Development and comparison of circulation type classifications using the COST 733 dataset and software

Andreas Philipp,[a*] Christoph Beck,[a] Radan Huth[b] and Jucundus Jacobeit[a]

[a] *Institute of Geography, University of Augsburg, Germany*

[b] *Department of Climatology, Institute of Atmospheric Physics, Praha 4, Czech Republic*

ABSTRACT: In order to examine correspondence between different methods for circulation type classification, a dataset of classification catalogs for 12 different European regions has been created using a specially developed software package. Twenty-seven basic automatic classification methods have been applied in several variants to different input datasets describing atmospheric circulation. Together with six manual classifications a total of 33 methods are available for intercomparison.

Pattern correlation, frequency time-series correlation and the adjusted Rand index have been used for comparison. Highly significant correspondence has been detected only for two clustering techniques while the remaining classification methods show surprisingly low similarity. A Monte-Carlo test with 1000 classifications of randomly defined types even shows that most of the methods are not more similar among each other than any arbitrarily chosen types.

The predominant dissimilarity between the methods is interpreted to be a result of a lack of inherent structures of the input data. Only simulated annealing clustering and self-organizing maps get nearly identical results because they can optimally fit the partitioning to the outer shape of the data cloud in the phase space. Also methods based on pre-defined types come to very different results because small changes in the definition of thresholds may lead to large differences in the partitioning.

It is concluded that because of the missing inner structure of the data there is no clear statistical reason to prefer any of the examined methods. For practice in synoptic climatology this means that finding a suited classification for a certain purpose may require a broad comparison of methods. The software package cost733class for development, comparison and evaluation of classifications which was developed and used in this study is available at http://cost733.geo.uni-augsburg.de to facilitate this task.

KEY WORDS circulation type classification; weather types; Rand index; pattern correlation; manual classification; threshold-based classification; principal component analysis; cluster analysis

1. Introduction

Weather and circulation type classifications are a well-established tool for many applications in synoptic climatology, ranging from support for weather forecasting to climate model validation or downscaling (Huth *et al.*, 2008). The most prominent ones are the Lamb classification for the British Isles and the Hess and Brezowsky classification for Central Europe. However, because of the availability of high computing capacities the number of eligible methods for classification has grown considerably and is still increasing (e.g. Jolliffe and Philipp, 2010), while the differences between the classification results of the various methods remain somewhat nebulous. Thus, a question of major concern in synoptic climatology is whether it makes a difference to choose a certain classification method in favour of another or whether the differences are marginal. In practice, different classifications are commonly compared by visual inspection of the similarity between the spatial mean circulation type patterns. However, several authors have addressed the question on the diversity of classification results produced by different methods using statistical methods. For example, Huth (1996) compared five methods detecting distinct differences between them, while Jones *et al.* (1993) compared the Lamb and an automated approach and found large accordance; and Stehlík and Bárdossy (2003) compared manual with automated classifications and found them not to be independent. Seemingly, the very basic question of whether different classification methods lead to considerably different classification results when applied to atmospheric circulation data, and how large this difference might be, has been not definitely answered yet and remains for further clarification. A main reason for this might be that no considerably large set of classification methods has been compared systematically until now, for different regions and for different method configuration parameters such as number of types or input variables. However, this task

* Correspondence to: A. Philipp, Institute of Geography, University of Augsburg, Augsburg, Germany. E-mail: a.philipp@geo.uni-augsburg.de

has been addressed within the recently finished COST Action 733 ‘Harmonisation and Applications of Weather Type Classifications for European Regions’.

Within the framework of this COST Action, the attempt was made to collect and create a large number of different circulation type classifications in order to compare them and to achieve indications of the differences and usability of the classification results, which are recorded by the classification *catalogs*, i.e. the nominally scaled time series of type numbers or names representing the compressed information on atmospheric states.

Thus, a dataset of classification catalogs, called *cost733cat*, has been compiled by an open source software package called *cost733class* that has been developed especially for easily creating, comparing and evaluating classifications in several variants. The software can be controlled by command line arguments concerning input data, classification methods and configuration variants, which made it possible to establish a rather large sample of classification catalogs for comparison.

In contrast to an early version of the cost733cat dataset (*cost733cat-1.2*) which has been described in Philipp *et al.* (2010), the successor (*cost733cat-2.0*) allows for a more systematic comparison of methods by including additional classification configuration variants. Thus, seasonal- *versus* full-year classifications can be compared as well as single-day- *versus* 4-day-sequence classifications and the usage of additional circulation variables. The present state of the collection (*cost733cat-2.1*) finally includes additional methods: the Lamb (1972) weather types, hierarchical cluster analysis and Gaussian mixture models, leading to a total of 33 different methods or algorithms, and may be considered as rather complete. This comprehensive data base now allows a systematic examination of classification methods and puts conclusions about diversity of the catalogs on a more solid basis.

Concerning the type of classification methods, several categories may be discerned. Above all, they can be categorized into *weather type classifications* which account for several climate (or weather) variables at once for creating classes of atmospheric states and pure *circulation type classifications* including only circulation variables such as air pressure or derivatives thereof or large-scale wind components.

Furthermore, a distinction can be made between the *environment-to-circulation* approach and the *circulation-to-environment* approach in synoptic climatology as defined by Yarnal (1993). For the former, the range of values of a target variable, like precipitation or temperature, is divided into categories (often two, one below and one above average) for which the average circulation patterns are generated by compositing. In contrast, the *circulation-to-environment* approach only considers circulation variables for determining classes and the target variable is addressed afterwards. The latter is generally suited for a broad bandwidth of applications and not specialized for a certain target variable at a certain location. Finally there are hybrid methods where circulation and target variables are classified together.

Although the developed software is able to produce all of these kinds, in the present study only general circulation type classifications (according to the circulation-to-classification approach) are examined because the work in COST action 733 had a broad bandwidth of application fields for the classifications in mind Huth *et al.* (2008).

Conceptionally, two major principles of circulation type classification methods may be discerned: methods working on pre-defined types and methods producing derived types. The former are based on the idea to distinguish between large-scale zonal, meridional or cyclonic/anticyclonic flow which is the main synoptic controlling factor for the mid-latitude climate. The first group defines the types subjectively and assigns manually the elements to these types, called manual classifications (MAN). The second group uses numeric thresholds for type definition and automates the step of assigning the elements, called threshold-based methods (THR). The other methods do not use pre-defined types but try to find structures within the input data and to derive corresponding types. The first group doing so are methods based on principal component analysis (PCA). A second group uses the so-called leader algorithm (LDR), which is a cost-effective technique for data mining concerning computer capacity (also known as correlation-based methods, Yarnal, 1993), while the third group uses iterative algorithms for finding optimized partitionings concerning low within-type variance (OPT). The fourth technique tries to explain the empirical distribution of the attributes of the elements by a mixture model of several distributions, each representing a different process or class (MIX). Finally and only for the sake of statistical comparisons, a group of methods based on random processes can be established (RAN).

Of course there is no reason, why methods of the groups with pre-defined types (MAN, THR) should come to the same results as methods for data mining (PCA, LDR, OPT, MIX). However, *within* the groups of methods for pre-defined types, it may be assumed that there is at least some correspondence as they are commonly based on similar concepts of synoptic climatology defining the types according to the main directions of large-scale flow. Accordingly, the data mining methods may come to similar derived types if they are able to find the dominant structures of the data distribution.

This paper introduces the dataset of classification catalogs cost733cat derived by application of the mentioned method groups and evaluates their similarity in order to prove these assumptions. The article is structured as follows: After a brief introduction of the available classification methods and the software in Section 2, the input dataset and the classification configuration variants are described in Section 3. Then Section 4 introduces three different metrics to describe correspondence between classifications, while Section 5 presents the comparison results leading to conclusions in Section 6.

Table 1. Classification methods in the cost733cat dataset: the listing includes commonly used names of the methods, abbreviations used within this paper, the dataset version when the referring method was included, the methodological group (see below for abbreviations) and the key references.

Number	Classification name	Abbreviation	Version	Group	Key reference
1	Hess-Brezowsky Grosswetterlagen	GWL	1.0	SUB	Hess and Brezowsky, 1952
2	Lamb	LMB	2.1	SUB	Lamb, 1972; Jones *et al.*, 1993
3	Peczely	PEC	1.0	SUB	Peczely, 1957
4	Perret	PER	1.0	SUB	Perret, 1987
5	Schueepp	SUE	1.0	SUB	Schueepp, 1979
6	ZAMG	ZMG	1.0	SUB	Lauscher, 1985
7	Grosswettertypes	GWT	1.0	THR	Beck, 2000; Beck *et al.*, 2007
8	Jenkinson–Collison classification	JCT	1.0	THR	Jenkinson and Collison, 1977
9	Litynski	LIT	1.0	THR	Litynski, 1969
10	Objektive Wetterlagenklassifikation	WLK	1.0	THR	Dittmann *et al.*, 1995
11	Kruizinga	KRZ	1.0	PCA	Kruizinga, 1979
12	Principal components extreme scores	PXE	1.0	PCA	Esteban *et al.*, 2005
13	T-mode PCA obliquely rotated	PCT	1.0	PCA	Huth, 1993
14	T-mode PCA orthogonally rotated	PTT	2.0	PCA	Philipp, 2009
15	Lund	LND	1.0	LDR	Lund, 1963
16	Kirchhofer	KIR	1.0	LDR	Kirchhofer, 1974; Blair, 1998
17	Erpicum	ERP	1.0	LDR	Erpicum *et al.*, 2008
18	Ward's method	HWD	2.1	HCA	Murtagh, 1985; Ward, 1963
19	Single Linkage	HSL	2.1	HCA	Murtagh, 1985
20	Complete Linkage	HCL	2.1	HCA	Murtagh, 1985
21	Average Linkage	HAL	2.1	HCA	Murtagh, 1985
22	Mc Quitty's Method	HMQ	2.1	HCA	Murtagh, 1985; McQuitty, 1966
23	Median method	HMD	2.1	HCA	Murtagh, 1985
24	Centroid method	HCN	2.1	HCA	Murtagh, 1985
25	k-means (random start partitions)	KMN	2.1	OPT	Murtagh, 1985
26	k-means (Ward's start partition)	CAP	2.0	OPT	Yarnal, 1993
27	k-means (PXE start partition)	PXK	1.0	OPT	Esteban *et al.*, 2005
28	k-means (differing start partitions)	CKM	1.0	OPT	Enke and Spekat, 1997
29	k-means (most differing start partitions)	DKM	2.0	OPT	Enke and Spekat, 1997
30	k-medoids	KMD	2.1	OPT	Kaufman and Rousseeuw, 1990
31	Self-organizing feature maps	SOM	1.0	OPT	Michaelides *et al.*, 2007
32	Simulated annealing (SANDRA)	SAN	1.0	OPT	Philipp *et al.*, 2007
33	Gaussian mixture model	MXG	2.1	MIX	Hartigan, 1975
34	Random classification	RDM	2.1	RAN	For description see text
35	Random medoid classification	RAM	2.0	RAN	For description see text

The methodological groups cover subjective methods (SUB, the only group not produced by the software), threshold-based methods (THR), methods based on principal component analysis (PCA), leader algorithms (LDR), hierarchical cluster analysis (HCA), optimization algorithms (OPT), mixture models (MIX) and methods based on random processes (RAN) which are available for comparison purposes.

2. Classification methods and software

In order to achieve a reference set of widely used classification catalogs six manual classifications have been compiled into the collection. The 27 remaining methods have been programmed as subroutines within the software package *cost733class*, in order to apply it to the same input datasets and use the same classification parameters (in particular the numbers of types) where possible. In this way the differences among classification results may be ascribed to the classification algorithms itself and not to different ways they are used. The classification method can be chosen by a command line argument for the software including the abbreviation in Table 1 which feeds any input data configuration into the respective routine for classification. At the time of writing, 27 automated classification methods are implemented including the most often used techniques (Huth *et al.*, 2008). The detailed description of all methods is beyond the scope of this article. For this purpose we direct the reader to the key references given in Table 1. However, in order to get an overview we briefly describe the method groups in summary as mentioned in column *group* of Table 1.

2.1. Manual classifications (MAN)

In order to compare the automated classifications with established classifications, six manual catalogs have been included into the dataset (Table 1). As manual classifications are not modifiable, e.g. concerning the data pre-processing, the region or the number of types (k hereafter) as it is the case for automated classifications, their comparability is limited to some degree. Additionally there are differences concerning so-called *unclassifiable* days which usually do not exist for automated methods or differences concerning the period covered by the original catalogs. Thus, the Hess and Brezowsky, 1952 classification, e.g. consists of 29 *Großwetterlagen* (which may be pooled to 10 *Großwettertypen*) and 1 class for *undetermined* cases. It covers the period 1881 to present

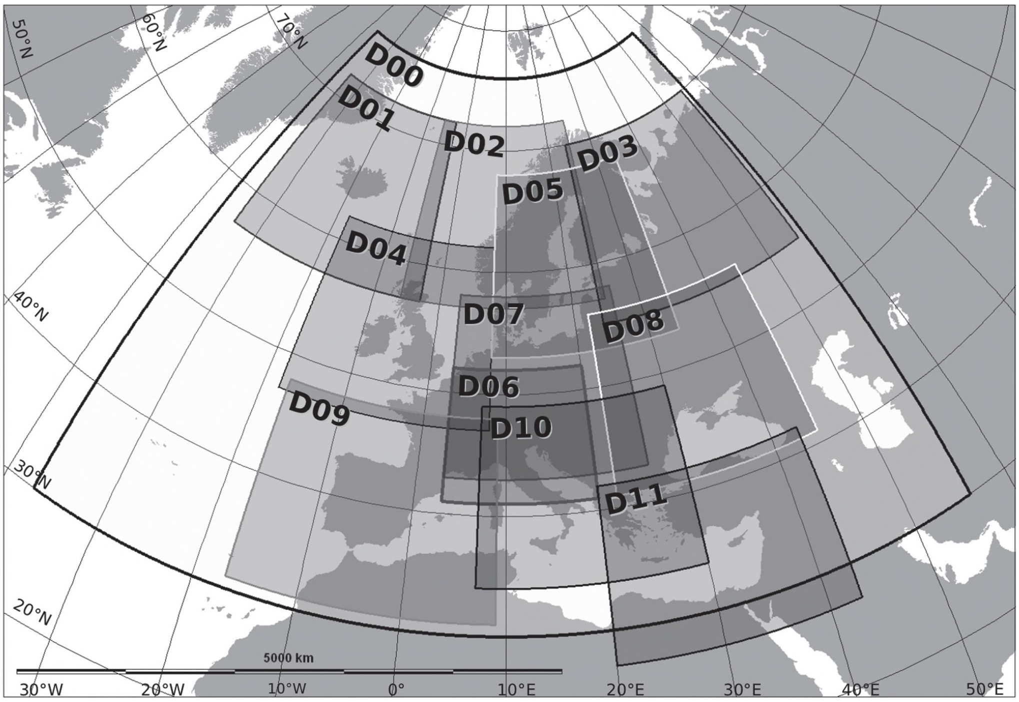

Figure 1. Spatial domains used for classification with automated classification methods in COST action 733 in Lambert equal area map projection.

and could therefore be included into the collection without modification. In contrast, the Lamb, 1972 catalog originally ends in 1997. This required a completion, which has been achieved by assigning all days not classified originally, by the minimum Euclidean distance to the sea level pressure (SP hereafter) patterns of the 26 types for the domain D04 (Great Britain, see Figure 1). This procedure, however, is only reasonable when omitting the class for *undetermined* cases. Further, for comparison, it is possible to pool the original 26 Lamb (1972) types to 8 directional types, 1 cyclonic and 1 anticyclonic type, leading to 10 types, or to 8 directional cyclonic and 8 directional anticyclonic types besides pure cyclonic and pure anticyclonic types, leading to 18 types. In contrast, a pooling to around 18 types seems less reasonable for the Hess and Brezowsky (1952) catalog, because it is less systematic. The other catalogs of Peczely (1957); Perret (1987); Schueepp (1979) and Lauscher (1985) have been adopted without modification. However, it is hard to relate the input data used for classification, to one of the datasets used in this study, concerning not only the circulation variable but also the region. Therefore, we treat them as a kind of *general* classification and compare them to all variants if only the number of types is approximately the same. Thus, the manual catalogs are generally less comparable than the automated methods. However, owing to their importance, they are included in the collection as reference.

2.2. Threshold-based classifications (THR)

The first group of automated classification methods is based on the concept of subjectively pre-defined types as it is done for the manual classifications. However, the assignment of cases to the classes is done objectively and automated by using threshold values for certain indices discriminating types from each other. Often three indices are used and they are often divided into three states (low, intermediate and high) leading to originally $3^3 = 27$ types. Variants with lower numbers of types (often 9 and 18 types are described) are derived by using fewer indices or fewer states. The indices themselves often represent large-scale flow directions (zonal and meridional) and vorticity (or high/low central pressure) and are derived from one single circulation map, where SP is most often used. The latter is the reason why these methods are not applicable for multi-parameter datasets, i.e. when circulation maps from two or more levels or variables should be classified together. However, a very special classification method is the WLK method of Dittmann *et al.* (1995), where the input variables are prescribed to be wind components and in the original form even a moisture parameter. Because it is not possible to apply this method, e.g. using SP alone, it is also not directly comparable to other classifications. However as with the manual classifications it is included in the dataset, keeping its special role in mind.

2.3. Classifications based on principal component analysis (PCA)

PCA (or empirical orthogonal functions) can be utilized for discrete classification either in T-mode or in S-mode; the former finding typical patterns (scores) and describing the degree of realization by loadings which can be divided in classes. In contrast, the latter finds typical modes of temporal variability (scores) in the data field which are realized at certain locations to a lesser or larger degree as described by the loadings. In this case the scores are used as classification index. Additionally they can differ by the type of rotation. Thus, the PCT method uses oblique rotation while the PTT uses the orthogonal VARIMAX rotation (both in T-mode) and KRZ (S-mode) uses no rotation at all while PXE (also S-mode) uses VARIMAX again.

2.4. Classifications based on the leader algorithm (LDR)

The leader algorithm is a predecessor of clustering algorithms, developed for low compute capacities Murtagh, 1985. Methods based on this algorithm are also known as 'correlation-based methods' (e.g. Yarnal *et al.*, 2001; Barry and Carleton 2001); however, the principle of finding groups can be applied for other distance/similarity metrics too, not only for correlation coefficients. Thus, following Murtagh (1985), the term 'leader algorithm' is preferred. The idea is to find representative key patterns for each type (i.e. the leader) by counting the number of elements with the similarity to the potential key pattern exceeding a certain threshold. The first key pattern is the one with the leading number of similar elements, the second key pattern is the one with the highest number after removing the elements similar to the first, and so on. After finding the key patterns, all elements are assigned to their most similar key pattern. No further optimization is done. Although modern computer capacities allow for exhaustive iterations for optimizing, the leader algorithm is still used today. Besides using the correlation coefficient as similarity metric (method of Lund, 1963), other ones are also used, e.g. the so-called *Kirchhofer* score, taking similarity in all parts of the map into account (Kirchhofer, 1974; Blair, 1998), or a pressure gradient metric (Erpicum *et al.*, 2008).

2.5. Hierarchical cluster analysis (HCA)

Another wide spread group of classification methods are the clustering algorithms or, more precisely, algorithms for cluster analysis. The first subgroup consists of hierarchical clustering algorithms, i.e. in a hierarchy of procedure steps, the two most similar clusters are combined, where at the beginning each element is located separately in its own cluster. At the step where only k clusters are left the procedure is stopped. As with the leader algorithm no further optimization is done. In this group, only agglomerative (and not divisive) algorithms are included from the routine of Murtagh (1985) which differ just by the kind of similarity metric used to find the pair of clusters to combine at each step.

2.6. Algorithms for optimizing partitions (OPT)

The second group of clustering algorithms works non-hierarchically, i.e. existing groups may be split up again during the procedure in order to further reduce within-type variance and reach an optimal partitioning. Because there is no other way to find the globally optimal partitioning of a sample of elements except of trying all possible combinations, which is computationally impossible for usual sample sizes, several approaches have been developed. Besides variants of the k-means algorithm (often differing only by the starting partition), also heuristics for optimization, like simulated annealing or self-organizing feature maps are included in this group.

2.7. Mixture models (MIX)

Very different from the methods before is the concept of mixture models. The idea behind this clustering technique is to describe the distribution of the elements in the multidimensional phase space (as defined by the number of attributes of the objects) as an overlay or mixture of k-multidimensional Gaussian distributions, where k is the number of types. The optimal assignment of the objects to the k types is estimated iteratively using an expectation–maximization (EM) algorithm. Thus, the optimization is not for minimal within-type variance (as it is the idea for the clustering algorithms mentioned above) but for optimal fit to the distributions.

2.8. Randomized classifications (RAN)

Two kinds of randomized classification algorithms are introduced to be able to compare the rational classification methods mentioned above with those which are products of random processes. The first method RDM defines the catalog number of all objects as a random number retrieved by a random number generator. The second method RAM (random medoid) introduces some structure into the classification scheme, as only the key element of each class is selected by random from the sample of objects to classify. Thus for each type one object is selected by random and all remaining objects are assigned to this key element by the minimum Euclidean distance. This method is used in Section 5.3 to generate samples for Monte-Carlo simulations.

3. Input data and classification variants

For all automated classification methods daily (12 UTC) ERA40 circulation data fields (Uppala *et al.*, 2005) have been used. The period covered by this dataset extends from September 1957 to August 2002. In order to assess regional variations when comparing the catalogs, subsets for 12 different spatial domains over Europe have been extracted, which differ by location and extent, as shown in Figure 1. While the large domain D00 includes all of Europe at a resolution of $2^\circ \times 3^\circ$, the smaller, regional domains D01 to D11 are resolved at $1^\circ \times 1^\circ$.

Furthermore, all methods are, in additional variants, applied to different atmospheric variables and their combinations, shown in Table 2, where possible.

While, as basic variants, all methods are carried out using solely SP (in Table 2) fields, additional variables are used either alone, in case of the 500-hPa geopotential height (Z5), or in combination with SP: i.e. SP together with Z5, together with thickness between 500 and 850 hPa (T5) and with vorticity of the 500 hPa field

Table 2. Variable combinations used as input data for automated classification methods.

Abbreviation	Description
SP	Sea level pressure
Z5	Geopotential height of the 500 hPa level
SP-Z5	SP and Z5 combined
SP-K5	SP and thickness between 500 and 850 hPa geopotential height
SP-Y5	SP and vorticity at the 500 hPa level
SP-Z5-Y5-K5	SP, Z5, Y5 and K5 combined

(Y5), resulting in three additional variants. Finally, one variant includes all five variables for classification (SP-Z5-Y5-T5) together. The combinations of the input variables together with the abbreviations used further on are listed in Table 2. In order to classify more than one variable field, the grid points of additional fields are treated just as additional attributes for each time step. However, as the physical units of the variables differ, this leads to a bias when calculating similarity metrics such as the Euclidean distance. Therefore, each field is normalized separately for each variable when combining two or more variables. A very special case, concerning the input data is the WLK classification, using originally wind components and geopotential height of two levels as well as humidity. In order to make it comparable to pure circulation type classification the humidity component can be avoided; however, the wind components remain as an incompatible speciality. Nevertheless, it is compared to classifications using the standard input variables mentioned above, keeping this exception in mind.

As the differences between methods may depend also on the number of types, all methods have been configured for 9, 18 and 27 types where possible. In some cases these numbers can be realized only approximately, leading to discrepancies of 2 at the maximum from the intended numbers mostly. However, in other cases like for the PXE or the PXK classification Esteban *et al.* (2005), it is not possible to produce classifications exceeding a certain number of types, as some classes stay empty.

All classifications are carried out on a daily basis for the ERA40 period 1st of September 1957 until 31st of August 2002. However besides the standard full-year classifications (YR), seasonal classifications (SE), i.e. using subsets for Winter (December, January, February), Spring (March, April, May), Summer (June, July, August) and Autumn (September, October, November) have been performed. In order to make these seasonal classifications comparable to the full-year classifications concerning the number of types, only seven types are created in each season, leading to a total of 28 which is close to the number of types realized for most of the classification methods in the standard variant (27).

Another time-related configuration variant relates to using not only single-day fields but also sequences of fields, i.e. each element is not only described by the atmospheric state of the actual day, but also by the fields of the three preceding days. Such a procedure has been found to be useful for applications dealing with persistent variables, as for example air temperature Philipp (2009). The construction of field sequences is done automatically by the software for any given sequence length, where a sequence length of 4 days (S04) has been chosen as an alternative to single-day classifications (S01).

A further way of integrating variance in time into one time step is to use a low-pass filter with Gaussian weights for a window of length 11 (F11). Finally a commonly used way of compressing and reweighting information of the input data is PCA in S-mode. For determination of the number of PCs to retain, the fraction of 90% of explained variance is chosen (P90), leading, e.g. to three PCs for the full-year SP data of domain D07.

Calculating the total number of possible classification variant combinations leads to 27 automated methods × 12 domains × 3 numbers of types × 6 input data variable combinations × 2 seasonal variants × 2 sequential variants × 2 filter variants × 2 PCA variants = 93 312 possibilities. In order to avoid this multitude of variants which have to be compared, we chose one standard variant as reference and changed only one parameter at a time for the estimation of its effect. This standard or reference variant is defined to use SP of domain D07 (Central Europe, see Figure 1) and the number of types $k = 27$. Further it is applied on full-year, single-day, raw data which are not filtered and not compressed by PCA.

4. Comparison metrics

4.1. Comparison of circulation patterns

The first and sometimes the most important feature for characterizing a circulation type classification is the spatial pressure pattern of the types. Commonly it is created by calculating the mean pattern (or composite) of all elements of the referring type, often called *centroid*, a term derived from cluster analysis, denoting the centre of the cloud of elements in the multidimensional phase space. Those mean patterns are frequently used to compare classifications, as done, e.g. by Crane and Barry (1988) for comparing the Kirchhofer types of an observational dataset with those of model output data, by Michelangeli *et al.* (1995) to determine similarity for the evaluation of robustness in cluster analyses or by Beck (2012) to compare the spatial structures of the circulation types with comparable weekly frequency distributions. Further, many authors compare their centroid patterns with those already documented in literature in order to confirm their results (e.g. Plaut and Simonnet, 2001). Therefore, the correspondence of the classifications according to pattern similarity in cost733cat is examined.

In order to quantify the correspondence of the spatial patterns of two classifications, the minimum Pearson correlation coefficient of the most similar pairs of patterns is used. To determine this metric, in a first step the most similar pairs of types are defined: among all possible pattern correlation coefficients, the highest defines the first and most similar pair of patterns. Among the remaining patterns, again the highest correlation coefficient defines the second pair, and so on, until all patterns of the first classification are assigned to their counterparts in the second classification. If there are more types in the first than in the second classification, the remaining types of the first classification are additionally assigned to the most similar patterns of the second classification. The reason to define the minimum of these selected correlation coefficients as a measure for similarity between the whole of the two classifications is that an overall similarity can only be declared if *all* pairs match significantly.

In order to check significance of the correlation coefficients between the centroid maps, spatial autocorrelation must be taken into account. As there is no independence of the circulation data in space, the number of cases (grid points) of the two samples (maps) is not suitable to determine the degrees of freedom of the correlation coefficient between the two. Instead, the adequate degrees of freedom must be determined depending on the auto-covariances, associated with the number and structure of the highs and lows in the maps. For a more detailed discussion see Legendre (1993), who provides a FORTRAN routine accounting for spatial autocorrelation in significance tests following the method of Dutilleul (1993). For the present study, significance of the similarity between two classifications is assumed if all pairs mentioned above are significantly correlated at least on the 0.1 alpha level.

An example for achieving this metric is given in Table 3. Although there are several very high and significant correlation coefficients between pairs of patterns, this cannot be observed for all patterns. Thus for pattern pair 25, the p-value of 0.1588 indicates a high probability to make an error when assuming that the $r = 0.6891$ is different from zero, owing to high autocorrelation that reduces the degrees of freedom from $df = 382$ grid points to $df = 3.5413$.

Table 3. Example of defining pattern similarity between Hess and Brezowsky and Lamb weather types using SP in domain 07.

Number	Pair	*r*	*df*	*p*-value
1	02–25	0.9878	5.6558	0.0000
2	16–01	0.9812	5.1054	0.0001
3	06–23	0.9798	4.9789	0.0001
4	13–08	0.9772	4.6208	0.0003
5	03–26	0.9753	4.8858	0.0002
6	28–22	0.9717	2.6192	0.0126
7	07–07	0.9589	5.3116	0.0005
8	05–14	0.9507	6.4986	0.0002
9	01–15	0.9472	7.0132	0.0001
10	24–12	0.9451	3.3292	0.0109
11	29–18	0.9428	3.4541	0.0102
12	22–03	0.9371	5.2899	0.0014
13	26–21	0.9367	1.9184	0.0741
14	20–11	0.9297	4.8800	0.0029
15	08–16	0.9278	5.5544	0.0015
16	14–05	0.9126	6.4385	0.0011
17	25–20	0.9120	3.1126	0.0281
18	27–19	0.9104	2.3270	0.0671
19	18–04	0.8912	5.2245	0.0059
20	10–06	0.8722	14.7166	0.0000
21	15–02	0.8643	6.0767	0.0053
22	12–09	0.7930	3.4827	0.0831
23	17–17	0.7706	8.4273	0.0074
24	30–24	0.7631	9.0364	0.0062
25	04–13	0.6891	3.5413	**0.1588**
26	23–10	0.6127	9.7657	0.0365
27	09–06	0.7535	11.7500	0.0021
28	11–10	**0.5773**	10.3074	0.0459
29	19–03	0.9266	5.0609	0.0025
30	21–11	0.9224	5.1332	0.0028

Twenty-six Lamb weather types are assigned to the 26 most similar Hess and Brezowsky types according to the maximum correlation coefficient r. The remaining Hess and Brezowsky types are additionally assigned to the most similar Lamb types. Among these is the lowest correlation coefficient (pair 28) which defines the overall pattern similarity metric for these two classifications with $r = 0.5773$ (bold number in column "r") which is significant on the 0.05 alpha level (p-value 0.0459). However, because of high spatial autocorrelation, the p-value for pair 25 (bold number in column "p-value") leads to the final result that similarity cannot be assumed for all pairs on the 0.1 alpha level.

4.2. Comparison of frequency time series

Another important characteristic of classifications is the variability over time, which may be described by the annual frequency of each type within the study period. Accordingly, correspondence between classifications can also be described by correlation coefficients between the type-specific frequency time series. An overall metric for similarity is therefore achieved in the same manner as done for the spatial patterns above, but replacing pattern correlation by frequency time-series correlation. In order to account for temporal autocorrelation, when estimating the significance of the correlation coefficients, the degree of freedom is reduced according to the effective sample size Werner (2002) as shown in the example in Table 4.

4.3. Comparison of partitionings

Besides using circulation patterns or frequency time series of the types for inter-comparison, there are metrics based on the classification catalogs directly. Steinley (2004) suggested the Rand (1971) index adjusted by Hubert and Arabie (1985) as the most appropriate metric for describing similarity between two classifications. The idea behind the Rand index is to count how many pairs of objects (days) are together in one class in both classifications (called quantity a) and how many are in two different classes in both classifications (quantity d). The sum of this degree of correspondence is then scaled to vary between 0 (no correspondence) and 1 (identity), by this sum plus the number of pairs of objects together in one class in the first classification but separated in the second (quantity b) and vice versa (quantity c):

$$\mathrm{RI} = \frac{a + d}{a + b + c + d} \qquad (1)$$

However, as any two classifications show some correspondence by chance, the Rand index has to be adjusted

Table 4. Example of defining time-series similarity between Hess and Brezowsky and Lamb weather types using SP in domain 07.

Number	Pair	r	df	Alpha level
1	30–01	0.5017	28.8102	0.0100
2	20–11	0.5011	45.0000	0.0100
3	01–07	0.4903	45.0000	0.0100
4	14–03	0.4798	41.5691	0.0100
5	10–14	0.4615	38.7367	0.0100
6	08–21	0.4059	44.8910	0.0100
7	29–04	0.4017	43.8358	0.0100
8	13–17	0.3998	45.0000	0.0100
9	03–18	0.3909	38.7462	0.0100
10	05–23	0.3768	36.6750	0.0100
11	06–20	0.3723	45.0000	0.0100
12	02–15	0.3581	40.0639	0.0500
13	24–13	0.3523	44.7714	0.0100
14	28–12	0.3513	45.0000	0.0100
15	04–26	0.3501	45.0000	0.0100
16	22–02	0.3424	45.0000	0.0100
17	15–10	0.3133	45.0000	0.0500
18	09–05	0.3024	43.7891	0.0500
19	27–16	0.2848	40.5995	0.1000
20	26–24	0.2307	45.0000	0.1000
21	12–22	0.1914	40.0628	1.0000
22	07–09	0.1443	45.0000	1.0000
23	11–06	0.1264	41.6326	1.0000
24	16–08	0.0945	39.9222	1.0000
25	17–19	0.0904	39.9569	1.0000
26	23–25	**0.0493**	37.2295	**1.0000**
27	18–01	0.2348	44.6181	1.0000
28	19–13	0.3091	43.7114	0.0500
29	21–18	0.2243	37.8046	1.0000
30	25–13	0.2552	43.8292	0.1000

Twenty-six Lamb weather types are assigned to the 26 most similar Hess and Brezowsky types according to the maximum correlation coefficient r. The remaining Hess and Brezowsky types are additionally assigned to the most similar Lamb types. The lowest correlation coefficient (pair 26) which defines the overall time-series similarity metric for these two classifications with $r = 0.0493$ (bold number in column "r") is not significant on the 0.05 alpha level (bold number in column "Alpha level", tested with confidence intervals). The number of degrees of freedom is partially reduced for several pairs from originally $df = 45$ $(n-1)$ to a minimum of $df = 28.81$ (pair 1) according to temporal autocorrelation. The last column shows the alpha level on which a correlation $r \neq 0$ can be assumed. Overall, a significant similarity between these two classifications cannot be assumed.

to be 0 if the correspondence is as high as expected by chance. Among several other variants the adjustment method of Hubert and Arabie (1985) has been accepted to be the correct one Steinley (2004):

$$\mathrm{ARI_{HA}} = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]} \tag{2}$$

On the basis of Monte-Carlo simulations with varying contingency tables for examination of the properties of the $\mathrm{ARI_{HA}}$, Steinley (2004) suggested that an $\mathrm{ARI_{HA}} > 0.90$ reflects *excellent* correspondence, an $\mathrm{ARI_{HA}} > 0.80$ *good* correspondence, an $\mathrm{ARI_{HA}} > 0.65$ *moderate* correspondence and an $\mathrm{ARI_{HA}} < 0.65$ *poor* correspondence. Further, a Monte-Carlo method for testing the significance of an $\mathrm{ARI_{HA}}$ value being higher than a certain level is suggested by Steinley (2004) which has been programmed within the *cost733class* software. Among 1000 perturbed contingency tables, reflecting the observed level of overlap of the two partitions to compare, the number of the corresponding $\mathrm{ARI_{HA}}$ values exceeding a certain value is counted (for details see Steinley 2004). This number reflects the likelihood that the certain $\mathrm{ARI_{HA}}$ value is reached by chance. Thus, it is possible to test the null hypothesis that the observed $\mathrm{ARI_{HA}}$ is equal to a certain reference value, in particular the values given above indicating the correspondence to be *excellent*, *good*, *moderate* or *poor*.

However these levels of correspondence are proposed for a very general rating. Although they give a rough idea about what to expect, e.g. that *poor* correspondence might indicate almost independence and *excellent* correspondence almost identity, a concrete reference is missing. For example, it would be helpful to test an observed $\mathrm{ARI_{HA}}$ exceeding $\mathrm{ARI_{HA}}$ values resulting from a random process. In particular, a considerable observed $\mathrm{ARI_{HA}}$ value should exceed a percentile threshold (depending on the level of significance) from reference $\mathrm{ARI_{HA}}$ values resulting from the comparison of randomly defined circulation types which can be created by the RAM method introduced in Section 2.8. Therefore, a Monte-Carlo simulation is implemented generating 1000 RAM classifications (see Section 2.8) for each classification variant. The distribution of the resulting reference $\mathrm{ARI_{HA}}$ values of all combinations of the 1000 classifications then allows to test whether any observed $\mathrm{ARI_{HA}}$ value is significantly higher than those based on the random process. If an observed $\mathrm{ARI_{HA}}$ value does not exceed the 95th percentile of this reference distribution we can assume with a likelihood of 95% that it is not different to the RAM-$\mathrm{ARI_{HA}}$ values (null hypothesis), whereas an exceedance indicates a significantly higher $\mathrm{ARI_{HA}}$ value (alternative hypothesis).

5. Correspondence among classification methods

In order to evaluate the similarity of classifications achieved by the 33 different methods and algorithms, the resulting catalogs have been compared among each other using the standard configuration described in Section 3. As this configuration is fixed for all automated methods concerning numbers of types, input data and pre-processing, the differences should emerge purely as a function of the classification algorithm. The only exceptions are the manual methods, where no clear definition of the space domain and the input data is available and those methods which cannot be applied to the number of types of 27 as well as WLK which includes special input data.

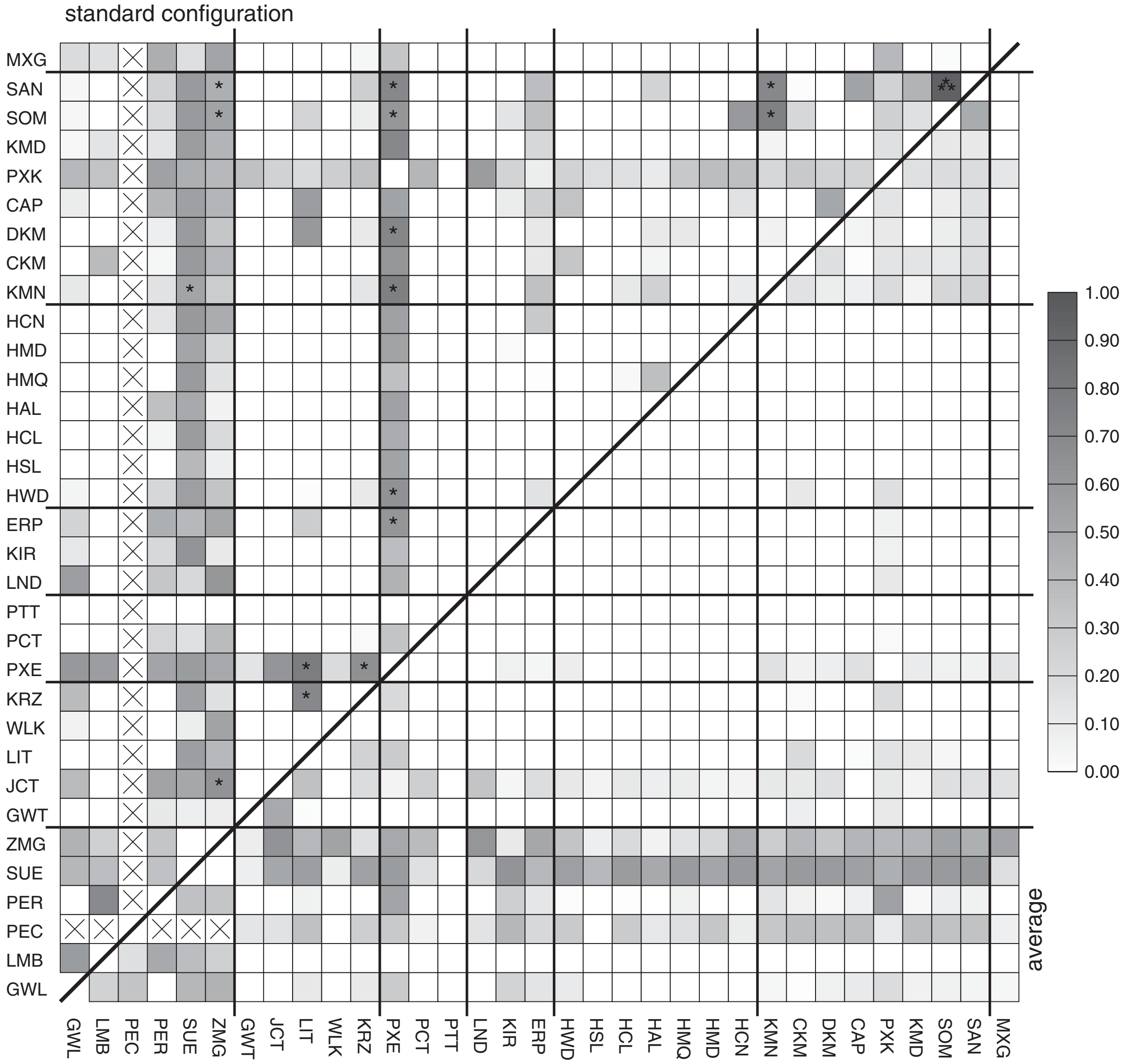

Figure 2. Comparison of classification methods by minimum centroid pattern correlation coefficients of most similar pairs of types. Upper left triangular part of matrix: standard configuration defined in Section 3, lower right triangular part of matrix: average over all configuration variants as listed in Table 5. Shading denotes correlation strength from $r = 0$ (white) to $r = 1.0$ dark grey, stars in the upper left half denote significance level (one for alpha level 0.1, two for 0.05, three for 0.01). Crosses mark combinations where one of the classification methods is not applicable.

Figures 2–4 show the comparison metrics based on pattern correlations, frequency time-series correlations and the adjusted Rand index, for the pairwise comparison of all 33 classification methods. In each figure the upper left triangular part shows the values for the standard configuration (Section 3). Furthermore, in order to exclude interpretation errors based on an unfortunate selection of classification configuration parameters, the comparison is extended to the other classification variants. For all these variants in all domains the average similarity metric for each pair of classification methods is calculated and displayed in the lower right triangular part of Figures 2–4.

The maximum number of variants used for averaging includes classifications for the 12 domains, each applied for the 12 variants listed in Table 5, leading to 144 cases. However, as, e.g. some manual classifications are invariant, the minimum number is 1, while the threshold methods using only one gridded input field can be applied in 60 variants. Thus these mean values differ concerning the underlying sample size. Nevertheless they are reflecting the central tendency of correspondence and are therefore shown additionally in Figures 2–4. Over all, the three metrics for describing similarity show somewhat different results.

5.1. Pattern similarity

For pattern similarity within the standard configuration (Figure 2, upper left triangular part) some high correlation coefficients are reached, several of them are significant. A salient feature can be observed for the manual methods PER, SUE and ZMG, for the PCA method PXE and for its related optimization method PXK (see Table 1 for abbreviations): they show relatively high pattern similarity with most of the other methods even though the latter only partly shows similarity among each other. The reason for this discrepancy can be determined to be

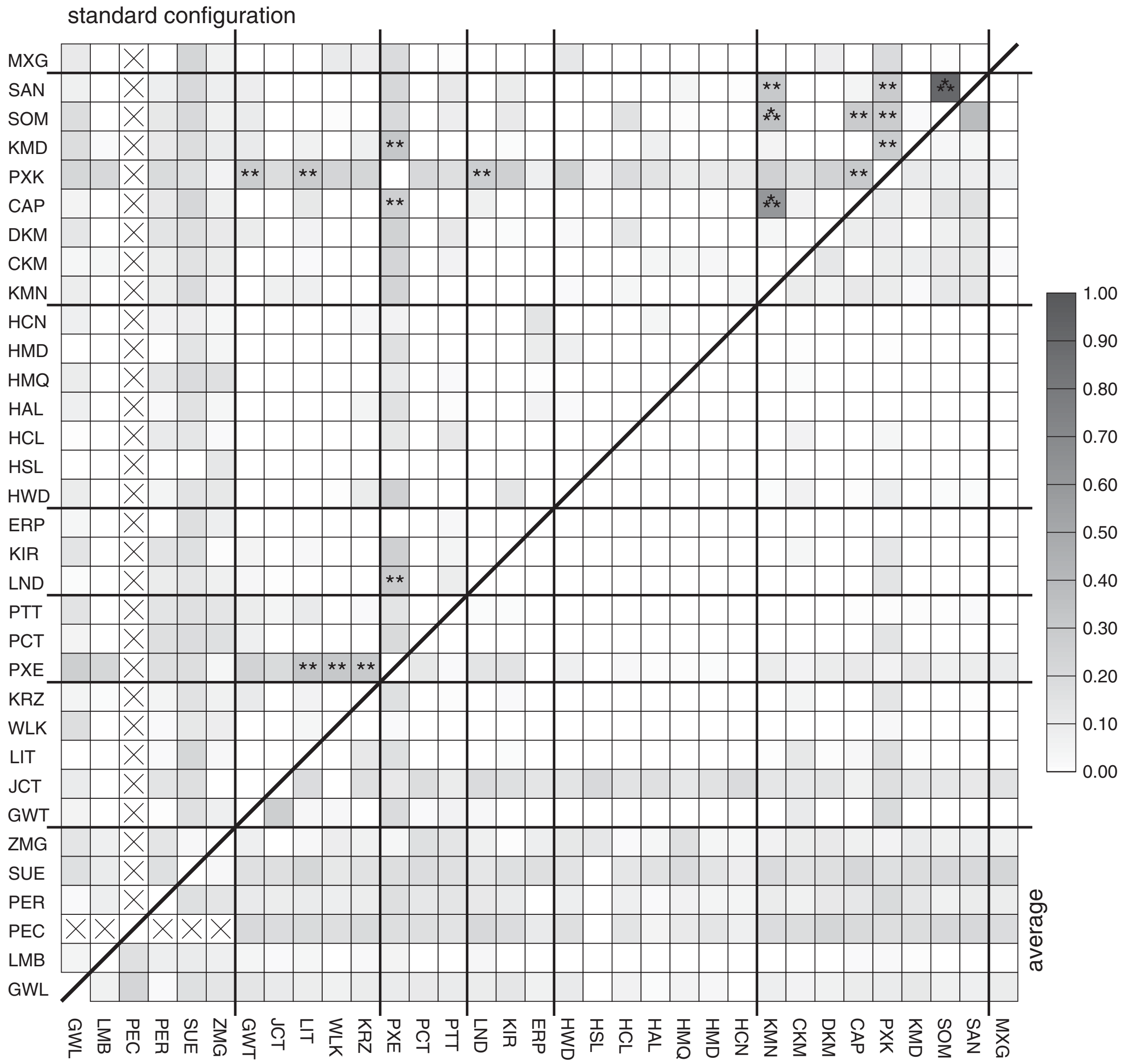

Figure 3. Comparison of classification methods by frequency time-series correlation coefficients. Upper left triangular part of matrix: standard configuration defined in Section 3. Lower right triangular part of matrix: average over all configuration variants as listed in Table 5. Shading and crosses as in Figure 2. Stars in the upper left half denote significance level (one for alpha level 0.1, two for 0.05 and three for 0.01).

the different numbers of types realized by the methods. Thus PXE effectively has only 16 types because types 17–27 are empty. Further, in the group of manual methods GWL has 30 types, LMB 26, PER 31, SUE 40 and ZMG even 43 types instead of the 27 types of the standard configuration. Since this metric selects the weakest correlation coefficient among the most similar pairs of patterns, the reason for these high metrics is apparently the larger flexibility to find a higher minimum correlation when the numbers of types are unequal. This artificial effect seems to be so strong, that this metric should actually not be referred to if unequal numbers of types exist. Moreover, this points out the risk of misinterpretation when comparing patterns for determining similarity of classifications since spatial patterns appear to show high similarity just by chance. In contrast, no such artefact is evident for the rest of the automated methods which all have the exact number of 27 types (except of the PXK method which has also only 16 types because it is directly derived from PXE). Apart from that some correspondence can be noticed within the group of manual methods (e.g. between GWL and LMB) although statistically not significant and significantly for the threshold-based methods LIT and KRZ. Also among some optimization methods similarity can be observed, significantly for SAN with SOM, two advanced clustering methods, and between both to KMN, the *k*-means clustering procedure initialized by random starting partitions. Other combinations show only low and insignificant similarities.

The more or less same pattern, although on a lower level due to smoothing effects, can be observed for the average pattern similarity over all configuration variants (Figure 2, lower right triangular part). Noteworthy correspondence remains for the manual classifications with high numbers of classes for the same artificial reason as explained above. Apart from that, JCT shows some resemblance to GWT and LIT (method group THR)

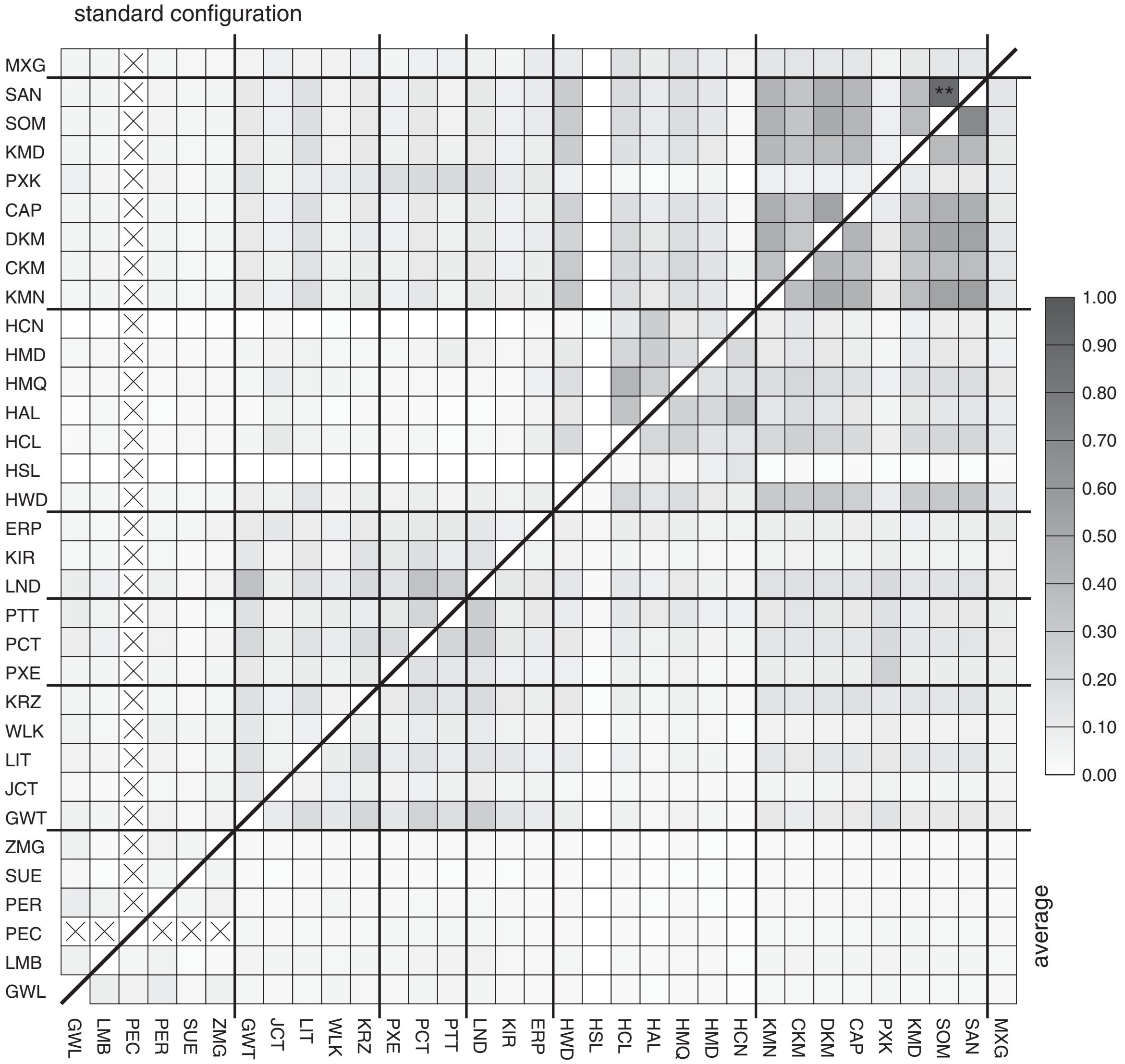

Figure 4. Comparison of partitionings of classifications methods by the adjusted Rand index. Upper left triangular part of matrix: using the standard configuration defined in Section 3. Lower right triangular part of matrix: average over all configuration variants as listed in Table 5. Shading and crosses as in Figure 2. Stars in the upper left half denote correspondence level: one for $\mathrm{ARI}_{\mathrm{HA}} > 0.65$ (moderate), two for $\mathrm{ARI}_{\mathrm{HA}} > 0.8$ (good) and three for $\mathrm{ARI}_{\mathrm{HA}} > 0.9$ (excellent), tested for significance on the 0.05 alpha level.

and within the group of optimization methods there are slightly increased values also for the other configurations, the highest being 0.49 for SOM and SAN.

5.2. Frequency time series

The matrix of the minimum correlation coefficients between the most similar frequency time series of the types for the standard configuration (upper left triangular part of Figure 3) shows generally low, however also some significant coefficients. This metric seems to be less distorted by different numbers of types than the pattern-correlation-metric, since the manual methods show less similarity to other methods any more. However, as is apparent for method PXE and, although weaker, for PXK, the effect still exists. Further, a noticeable frequent and partially significant correspondence exists among several of the optimization methods, except of the CKM, DKM and KMD methods. However, when looking at the average values (Figure 3, lower right triangular part), only the similarity between SOM and SAN is strong and systematic enough to persist (0.39).

5.3. Adjusted Rand index

Finally Figure 4 shows the adjusted Rand index. Looking at the upper left triangular part for the standard configuration, the most striking difference to the other metrics is that no artificial similarity is displayed any more for PXE and other methods, which demonstrates the robustness of this metric concerning differing numbers of types. Apart from that, the overall level of similarity is again distinctively low. Slightly increased similarity is indicated among the threshold (THR) based, PCA-based methods and the leader (LDR) algorithms. Also, among the hierarchical clustering methods (HCA) some little correspondence can be observed (except for the single linkage method HSL) as well as between

Table 5. Classification variants.

Number	Variant	Code
1	Reference: raw SP data and $K = 27$ types	K27_YR_S01_F00_P00_SP
2	As #1 but K = 18 types	**K18**_YR_S01_F00_P00_SP
3	As 1 but K = 9 types	**K09**_YR_S01_F00_P00_SP
4	As #1 but 500 hPa height instead of SP	K27_YR_S01_F00_P00_**Z5**
5	As #1 but SP and Z5	K27_YR_S01_F00_P00_**SP-Z5**
6	As #1 but SP and vorticity at 500 hPa	K27_YR_S01_F00_P00_**SP-Y5**
7	As #1 but SP and thickness 500–850 hPa	K27_YR_S01_F00_P00_**SP-K5**
8	As #1 plus 500 hPa vorticity and thickness	K27_YR_S01_F00_P00_**SP-Z5-Y5-K5**
9	As #1 but 4 day sequences	K27_YR_**S04**_F00_P00_SP
10	As #1 but with 11 pt. low-pass filter	K27_YR_S01_**F11**_P00_SP
11	As #1 but with PCA data compression	K27_YR_S01_F00_**P90**_SP
12	As #1 but separately for the four seasons	K27_**SE**_S01_F11_P00_SP

HCA and OPT methods, although on a nearly negligible level. Only the ARI_{HA} values within the optimization group are higher than 0.33, except of PXK. However, again, the only similarity exceeding better than *poor* correspondence with confidence, exists between SAN and SOM (*good* correspondence with $\text{ARI}_{\text{HA}} = 0.89$). Further, some, although *poor*, similarity can be stated between the HWD hierarchical clustering method and the non-hierarchical methods, explainable because the HWD is the only hierarchical method aimed to reduce within-type variance. No explanation can be found for the slightly increased, but also *poor* correspondence between LND and GWT ($\text{ARI}_{\text{HA}} = 0.35$) and LND and PCT ($\text{ARI}_{\text{HA}} = 0.35$ again).

The maximum mean adjusted Rand index (lower right triangular part of Figure 4) is 0.70, now indicating only *moderate* correspondence between SAN and SOM. The second and third highest values are 0.55 and 0.53 between KMN and SAN and SOM, denoting *poor* correspondence. All other average values are below 0.40.

In summary, it can be stated that unexpected low correspondence between the classification methods is observed in general. Even the optimization methods which are related very closely show only low similarity with the only exception of SAN and SOM. Of course it should not be expected that classifications with predefined types (method groups MAN and THR) working with main directions of advection, are similar to those methods where the types emerge during the classification process itself (PCA, LDR, OPT and MIX). However, the markedly low similarity *within* these two conceptional groups of methods is surprising. Hence, this finding gives rise to the question of whether the classification methods show significantly more correspondence in defining or detecting structures in the datasets than any randomly defined partitioning of the data. Therefore the Monte-Carlo test as described in Section 4 is applied to the standard configuration as well to all other variants.

Figure 5 illustrates the Monte-Carlo test for the standard configuration. It shows the kernel density estimate of the Rand index values among 1000 RAM catalogs (solid curve) compared with the Rand index values among the deliberate methods (histogram). Unexpectedly, the location of the distribution of the Rand index for the deliberate methods appears to be distinctly shifted towards lower values compared to the RAM classifications. Only a minority (30 of 528) of the observed ARI_{HA} values exceed the 95th percentile of the RAM-ARI_{HA} distribution (which is 0.28), indicating that they differ significantly. They are all from the method groups HCA and OPT with the only exception of GWT-LND. This means, apart from the similarity between some of the optimization methods, that the deliberate methods do not have more in common than the classifications based on randomly chosen types.

The phenomenon of RAM showing even more correspondence than other classifications in Figure 5 can be explained by the use of one and the same distance metric (Euclidean Distance) in all RAM classifications, while the observed ARI_{HA} values include those where methods using different distance metrics are compared. However, this does not indicate that this test is ill-conditioned due to the RAM-ARI distribution being biased towards too high values. If a high ARI_{HA} value would be caused by using the same distance metric alone, then all pairs of methods having the distance metric in common should be distinctively more similar than the RAM catalogs, which is not the case. Therefore, differences in the conception of the classification methods must be the reason for the distinct dissimilarities.

Again, this result is not only true for the standard classification configuration. As shown in Figure 6 the correspondence among the deliberate classification methods (light grey boxes in Figure 6) is in no case generally higher than the correspondence among the randomized classifications) dark grey boxes in Figure 6).

Even those data pre-processing procedures reducing the detail level (or noise) of the data (e.g. the PCA pre-processing), do not lead to systematically stronger resemblance between the catalogs. The only configuration variant resulting in considerably strong similarity among catalogs is the seasonal classification (line 27_SE_S01_SP in Figure 6). Probably the reduced sample size for classifications applied separately to the four seasons gives less scope for different partitionings. The same is true for the second noticeable variant, the

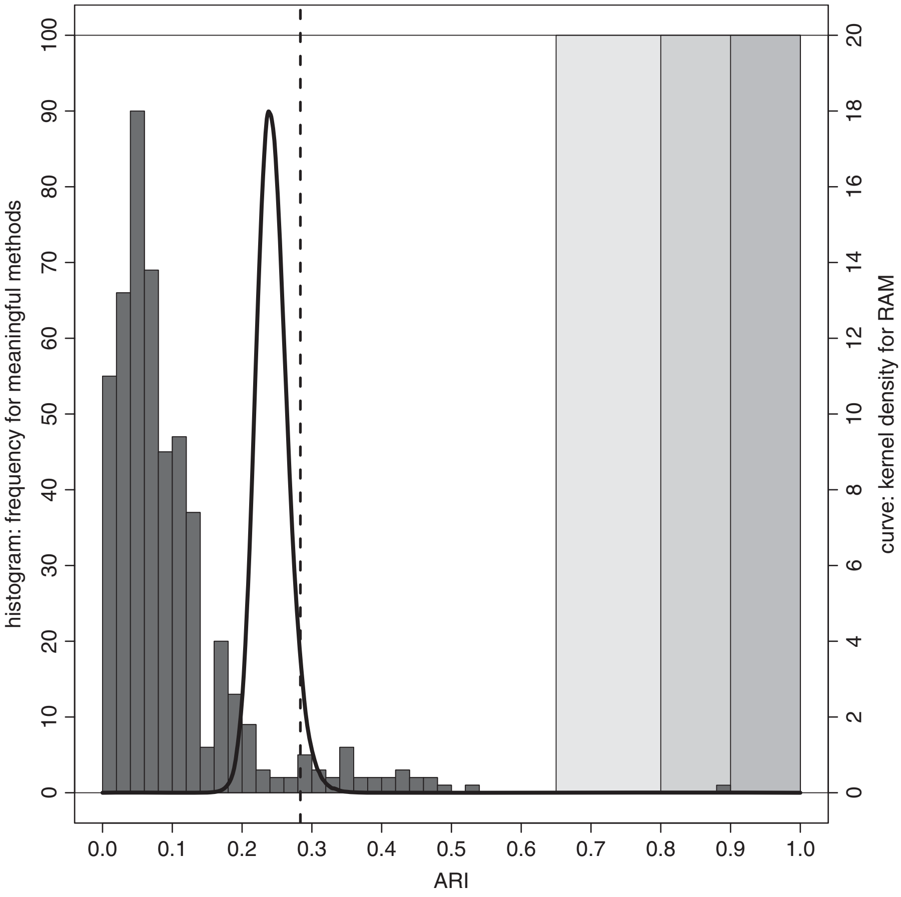

Figure 5. Distribution of adjusted Rand index values comparing deliberate classification methods (histogram, left axis) and classifications with randomly chosen circulation types (kernel density estimate curve, right axis) for the standard configuration. Grey boxes indicate the thresholds for moderate correspondence with $\mathrm{ARI_{HA}} > 0.65$ (light grey), good correspondence with $\mathrm{ARI_{HA}} > 0.80$ (intermediate grey) and excellent correspondence with $\mathrm{ARI_{HA}}>0.90$ (dark grey). All values $\mathrm{ARI_{HA}} < 0.65$ denote poor correspondence. The dashed vertical line indicates the 95th percentile of the RAM-$\mathrm{ARI_{HA}}$ values.

classifications with nine types (and less pronounced the classifications with 18 types). Apparently the smaller number of types leads to a higher chance for overlapping partitions. However, only outliers reach noteworthy similarity and the vast majority of cases fall below an $\mathrm{ARI_{HA}} < 0.65$. All pairs of methods showing at least in one case moderate correspondence ($\mathrm{ARI_{HA}}$ 0.65) are listed and ranked in Table 6. The number of cases consists of 12 spatial domains times 12 configuration variants, thus no combination shows moderate or better correspondence in the half of the cases (maximum of 63 out of 144 cases); however, SAN and SOM are distinctively more often (in 63 cases) similar than all others (below 27 cases). The methods of all pairs in the list are either from the group of optimization methods (OPT) or from the group of hierarchical clustering methods (HCA). It is noteworthy that hierarchical clustering methods show similarity only *within* the method group and not to non-hierarchical methods and vice versa. No other classifications show any noteworthy correspondence among all variants.

As a side note it can be observed that the distribution of the $\mathrm{ARI_{HA}}$ values of the RAM comparisons seems to justify the definition of Steinley (2004) calling $\mathrm{ARI_{HA}}$ values below 0.65 poor. In only one configuration variant (nine types) the RAM similarities exceed this threshold.

6. Discussion and conclusions

A comprehensive dataset of classification catalogs for 12 European domains has been compiled using a specially developed open source software package. The dataset and software allow the hitherto largest systematic comparison of synoptic classification algorithms and their configuration variants. Correspondence between the classification methods is examined by using fixed input datasets and only varying the classification algorithm. This is done for 12 different classification configuration variants and for 12 different spatial domains. Metrics reflecting the similarity between the circulation type patterns, the annual type frequency time series and the partitioning, by means of the adjusted Rand index, are used. Significance tests for the latter allow for exact stratification into poor, moderate, good or excellent correspondence between a pair of classification catalogs and to compare them to randomly chosen types.

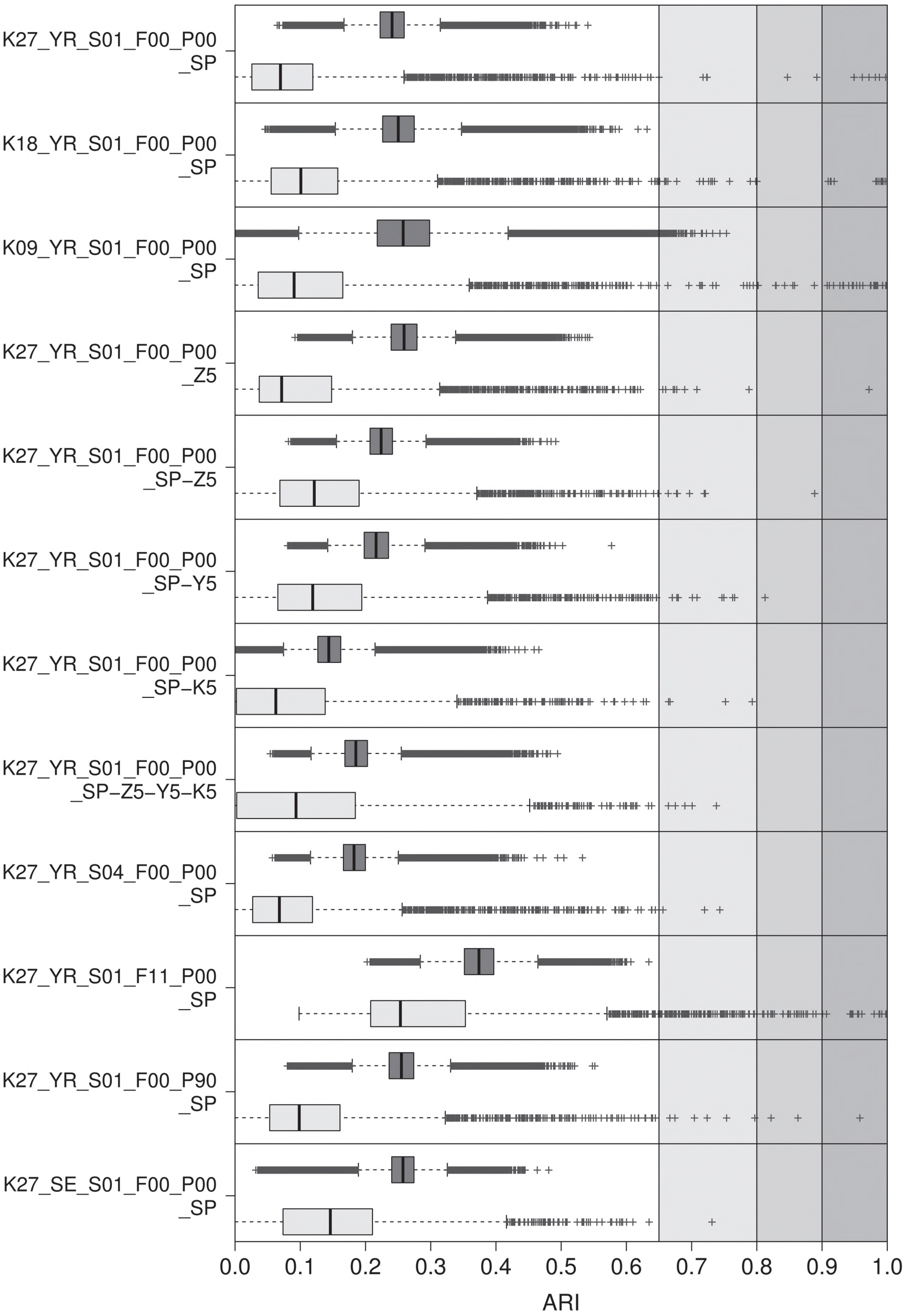

Figure 6. Adjusted Rand index values (abscissa) comparing the different classification methods with each other separately for the classification configurations (ordinate) as listed in Table 5. Light grey box plots include the values comparing deliberate methods for all 12 spatial domains. Dark grey box plots represent the $\mathrm{ARI}_{\mathrm{HA}}$ values for 1000 RAM classifications for all 12 spatial domains respectively. The maximum extension of the box-plot whiskers has been increased from 1.5 to 2 times the inter-quartile distance in order to reduce the number of outliers. Black vertical lines show the 95th percentile of the RAM-$\mathrm{ARI}_{\mathrm{HA}}$ distributions.

Surprisingly, there is nearly no considerable correlation between the methods and algorithms except for some hierarchical clustering methods (HCA) and except for the optimization methods (OPT). In case of HCA methods, the similarity is mainly due to very unequally large groups, which largely reduces the usefulness of the resulting catalogs. However, at least the methods of the OPT group, which differ in most cases only concerning the starting partition, should by definition come to same result. However, the nature of the k-means algorithm to stop in local minima of the optimization function leads often only to moderate or even poor correspondence. The only exception is the simulated annealing clustering (SAN) and the self-organizing map (SOM) approach, both are computationally very extensive algorithms designed to skip local minima and approach to the global optimum of the function minimizing within-type variance. Their correspondence reaches *good similarity* in individual variants and averages $\mathrm{ARI}_{\mathrm{HA}} = 0.7$ (*moderate* similarity), while k-means with random

Table 6. Pairs of methods showing noteworthy correspondence ($ARI_{HA} > 0.65$) ranked by the number of cases out of 12 domains times 12 classification configuration variants. All other pairs show poor correspondence ($ARI_{HA} < 0.65$). Note that all these cases are the same as the outliers with $ARI_{HA} > 0.65$ in Figure 6.

pair of methods	cases of ARI > 0.65
SOM-SAN	63
KMN-SAN	26
DKM-SAN	22
KMN-SOM	21
KMN-CAP	15
DKM-SOM	15
CAP-SAN	15
KMN-DKM	14
DKM-CAP	13
CAP-SOM	13
CKM-DKM	10
HSL-HCN	9
CKM-SAN	9
CKM-CAP	8
HAL-HCN	6
KMN-CKM	5
CKM-SOM	5
KMN-KMD	1
KMD-SOM	1
KMD-SAN	1
HSL-HMD	1
HSL-HAL	1
HMD-HCN	1
CKM-KMD	1
CAP-KMD	1

starting partitions (KMN) shows an even *poor* average correspondence to both of them. In terms of cases exceeding an $ARI_{HA} = 0.65$, SOM and SAN count 63 times moderate or better correspondence, while KMN and DKM count 26 and 22 times noteworthy similarity to SAN, out of 144 cases. All others correspond only seldom and are on average independent of one another (average $ARI_{HA} < 0.4$). Moreover, except of the mentioned cases, the similarity of the vast majority of pairs of methods is not significantly higher than that of classifications based on randomly defined circulation types.

In a further comparison experiment (not shown) very low similarity between classifications has also been found if the algorithm is fixed and only the data input and pre-processing is varied. Thus, if, e.g. the SP input data are either complemented or replaced by geopotential height of the 500 hPa level (Z5), the same method comes to very different classification results with Rand index values on a comparable low level as shown above.

Thus, the main message from this study is that almost all classification methods, applied for classification of circulation data, come to extremely different results.

In order to find reasons for this overall low level of similarity it is useful to look at the structure of the input data used for classification. Figure 7 shows the 16 436 days as dots in a scatter plot which is spanned by the x and y axis representing the scores of the first and second principal component of the SP data for domain D07 (Central Europe).

Looking at the upper panel of Figure 7 it might become clear why there is such a low level of correspondence between the methods: obviously there is no evidence of an inner structure of the data. Beside the centre of the cloud (the climatological mean) there are neither any areas of increased point density, indicating locations of preferred types nor areas of decreased density suggesting a border between two adjacent types. This explains why all data mining methods will fail to detect the same clusters of preferred occurrences of cases, simply because they are absent (see also Christiansen, 2002, 2007; Stephenson *et al.*, 2004; Philipp *et al.*, 2007 or Fereday *et al.*, 2008).

However, this cannot explain why a few of the hierarchical and a few of the non-hierarchical clustering methods show some correspondence at all. In case of the non-hierarchical clustering methods the explanation is the so-called snow-balling effect, i.e. if there is no structure in the dataset non-hierarchical cluster analysis tends to produce one large type and very small remaining types. In case of the single linkage algorithm (HSL) this leads, e.g. for the standard configuration to a size of 16 408 days for the first class and type sizes of 2 or 1 for all the remaining types. It is clear that if any other method shows such a behaviour that the overlap between partitions is large which is accounted for by the similarity metrics. Thus, the hierarchical clustering methods achieve similar results but they are obviously not useful for applications in synoptic climatology, when applied to datasets as done in this study (note that hierarchical clustering has successfully been applied to a sample of other meteorological variables, e.g. by Kalkstein *et al.*, 1987). However the snow-balling effect cannot explain the similarity of the non-hierarchical methods. As illustrated in the lower panel of Figure 7 the type sizes are generally rather equally distributed, which is also the case for application in higher dimensional phase spaces. This means that some feature of the input data should exist determining a more or less similar partitioning.

Looking at the shape of the point cloud in Figure 7 (upper panel) it becomes apparent that it is not a strict circular shape but, e.g. covers a larger area on the left half of the plot (quadrants II and III) compared to the right half (quadrants I and IV). Accordingly the marginal distributions are both positively skewed (the PC 1 score distribution only slightly). For algorithms trying to minimize within-type variance, such irregularities of the shape and distribution constrain the position of boundaries between the types. This is demonstrated in Figure 8 where the artificially deformed point cloud shows extreme irregularities in shape and therefore a clear constraint for classification. In this case all methods from the OPT group will come to closely related results.

In reality, these irregularities are distinctively smaller and only those algorithms which are able to find highly

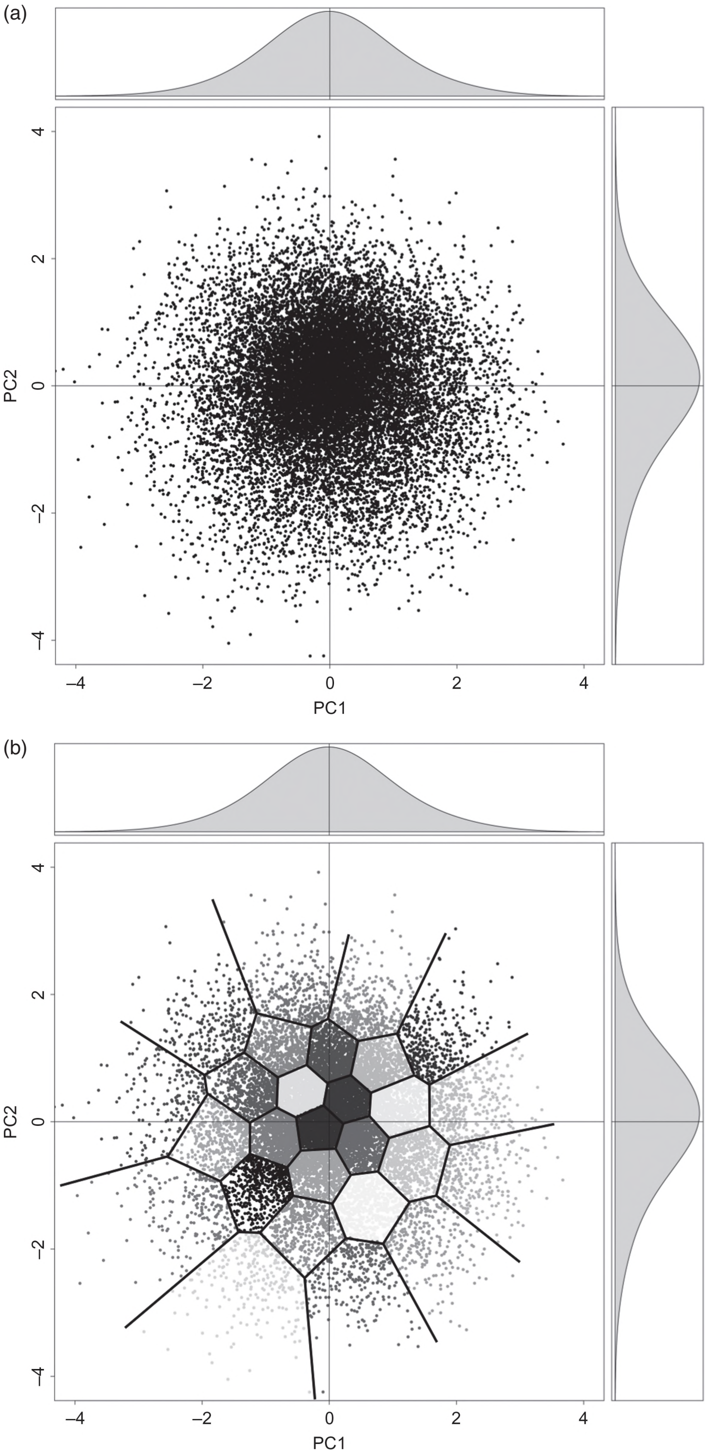

Figure 7. Scatter plot of sea level pressure data in domain D07 for all 16 436 days (points). The position of each point in this two-dimensional phase space is defined by the scores of the first two principal components (PCs) of the data. Upper panel: point cloud distribution supplemented by marginal kernel density distributions (filled curves, scaled to maximum density of both). Lower panel: the same data but classified by non-hierarchical cluster analysis of the scores for 27 types. Classification of each point is indicated by grey scales and boundary lines.

optimized solutions show noteworthy correspondence. Actually this is the case for SAN and SOM, both are designed to overcome local minima in the optimization function, while e.g. the k-means algorithm often converges within local minima and is therefore unable to fit the partitioning to the irregularities in the shape of the cloud.

Another question is why the methods using pre-defined types are so dissimilar. A possible answer may be also the missing inner and weak outer structure of the point cloud. If we think of the type definition as the drawing of a boundary plane between the types in a very dense multidimensional point cloud, then small changes in the position and orientation of the plane will result in a

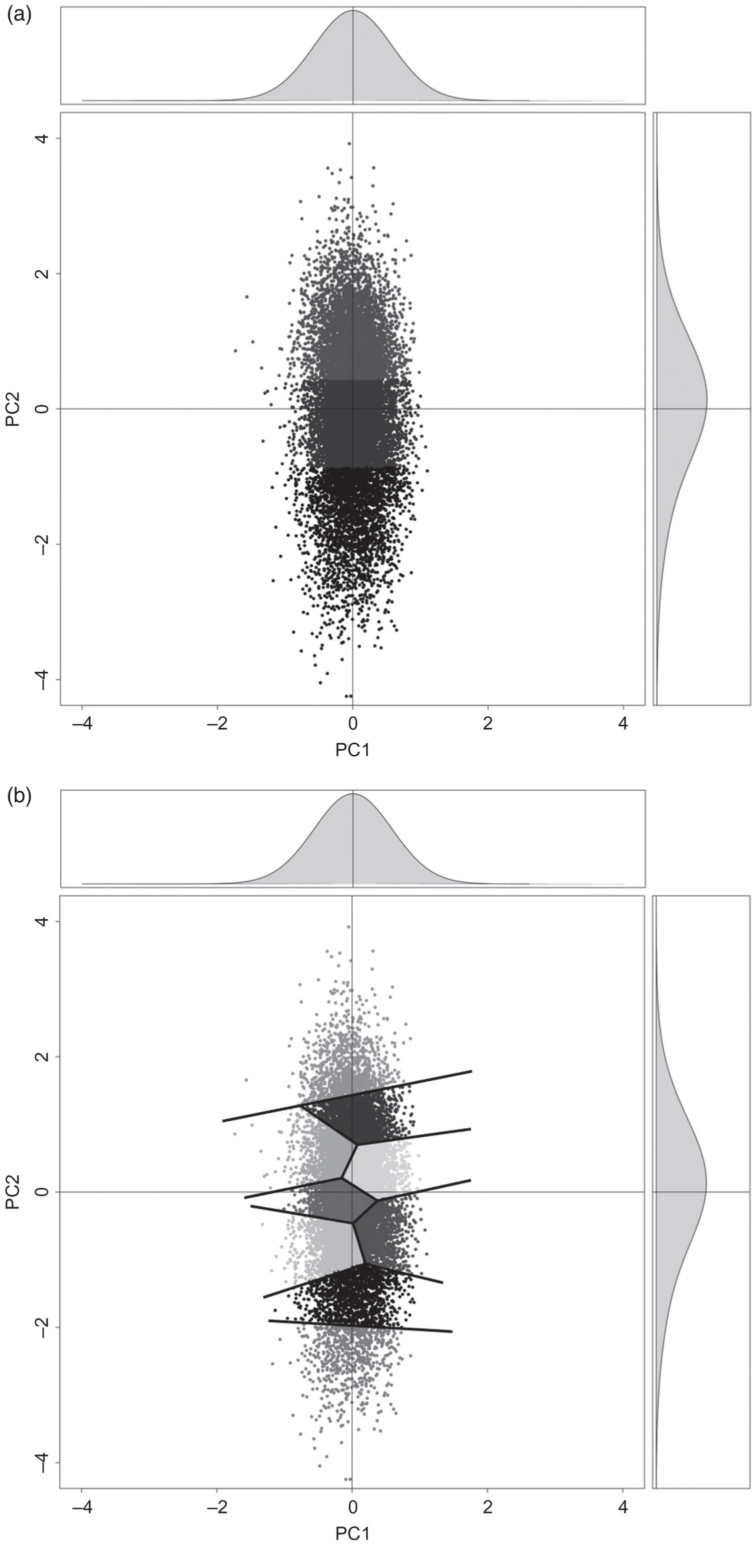

Figure 8. Illustration of type definition constraint by the shape of the data point cloud: scatter plot of artificially transformed sea level pressure data (scores of PC 1 of Figure 7 scaled by 0.3) classified by non-hierarchical cluster analysis using three (upper panel) and nine classes (lower panel). Classification of each point is indicated by grey scales and boundary lines.

relatively large change in the population of the types. Thus if, e.g. there is no clear preference of westerly type situations in the data and if the transitions to neighboured situations (into all directions of the multidimensional space) are smooth and seamless, high diversity between the referring types of two classifications based on slightly different type definitions may occur. Further, different threshold metrics are used by the threshold-based methods, like correlation coefficients to prototype patterns in GWT, wind directions and pressure in WLK, gradients in JCT and so on. This would lead to small differences if the areas of the boundaries between two types would be clear and sparsely populated. However in a dataset without such structures the opposite is the case.

Thus, the generally observed low similarity can be explained by the low degree of structure in the datasets, leading to the question of whether it is suitable in synoptic climatology to apply classification methods on such datasets at all or in other words whether such datasets are classifiable. The opinion of the authors to this question is clearly 'yes', every sample may be classified or partitioned. Classification methods generally help to reduce the complexity of datasets and are indispensable for finding systematic relations, regardless of whether the data show structure or not. It is always useful to focus parts of a sample (i.e. circulation types in this case) rather than to always consider only the whole of a collective of objects. Thus any classification as shown above may be used in a descriptive way. However this should be done without the expectation that any definition of the types has, from the statistical point of view, a higher justification than others. This point of view indeed may lead to a change of paradigm concerning the role of circulation type classifications in synoptic climatology.

Nevertheless, any classification may be useful even though it will be independent from others. On the one hand this is true for applications where the precise partition decision itself may not be the most critical process. For example, if the main flow direction and its impact, e.g. for air pollution, is in the focus, it is less relevant how many categories include this direction and how they look like in detail. Instead, the most important outcome would be to identify the overall weather situation with impact on the target variable represented by the categories. On the other hand, the large diversity of classification catalogs may offer an increased probability to find a well-suited classification also for applications where the detail matters, e.g. for precipitation analysis. For this purpose the cost733class software package which was used in this study offers the opportunity to develop and evaluate specialized classification catalogs for all kind of synoptic analysis. In addition to the catalog dataset it is available at http://cost733.geo.uni-augsburg.de.

References

Barry RG, Carleton AM. 2001. *Synoptic and dynamic climatology*. Routledge: London and New York, NY, 620 pp.

Beck C. 2000. Zirkulationsdynamische Variabilität im Bereich Nordatlantik-Europa seit 1780 (variability of circulation dynamics in the North-Atlantic-European region). *Würzburger Geographische Arbeiten* **95**: 350 pp.

Beck C. 2012. Are there weekly cycles in occurrence frequencies of large-scale circulation types? *Atmos. Sci. Lett.* **13**: 238–243.

Beck C, Jacobeit J, Jones PD. 2007. Frequency and within-type variations of large scale circulation types and their effects on low-frequency climate variability in central Europe since 1780. *Int. J. Climatol.* **27**: 473–491.

Blair D. 1998. The Kirchhofer technique of synoptic typing revisited. *J. Climatol.* **18**: 1625–1635.

Christiansen B. 2002. On the physical nature of the Arctic Oscillation. *Geophys. Res. Lett.* **29**: 1805, doi: 10.1029/2002GL015208.

Christiansen B. 2007. Atmospheric circulation regimes: Can cluster analysis provide the number? *J. Clim.* **20**: 2229–2250.

Crane RG, Barry RG. 1988. Comparison of the msl synoptic pressure patterns of the arctic as observed and simulated by the GISS general circulation model. *Meteorol. Atmos. Phys.* **39**: 169–183.

Dittmann E, Barth S, Lang J, Müller-Westermeier G. 1995. Objektive Wetterlagenklassifikation (objective weather type classification). *Ber. Dt. Wetterd.* 197.

Dutilleul P. 1993. Modifying the t-test for assessing the correlation between two spatial processes. *Biometrics* **49**: 305–314.

Enke W, Spekat A. 1997. Downscaling climate model outputs into local and regional weather elements by classification and regression. *Clim. Res.* **8**: 195–207.

Erpicum M, Mabille G, Fettweis X. 2008. Automatic synoptic weather circulation types classification based on the 850 hpa geopotential height. In *Abstracts COST 733 Mid-term Conference, Advances in Weather and Circulation Type Classifications & Applications 22â"25 October 2008 Krakow, Poland*, 33.

Esteban P, Jones PD, Martin-Vide J, Mases M. 2005. Atmospheric circulation patterns related to heavy snowfall days in Andorra, Pyrenees. *Int. J. Climatol.* **25**: 319–329.

Fereday DR, Knight JR, Scaife AA, Folland CK, Philipp A. 2008. Cluster analysis of North Atlantic/European circulation types and links with tropical Pacific sea surface temperatures. *J. Clim.* **21**(15): 3687–3703.

Hartigan J. 1975. *Clustering Algorithms*. *Wiley Series in Probability and Mathematical Statistics*. John Wiley: New York, NY; 351.

Hess P, Brezowsky H. 1952. Katalog der Großwetterlagen Europas (catalog of the European large scale weather types). *Ber. Dt. Wetterd. in der US-Zone* **33**: 39 pp.

Hubert L, Arabie P. 1985. Comparing partitions. *J. Classif.* **2**: 193–218.

Huth R. 1993. An example of using obliquely rotated principal components to detect circulation types over Europe. *Meteorol. Z.* **2**: 285–293.

Huth R. 1996. An intercomparison of computer-assisted circulation classification methods. *Int. J. Climatol.* **16**: 893–922.

Huth R, Beck C, Philipp A, Demuzere M, Ustrnul Z, Cahynová M, Kyselý J, Tveito O-E. 2008. Classifications of atmospheric circulation patterns: recent advances and applications. *Ann. N. Y. Acad. Sci.* **1146**: 105–152.

Jenkinson AF, Collison FP. 1977. An initial climatology of gales over the North Sea. *Synoptic Climatology Branch Memorandum 62, Meteorological Office, Bracknell*.

Jolliffe I, Philipp A. 2010. Some recent ideas in cluster analysis. *Phys. Chem. Earth* **35**: 309–315.

Jones PD, Hulme M, Briffa KR. 1993. A comparison of lamb circulation types with an objective classification scheme. *Int. J. Climatol.* 655–663.

Kalkstein LS, Tan G, Skindlov JA. 1987. An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Clim. Appl. Meteorol.* **26**: 717–730.

Kaufman L, Rousseeuw PJ. 1990. *Finding Groups in Data an Introduction to Cluster Analysis*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley: New York, NY; 342.

Kirchhofer W. 1974. Classification of European 500 mb patterns. *Arbeitsbericht der Schweizerischen Meteorologischen Zentralanstalt, Zurich, Switzerland* **43**: 1–16.

Kruizinga S. 1979. Objective classification of daily 500 mbar patterns. In *Preprints Sixth Conference on Probability and Statistics in Atmospheric Sciences, Banff, Alberta*., American Meteorological Society, Boston, MA, 126–129.

Lamb HH. 1972. British Isles weather types and a register of daily sequence of circulation patterns, 1861–1971. *Geophys. Memoir.* **116**: 85L.

Lauscher F. 1985. Klimatologische Synoptik Österreichs mittels der ostalpinen wetterlagenklassifikation (synoptic climatology of Austria based on the eastern- alpine weather type classification). *Arbeiten aus der Zentralanstalt für Meteorologie und Geodynamik* **64**: 65 pp.

Legendre P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology* **74**: 1659–1673.

Litynski J. 1969. A numerical classification of circulation patterns and weather types in Poland. *Prace Panstwowego Instytutu Hydrologiczno-Meteorologicznego* **97**: 3–15.

Lund IA. 1963. Map-pattern classification by statistical methods. *J. Appl. Meteorol.* **2**: 56–65.

McQuitty L. 1966. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educ. Psychol. Meas.* **26**: 825–831.

Michaelides S, Liassidou F, Schizas C. 2007. Synoptic classification and establishment of analogues with artificial neural networks. *Pure Appl. Geophys.* **164**: 1347–1364.

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

Murtagh F. 1985. *Multidimensional Clustering Algorithms, Volume 4 of COMPSTAT Lectures*. Physica-Verlag: Würzburg.

Peczely G. 1957. Grosswetterlagen in Ungarn. *Kleinere Veröffentlichungen der Zentralanstalt für Meteorologie* **30**: 86 pp.

Perret R. 1987. Une classification des situations météorologiques à usage de la prévision (a classification of meteorological situations for use in prediction). *Veröffentlichungen der schweizerischen Meteorologischen Anstalt* **46**: 127 pp.

Philipp A. 2009. Comparison of principal component and cluster analysis for classifying circulation pattern sequences for the European domain. *Theor. Appl. Climatol.* 31–41.

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

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

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

Rand WM. 1971. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**: 846–850.

Schueepp M. 1979. Witterungsklimatologie - Klimatologie der Schweiz iii (weather climatology - climatology of Switzerland iii). *Beihefte zu den Annalen der Schweizerischen Meteorologischen Anstalt*, 93.

Stehlík J, Bárdossy A. 2003. Statistical comparison of European circulation patterns and development of a continental scale classification. *Theor. Appl. Climatol.* **76**: 31–46.

Steinley D. 2004. Properties of the Hubert-Arabie adjusted rand index. *Psychol. Methods* **9**: 386–396.

Stephenson D, Hannachi A, O'Neill A. 2004. On the existence of multiple climate regimes. *Q. J. R. Meteorol. Soc.* **130**: 583–605.

Uppala SM, Kallberg PW, Simmons AJ, Andrae U, Bechtold VDC, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, Berg LVD, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, Mcnally AP, Mahfouf J-F, Morcrette J-J, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**: 2961–3012.

Ward JH. 1963. Hierachical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**: 236–244.

Werner PC. 2002. Zur Berücksichtigung der Persistenz in meteorologischen Zeitreihen. *PIK Rep.* **75**: 43–54 (in German).

Yarnal B. 1993. *Synoptic Climatology in Environmental Analysis*. Belhaven Press: London.

Yarnal B, Comrie AC, Frakes B, Brown DP. 2001. Developments and prospects in synoptic climatology. *Int. J. Climatol*. **21**: 1923–1950.