	31.1
EM(*w*)	87.1	72.3	40.5	59.8	37.2	41.7	57.8	55.7	**28**

[a]Results are given as percent. Minimum values are in bold.

data (Z500 and SLP) and the *w*-CVs; (7) the 10-nearest neighbors method (knnA10), in which the 10 days whose large-scale atmospheric situations are the closest (in terms of Euclidean distance) to the day to attribute are determined, and then the day is attributed to the majority cluster within the 10 days [e.g., *Duda et al.*, 2001; *Toth*, 1991]; (8) a mixture model (MM) attribution method applied to the main principal components (≈95% of variance) of the atmospheric data, in which a Gaussian PDF is first determined for each 1959–1989 pattern, and then, for each day in 1990–2004, its PCs are calculated (by projecting the large-scale data onto the factorial space) and applied to each PDF (the pattern selected for this day is the one for which the PDF is the highest); and (9) a mixture model attribution method applied to the *w*-CVs defined from the 1959–1989 time period (MMw), which is the same as the MM approach but on the *w*-CVs instead of the PCs. Hence, for each pair of methods (clustering and attribution), the percentage of bad classification (PBC) associated to 1990–2004 is computed. The results are presented in Table 6 where the best (i.e., minimum) PBC values are in bold. Classical *k*-means and EM approaches (i.e., applied to PCs) provide the smallest PBCs (15.8 and 22.6%), while CCM shows the highest (37% with MM). This was expected since the classical *k*-means and EM do not use at all the local-scale observed information, and thus, are purely defined in terms of large-scale features. On the opposite, CCM, *k*-means(*w*), and EM(*w*) were determined by incorporating some local characteristics. Hence, because the attribution methods are not allowed to use any local variables (neither in learning nor in projection), it is more difficult to retrieve the correct sequence of daily patterns from CVs-based results than from PCs-based ones. However, to understand how much these errors in classification of new days "cost" in terms of local modeling of precipitation, three cost functions have been developed.

[56] The first one characterizes the error made if we want to simulate precipitation at station s based on the mean characteristics of the cluster associated to a given day d. Indeed, d can be projected onto a "false" cluster with a different mean precipitation. This cost function is denoted daily intensity (DI) error and depends on s and d through

$$DI(s,d) = |\text{mean}(s, C_{\text{proj}}(d)) - Y(s,d)| - |\text{mean}(s, C_{\text{real}}(d)) - Y(s,d)|, \quad (12)$$

where $C_{\text{proj}}(d)$ is the cluster (correctly or incorrectly) associated to day d, $C_{\text{real}}(d)$ is the cluster where d should be projected to, mean(s, C) is the mean precipitation of station s for 1959–1989 days in cluster C, and $Y(s, d)$ is the precipitation observed at station s and for day d. The DI cost function allows us to quantify the difference between the error made in simulating a mean precipitation when we know the correct cluster ($|\text{mean}(s, C_{\text{real}}(d)) - Y(s,d)|$), and the error made when simulating a mean precipitation when we project d to a given (potentially wrong) cluster ($|\text{mean}(s, C_{\text{proj}}(d)) - Y(s,d)|$). Hence, this cost function is useful in a modeling context to evaluate the mean errors of simulations due to wrong association to the clusters. To summarize this information per station, a mean DI (MDI) cost function is used,

$$\text{MDI}(s) = \frac{1}{N}\sum_{d=1}^{N} DI(s,d), \quad (13)$$

where N is the number of winter days for 1990–2004. A regional MDI (RMDI) is also calculated to obtain one value for all the stations,

$$\text{RMDI} = \frac{1}{S}\sum_{s=1}^{S} \text{MDI}(s), \quad (14)$$

where S is the number of stations ($S = 7$ in this study). The MDI results are presented in the maps shown in Figure 11 for each clustering method and for the best attribution methods detailed in Table 6. By definition in equations (12)–(13) of MDI, an upward triangle in Figure 11 indicates a bigger difference between the observed precipitation and the mean precipitation of the associated clusters than between the observed precipitation and the mean precipitation of the clusters we should have. Thus, an upward (downward) triangle corresponds to an overestimation (underestimation) of the mean precipitation to be simulated in a modeling context. From Figure 11, we see that *k*-means and EM MDIs are relatively equivalent (both in magnitude and in the spatial structure) and better (i.e., smaller in absolute values) than the MDIs provided by the three CVs-based methods. Visually, the latter give the impression of an overestimation of the mean precipitation, which is reflected by the RMDI values (between 0.019 and 0.041 for the CVs-based methods, and around -0.01, slightly negative, for classical PCs-based ones). Those values also indicate that the attribution errors in CCM patterns induce a smaller regional cost in daily mean precipitation (RMDI = 0.019) than the attribution errors to patterns determined by *k*-means(*w*) and EM(*w*) (RMDI = 0.041). Note also that EM(*w*) and *k*-means(*w*) present a very similar structure, while CCM shows some differences. In equation (12), mean(s, $C_{\text{real}}(d)$) and mean(s, $C_{\text{proj}}(d)$) are here calculated with respect to the 1959–1989 precipitation data. Note that they also could have been calculated with respect to the 1990–2004 data. The latter method would allow to take into account the potential change of the mean precipitation of each station per cluster. However, the results of this approach (not shown) show similar structures and magnitudes of the DI and MDI

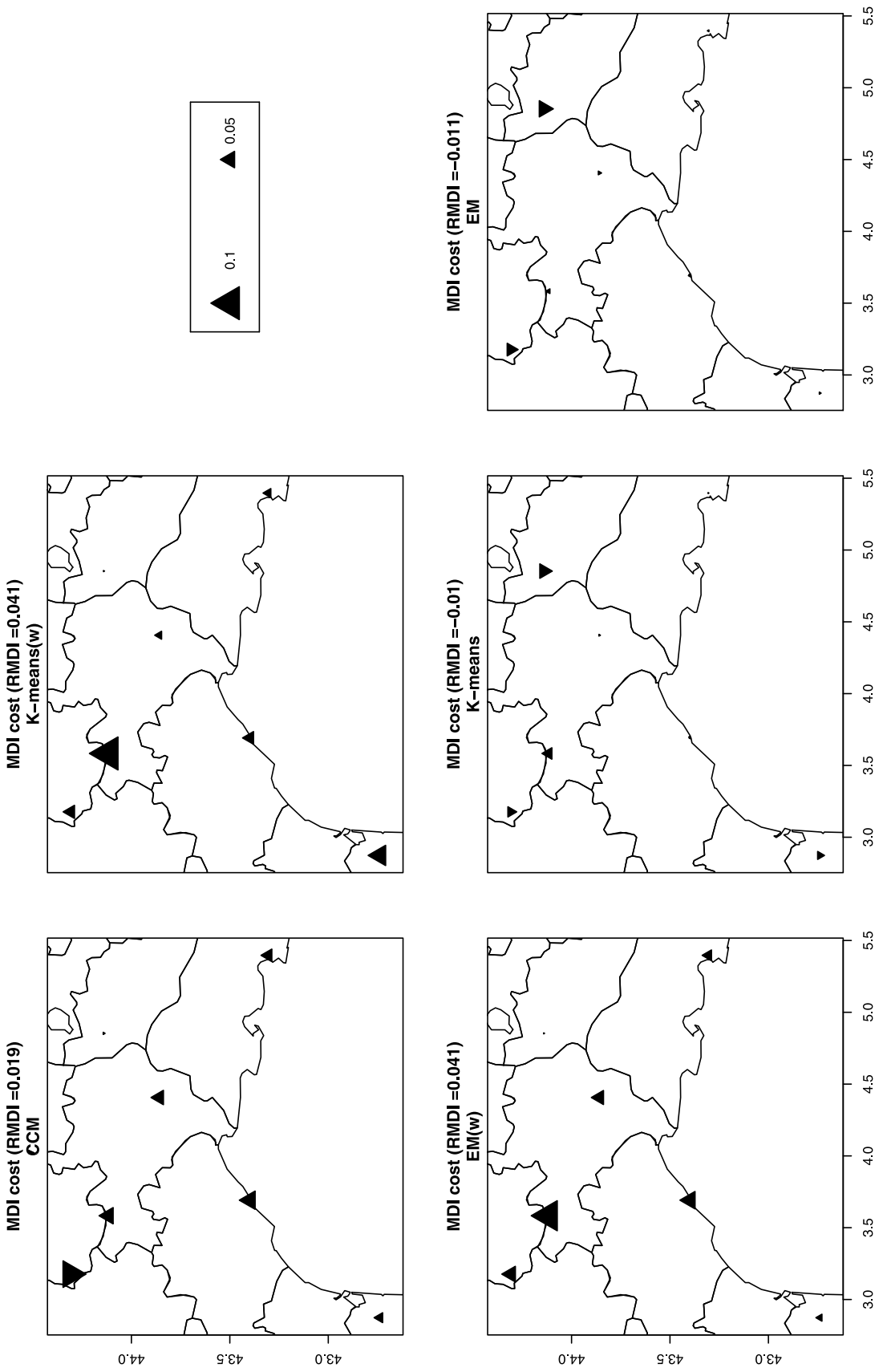

Figure 11. Mean daily intensity (MDI) errors, as defined in equation (13), for each station and each clustering method and for the best attribution methods detailed in Table 6. The global MDI (GMDI; see equation (14)) values are indicated in brackets.

values, meaning that the clusters are consistent between learning and projection periods.

[57] Another cost function is developed to evaluate the global error made in the rain occurrence probability due to misclassification. It corresponds to the sum (over the K clusters) of differences between the probability of rain occurrence in "real" cluster k (i.e., group of days we should retrieve), denoted $\Pr(\text{occ}|C_{k,\text{real}})$, and the probability in "projected" cluster k (i.e., group of days we actually put in cluster k by projection), denoted $\Pr(\text{occ}|C_{k,\text{proj}})$. Each probability is weighted to take into account the number of days in each cluster. This weighted global probability (WGP) cost function depends on the stations s through

$$\text{WGP}(s) = \frac{1}{N}\sum_{k=1}^{K}\left[(\text{Card}(C_{k,\text{proj}}) \times \Pr(\text{occ}|C_{k,\text{proj}})) - (\text{Card}(C_{k,\text{real}}) \times \Pr(\text{occ}|C_{k,\text{real}}))\right], \quad (15)$$

where $\text{Card}(C)$ is the cardinality of cluster C, i.e., the number of days within this cluster. The WGP results are presented in Figure 12, where a downward triangle corresponds to underestimation of the probability of rain occurrence. Overall, we see that all misclassification in patterns from any clustering method provide underestimation of those probabilities. However, if k-means and EM applied either to PCs or to CVs show equivalently small (i.e., good) values of WGP, the misclassification of CCM patterns implies a relatively high cost in occurrence probabilities, much bigger than for the other clustering methods.

[58] However, WGP does not characterize the cost in intensity of rainfall. This is performed by the conditional weighted global log-intensity (CWGLI) cost function defined as

$$\text{CWGLI}(s) = \frac{1}{N}\sum_{k=1}^{K}[(\text{Card}(C^*_{k,\text{proj}}) \times \log(\text{mean}(s, C^*_{k,\text{proj}}))) - (\text{Card}(C^*_{k,\text{real}}) \times \log(\text{mean}(s, C^*_{k,\text{real}})))], \quad (16)$$

where $\text{mean}(s, C^*_k)$ corresponds to the mean precipitation value for station s calculated only from the days with positive intensity of rain in cluster C_k, and $C_{k,\text{real}}$ and $C_{k,\text{proj}}$ are the groups of days that we should retrieve and the groups of days that we actually retrieve by projection, respectively, for cluster k. The maps of the CWGLI results are presented for each clustering method and their associated "optimal" attribution method in Figure 13. Upward triangles correspond to overestimation of log precipitation due to misclassification. In terms of "ranking" of the clustering methods based on the CWGLI costs, we retrieve about the same results as for WGP: k-means and EM applied to PCs show very small CWGLI values, bigger (but still good) values when applied to CVs, whereas CCM shows higher (mostly positive) CWGLI costs. Nevertheless, note that WGP and CWGLI values are globally of opposite signs, meaning that the potential underestimation of the probabilities of rain occurrence (i.e., downward WGP triangles) is generally counterbalanced by an overestimation of the rainfall intensities (i.e., upward CWGLI triangles).

5. Summary, Discussion, and Recommendations

[59] The results presented on MDI, WGP, and CWGLI and the differences noted between clustering methods are explicable by the content of information carried out by each clustering result. Indeed, if the CVs-based methods generally discriminate more the local-scale precipitation characteristics (this is due to the incorporation of precipitation information through the CCA), the PCs-based methods provide patterns less "separated" in terms of local precipitation (this is true, for example, for the extremes as shown in Tables 4 and 5. In other words, the features of precipitation are more similar from one PCs-based pattern to another than from one CVs-based pattern to another. Consequently, for a given misclassification (i.e., for the same PBC), the costs induced in terms of rain probability or intensity will be higher for "informative" patterns (i.e., from CVs-based methods) than for "similar" patterns (i.e., from PCs-based methods). Note that only seven rain gauges were used in this study. Increasing this number could provide slightly different patterns, leading also to different precipitation costs. However, one can remark that, technically, it does not matter if there is any correlation among the rain gauges. Indeed, strong correlation between rain gauges would be captured in the canonical variates resulting from the CCA. Hence, the CVs-based methods (EM(w), k-means(w), and CCM) would need fewer CVs to capture the main weather regimes associated to local-scale precipitation events. For the PCs-based methods (k-means and EM), that would make no change at all since those methods only work on large-scale atmospheric data. A similar remark holds for a greater number of stations. Indeed, more stations would imply more "correlated" data. Hence, if the "added" rain gauges are strongly correlated to the seven existing stations, the number of CVs resulting from the CCA would not increase too much. Hence, although the CCA based on those seven gauges captures a subsample of the "complete" spatial rainfall variability of the region, and a subsample of the "complete" correlation between the regional scale of interest and the large-scale atmospheric data, this "partial" variability/correlation would be present in an ideal complete local-scale data set. However, in practice, geographical locations of the stations may have importance. For example, station 4 (Le Massegros) shows extreme values that are not fully explained by any of the clusters. Although this result can come from very local-scale processes and effects (i.e., implying a lack of correlation between the large-scale variables (SLP, Z500) and intense precipitation at this rain gauge), this may also be due to a location too northern from the rest of the stations.

[60] If NCEP/NCAR data have been employed in this study, using ECMWF ERA-40 reanalyses [*Uppala et al.*, 2005] could lead to slightly different results, with potentially finer spatial structures due to the higher spatial resolution. However, this higher resolution would imply more statistical dimensions (i.e., variables). Hence, the attribution methods could then have more difficulties to retrieve the "correct" weather regimes. Moreover, ERA-40 reanalyses span a shorter time period, and end in 2002, which makes them unsuitable for potential operational seasonal prediction. Besides, NCEP/NCAR reanalyses have a spatial resolution closer to that of many of the "traditional" General Circulation Models (GCMs). Hence, the NCEP/NCAR resolution allows one to deal with technical conditions more similar to those that we would get when working with weather regimes in a future climate change context.

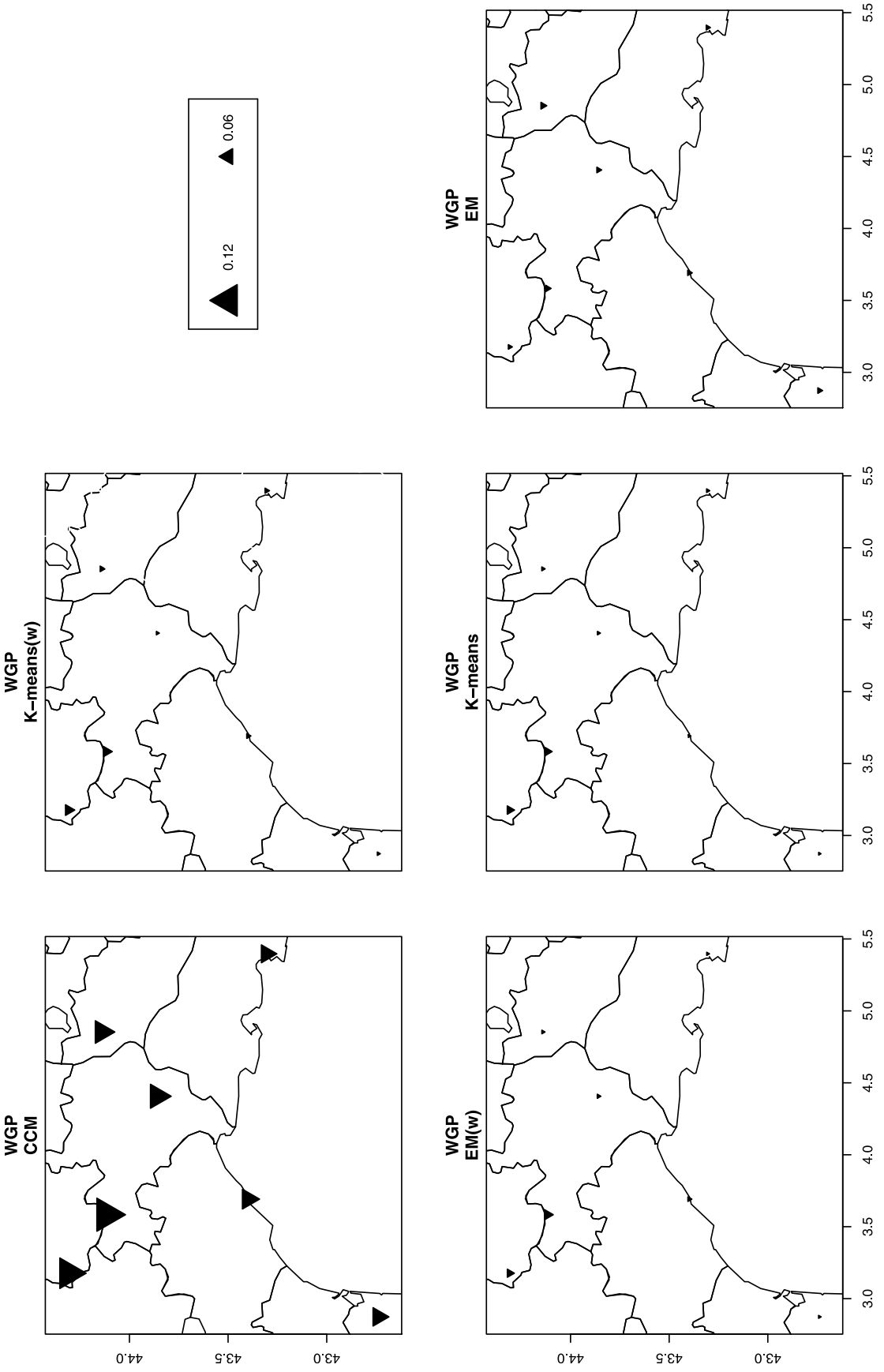

Figure 12. Weighted global probability (WGP) of rain occurrence costs induced by misclassification, as defined in equation (15).

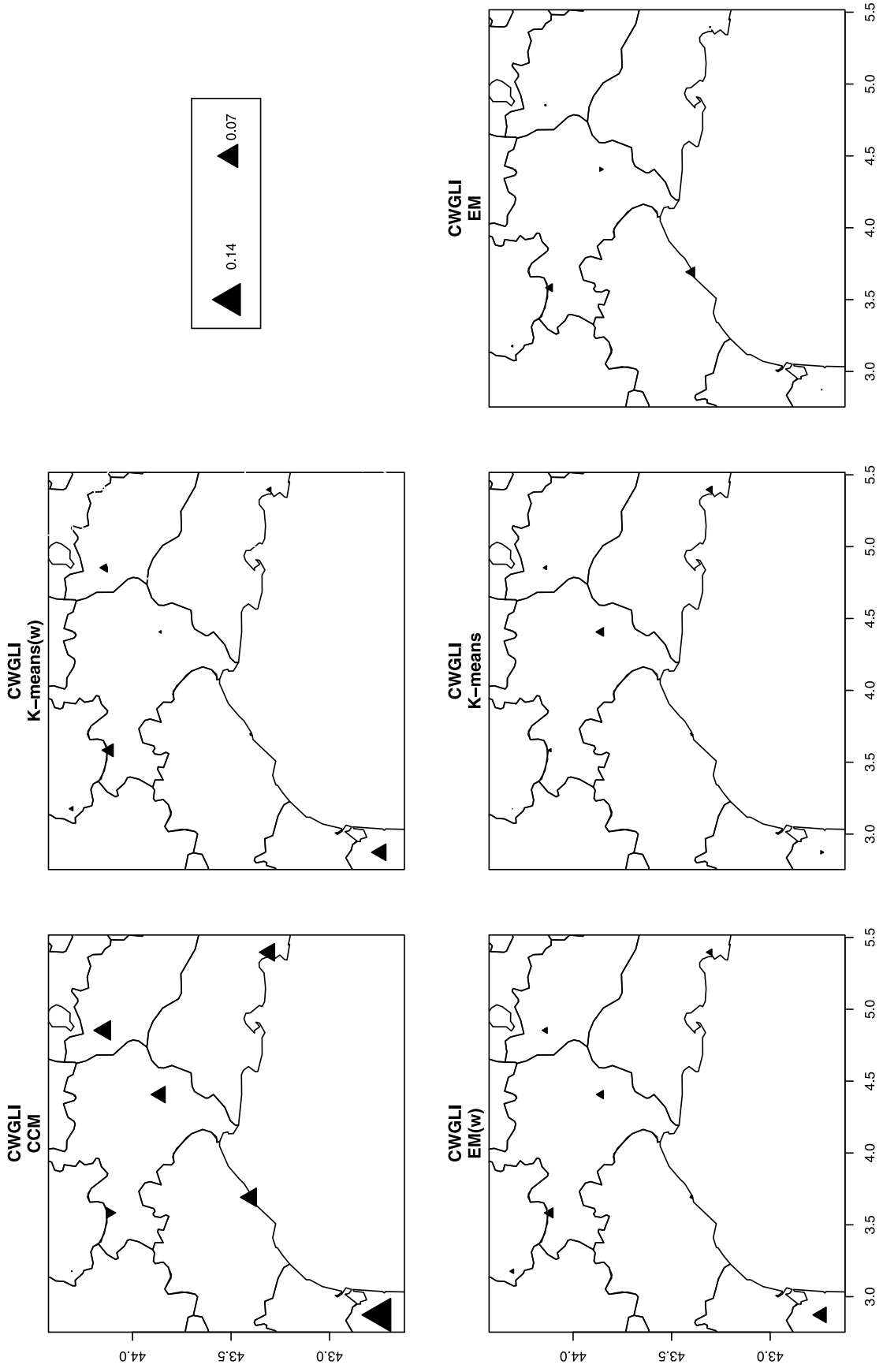

Figure 13. Conditional weighted global log intensity (CWGLI) errors due to misclassification, as defined in equation (15).

Table 7. Summary of the Regional Mean Costs for the Different Clustering Methods

Clustering Methods	CCM	*k*-Means	EM	*k*-Means(*w*)	EM(*w*)
RMDI	0.02	−0.01	−0.01	0.04	0.04
RWGP	−0.09	−0.02	−0.02	−0.02	−0.02
RCWGLI	0.06	0.01	0.01	0.03	0.03

[61] To summarize the information brought by the different cost functions, regional means are computed for WGP and CWGLI (denoted by RWGP and RCWGLI, respectively) in the same way as RMDI was defined in (14). Table 7 summarizes the values of the three regional costs for each clustering method.

[62] When such methods are used for classifying large-scale atmospheric circulation and local precipitation, we recommend that the following points are considered, to weigh the pros and cons of each method.

5.1. CCM

[63] Although the patterns obtained from CCM are well discriminated in terms of local precipitation, the percentage of bad classification is generally relatively high whatever the attribution method used. Moreover, those misclassifications can have important error costs when simulating precipitation (occurrence or intensity). Hence, CCM is not recommended to be used in a projection and modeling context. However, the local-scale information contained in the large-scale patterns makes CCM a very powerful tool in an analysis context to better understand the large-scale drivers of local-scale phenomena and particularly extremes.

5.2. Classical *k*-Means and EM (i.e., Applied to PCs)

[64] The patterns provided by those two approaches basically contain the same information. Although they provide consistent patterns in terms of the large-scale variables used for clustering, they usually do not discriminate very well the local-scale variable of interest (here precipitation) from one pattern to another. Indeed, contrarily to the CVs-based methods (CCM, *k*-means(*w*), and EM(*w*)), no local information is used in the definition of the patterns. Consequently, retrieving the pattern associated to the large-scale characteristics of a new given day is easier and, thus, their percentage of bad classification is lower. Moreover, the relative similarity from one pattern to another makes that a misclassification does not cost too much when modeling or simulating daily precipitation, with respect to the mean of the different clusters. In other words, those patterns used in a modeling context may have some ability in mean intensity but the precipitation variability could strongly be underestimated.

5.3. CVs-Based Methods (*k*-Means(*w*) and EM(*w*))

[65] Those two methods showed well-discriminated local precipitation characteristics from one large-scale pattern to another, particularly for extreme precipitation with EM(*w*) which allows to capture 100% of the precipitation events higher than the 99th with only two patterns (over seven) for two stations. Nevertheless, their PBC is better than the CCM PBC. Although their RMDI is high, it is mainly driven by one given station (Mont-Aigoual), famous to be difficult to model due to frequent intense events compared to other stations. However, the misclassification of new days implies a relatively small (i.e., good) cost in terms of rain occurrence probability (WGP), comparable to those from classical *k*-means and EM. Although less pronounced, the same conclusion holds for the cost in intensity.

6. Conclusions and Perspectives

[66] A comparison study between different clustering methods has been performed. The methods were designed to provide weather regimes based on large-scale NCEP-NCAR reanalyses over the Mediterranean basin for the winter season (NDJFM). Although, the results can change according to the months and season considered, the main point of this comparison was to see how much local-scale precipitation information was brought by the regimes from the different clustering methods. Five clustering algorithms were used, corresponding to three approaches: (1) the correlation clustering mixture (CCM, iteratively defining a CCA model per cluster); and the "classical" (2) *k*-means and (3) EM clustering methods applied either to the first principal components (about 95% of the variance) of the large-scale atmospheric variables (sea level pressure (SLP), and geopotential height a 500 mbar (Z500)), or to the *w* canonical variates obtained from a CCA performed on atmospheric data versus observed precipitation at seven rain gauges.

[67] In general, the *k*-means and EM algorithms applied to *w* CVs (*k*-means(*w*) and EM(*w*)) allow a good compromise between the two other approaches. Their percentage of bad classification is halfway between those from CCM and PCs-based methods, and their precipitation cost due to bad classification is generally lower than CCM's. Moreover, EM(*w*) provides regimes which very well discriminate local precipitation, particularly for the extremes.

[68] Those approaches, the evaluation of the pattern attribution performances, as well as the different precipitation cost functions developed here, have been implemented into an R package called "CCMtools" that should soon be freely available on the CRAN Web site (http://cran.r-project.org/) or upon request to the authors.

[69] This study could be extended in different ways. One of the most interesting studies to perform to pursue this analysis would be to compare one or several statistical downscaling methods (SDMs) based on the different WRs discussed in this article. Indeed, if SDMs are often conditioned by weather regimes, their influences are seldom investigated. WRs with high local-scale discrimination could bring valuable information within a downscaling context, but could also be implicitly attached to misclassification or difficult attribution of a new day to the WRs.

[70] From a methodological point of view, CCM is a *k*-means like algorithm that is sensitive to the initialization step. This could be improved by placing CCM in a simulated annealing context to have CCM converging toward a global optimum [*Philipp et al.*, 2007].

[71] Moreover, precipitation was the only focus of this analysis and other meteorological variables, such as temperature or wind, would certainly bring other regimes with different physical interpretation. Note that there is no a priori technical issue in using the discussed clustering methods for more than one variable (e.g., precipitation and wind at the same time). Only the interpretation of the re-

gimes can be more complicated but also depends on the analysis of interest.

[72] Although the large-scale atmospheric variables employed in this study (SLP and Z500) are quite usual to define weather regimes, other atmospheric variables could be better correlated to the local meteorological variable of interest (e.g., relative humidity, wind speed, temperature).

[73] More generally, the comparison framework we developed in this paper for southern France could be adapted to different regions of the world where local-scale effects may be predominant and not well captured by large-scale patterns.

[74] As final remark, we emphasize that there is no "miracle" nor "best" clustering method. According to the goal to reach and to the optimum weather regimes looked for (e.g., in terms of local discrimination, costs of misclassification), different solutions are possible. This study provides efficient tools to help the users to find this (or those) optimum WR(s), adapted to particular cases and variables.

[75] **Acknowledgments.** This work was supported by GIS-REGYNA, ANR MedUP, and CHAMPION projects. An R package called "CCMtools" has been developed for CCM clustering and tests on attribution performances. This package should soon be freely available on the CRAN Web site (http://cran.r-project.org/) or upon request to M. Vrac. The authors also would like to thank the two anonymous reviewers for their careful reading and constructive remarks throughout the review process.

References

Bárdossy, A., J. Stehlik, and H. J. Caspary (2002), Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules, *Clim. Res.*, *23*, 11–22.

Barnett, T., and R. Preisendorfer (1987), Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis, *Mon. Weather Rev.*, *115*, 1825–1850.

Breiman, L., J. Freidman, R. Olshen, and C. Stone (1984), *Classification and Regression Trees*, Wadsworth, Stamford, Conn.

Bunkers, M., J. Miller, and A. DeGaetano (1996), Definition of climate regions in the Northern Plains using an objective cluster modification technique, *J. Clim.*, *9*, 130–146.

Busuioc, A., R. Tomozeiu, and C. Cacciamani (2008), Statistical downscaling model based on canonical correlation analysis for winter extreme precipitation events in the Emilia-Romagna region, *Int. J. Climatol.*, *28*(4), 449–464.

Casola, J., and J. Wallace (2007), Identifying weather regimes in the wintertime 500-hpa geopotential height field for the Pacific-North American sector using a limited-contour clustering technique, *J. Appl. Meteorol. Climatol.*, *46*, 1619–1630.

Cassou, C. (2008), Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation, *Nature*, *455*, 523–527.

Davis, R., R. Dolan, and G. Demme (1993), Synoptic climatology of Atlantic coast northeasters, *Int. J. Climatol.*, *13*, 171–189.

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

Diday, E., Y. Ok, and A. Schroeder (1974), The dynamic clusters method in pattern recognition, in *Proceedings of the IFIP Congress 74*, edited by J. L. Rosenfeld, pp. 691–697, Elsevier, New York.

Duda, R., P. Hart, and D. Stork (2001), *Pattern Classification*, 2nd ed., John Wiley, Hoboken, N. J.

Fern, X., C. E. Brodley, and M. A. Fried (2005), Correlation clustering for learning mixtures of canonical correlation models, in *Proceedings of the Fifth SIAM International Conference on Data Mining*, edited by H. Kargupta, 439–448, SIAM, Newport Beach, Calif.

Fraley, C., and A. Raftery (2002), Model-based clustering, discriminant analysis and density estimation, *J. Am. Stat. Assoc.*, *97*, 611–631.

Gaffney, S., A. Robertson, P. Smyth, S. Camargo, and M. Ghil (2007), Probabilistic clustering of extratropical cyclones using regression mixture models, *Clim. Dyn.*, *29*, 423–440.

Hess, P., and H. Brezowski (1977), Katalog der grosswetterlagen Europas, *Ber. Dtsch. Wetterdienstes*, *15*, 1–14.

Hewitson, B., and R. Crane (2002), Self organizing maps: Applications to synoptic climatology, *Clim. Res.*, *26*, 1315–1337.

Hotelling, H. (1936), Relations between two sets of variates, *Biometrika*, *28*, 321–377, doi:10.1093/biomet/28.3-4.321.

Huth, R. (1996), An intercomparison of computer-assissted circulation classification methods, *Int. J. Climatol.*, *16*, 893–922.

Huth, R. (2001), Disaggregating climatic trends by classification of circulation patterns, *Int. J. Climatol.*, *21*, 135–153.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 370–471.

Klein Tank, A. M. G., et al. (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, *Int. J. Climatol.*, *22*, 1441–1453.

Lamb, H. (1972), British isles weather types and a register of daily sequence of circulation patterns, 1861–1971, *Geophys. Mem.*, *116*, 85 pp.

Leloup, J., M. Lengaigne, and J. P. Boulanger (2008), Twentieth century ENSO characteristics in the IPCC database, *Clim. Dyn.*, *30*, 277–291.

McLachlan, G., and D. Peel (2000), *Finite Mixture Model*, John Wiley, Hoboken, N. J.

Michelangeli, P., R. Vautard, and B. Legras (1995), Weather regimes: Recurrence and quasi-stationarity, *J. Atmos. Sci.*, *52*, 1237–1256.

Moron, V., A. Robertson, M. Ward, and O. Ndiaye (2008), Weather types and rainfall over Senegal. Part I: Observational analysis, *J. Clim.*, *21*, 266–287.

Pearson, K. (1894), Contributions to the theory of mathematical evolution, *Philos. Trans. R. Soc., Ser. A*, *185*, 71–110.

Philipp, A., P. Della-Marta, J. Jacobeit, D. Fereday, P. Jones, A. Moberg, and H. Wanner (2007), Long-term variability of daily north Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering, *J. Clim.*, *20*, 4065–4095.

Plaut, G., and E. Simonnet (2001), Large-scale circulation classification, weather regimes, and local climate over France, the Alps and western Europe, *Clim. Res.*, *17*, 303–324.

Plaut, G., E. Schuepbach, and M. Doctor (2001), Heavy precipitation events over a few alpine sub-regions and the links with large-scale circulation, 1971–1995, *Clim. Res.*, *17*, 285–302.

Pongracz, R., J. Bartholy, and I. Bogardi (2001), Fuzzy rule-based prediction of monthly precipitation, *Phys. Chem. Earth*, *9*, 663–667.

Sanchez-Gomez, E., and L. Terray (2005), Large-scale atmospheric dynamics and local intense precipitation episodes, *Geophys. Res. Lett.*, *32*, L24711, doi:10.1029/2005GL023990.

Sanchez-Gomez, E., L. Terray, and B. Joly (2008), Intra-seasonal atmospheric variability and extreme precipitation events in the European-Mediterranean region, *Geophys. Res. Lett.*, *35*, L15708, doi:10.1029/2008GL034515.

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

Sheridan, S., and L. Kalkstein (1998), Heat watch-warning systems in urban areas, *World Resour. Rev.*, *10*(3), 375–383.

Smyth, P., K. Ide, and M. Ghil (1999), Multiple regimes in Northern Hemisphere height fields via mixture model clustering, *J. Atmos. Sci.*, *56*, 3704–3723.

Toth, Z. (1991), Estimation of atmospheric predictability by circulation analogs, *Mon. Weather Rev.*, *119*, 65–72.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012.

Vannitsem, S. (2001), Toward a phase-space cartography of the short- and medium-range predictability of weather regimes, *Tellus, Ser. A*, *53*, 56–73.

Vrac, M., A. Chédin, and E. Diday (2005), Clustering a global field of atmospheric profiles by mixture decomposition of copulas, *J. Atmos. Oceanic Technol.*, *22*, 1445–1459.

Vrac, M., K. Hayhoe, and M. Stein (2007a), Identification and inter-model comparison of seasonal circulation patterns over North America, *Int. J. Climatol.*, *27*, 603–620, doi:10.1002/joc.1422.

Vrac, M., M. Stein, and K. Hayhoe (2007b), Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing, *Clim. Res.*, *34*, 169–184, doi:10.3354/cr00696.

Ward, J. (1963), Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, *58*, 236–244.

Wilks, D. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed., Elsevier, Oxford, U. K.

Yiou, P., and M. Nogaj (2004), Extreme climatic events and weather regimes over the North Atlantic: When and where?, *Geophys. Res. Lett.*, *31*, L07202, doi:10.1029/2003GL019119.

M. Vrac and P. Yiou, Laboratoire des Sciences du Climat et de l'Environnement, Centre d'Étude de Saclay, Orme des Merisiers, F-91191 Gif-sur-Yvette, France. (mathieu.vrac@lsce.ipsl.fr)