Evaluation and comparison of circulation type classifications for the European domain

Christoph Beck *, Andreas Philipp

University of Augsburg, Institute of Geography, Physical Geography and Quantitative Methods, Universitätsstr. 10, D-86135 Augsburg, Germany

1. Introduction

First and foremost circulation type classification (CTC) is a method for categorizing the continuum of atmospheric circulation into a reasonable and manageable number of discrete circulation types. Thus CTCs provide a tool for analyzing the variability of atmospheric circulation in terms of frequency changes of resulting circulation types on different temporal and spatial scales (e.g. Kysely and Huth, 2006; Schmutz and Wanner, 1998). Furthermore CTCs are used for investigating the relationship between the large-scale atmospheric circulation and regional to local scale climate parameters like temperature or precipitation (e.g. Beck et al., 2007; Jacobeit et al., 2003) and as well various non-climatic environmental variables (e.g. Demuzere et al., 2008). A comprehensive overview of different methodological approaches for CTC and their application in different fields of research is provided in a recent review paper (Huth et al., 2008). In view of the large variety of different CTCs in use Huth et al. (2008) state the need for the systematic evaluation and comparison of the varying CTCs in order to identify

* Corresponding author. Tel.: +49 821 5982129; fax: +49 821 5982264.
E-mail address: christoph.beck@geo.uni-augsburg.de (C. Beck).

general strengths and weaknesses of individual classification approaches and superordinate basic approaches. Such evaluation and comparison studies focusing on varying properties of CTCs are currently performed within the framework of the COST733 Action "Harmonisation and Applications of Weather Types Classifications for European Regions" (refer to http://cost733.org/ for more information). The main goal of these investigations is to quantify the performance of individual CTCs and basic methods in order to provide recommendations for the application of CTCs and for the development of new CTCs.

Different approaches towards the evaluation and comparison of CTCs are possible. For quantifying the similarity between two classifications (or more generally spoken – two partitions of one set of observations) in terms of concurrent assignment of individual atmospheric fields to circulation types several similarity indices can be calculated (e.g. Rand, 1971; Hubert and Arabie, 1985; Milligan and Cooper, 1986; Strehl and Gosh, 2002). As an alternative to the above-mentioned classical measures for comparing partitions the Shannon Entropy (Shannon, 1948) as a measure for the statistical dispersion of a discrete random variable has been utilized by James (2007) to examine the degree of correspondence between two CTCs. However the estimation of such metrics for CTCs (e.g. Stehlik and Bardossy, 2003; James, 2007; Huth et al., 2008) yields no information on what classification is better than the others as the natural partitioning of atmospheric circulation which could be used as a benchmark for performance assessment is unknown or may not even exist.

A rather pragmatic approach is to analyze how different CTCs perform when used in specific applications. Examples for such comparison studies related to climatological analyses can be found in Buishand and Brandsma (1997) or Philipp (2008). However most often only few classifications were compared within such application studies thus enabling no comprehensive assessment of the large variety of CTCs.

Basically the quality of a classification can be defined as its ability to arrange entities into groups (classes) that feature maximum internal homogeneity and at the same time maximum external dissimilarity. Accordingly the quantitative evaluation and subsequent comparison of CTCs may be performed by the estimation of variability within classes (circulation types) and separability between classes on the basis of suitable metrics. The so far most comprehensive comparison study of CTCs incorporating – among other classification characteristics – the basic property of separability between classes has been carried out by Huth (1996) for a sample of variants of five computer-assisted classification methods. Confined to the variable used for classification (700 hPa geopotential heights) Huth (1996) concluded that CTCs based on *k*-means cluster analysis are more powerful than other classification approaches concerning separability among resulting circulation types. However focusing on other characteristics other methods turn out to be superior (e.g. t-mode principal component with regard to the reproduction of predefined circulation types) pointing to the fact that there is no classification that performs best in all aspects.

In consideration of the fact that CTCs are employed in a wide variety of synoptic climatological studies trying to establish relationships between large-scale circulation dynamics and environmental variables it is desirable to quantify separability and within-type variability of CTCs not only with regard to the parameter utilized for classification (baric fields from different atmospheric levels) but as well for different associated variables. The relevance of within-type variability for the analyses of circulation–climate relationships using CTCs and the need for reducing climatic within-type variability have been stressed in several studies (e.g. Brinkmann, 1999; Beck et al., 2007).

Against this background evaluation and comparison studies within the framework of the COST733 Action focusing on separability and within-type variability as basic properties of CTCs have been extended in several ways compared to above-mentioned former studies. A selection of statistical metrics for estimating separability and within-type variability is determined for a comprehensive and consistent sample of daily CTCs for the period from September 1957 to August 2002 that has been developed within the COST733 Action (Philipp et al., 2010) covering a wide range of classification approaches. Evaluations are performed on the basis of one consistent data set not only for the parameter used for classification (mean sea level pressure) but as well for associated surface climate variables. To account for spatial variations in the performance of classifications all analyses are conducted for a number of spatial domains of varying size and location and furthermore – if possible – evaluation metrics are calculated not only over the whole domain but for individual locations (grid points) as well.

The paper is structured as follows. The data used for the evaluation and comparison studies are briefly introduced in Section 2. Section 3 illustrates the methodological approach. Selected results from the ongoing investigations within the framework of the COST733 Action are presented in Section 4. Finally in Section 5 main findings are discussed and preliminary conclusions are outlined.

2. Data

Two data sets have been used for performing the evaluation and comparison studies presented in this paper.

2.1. Circulation type data

Daily catalogues providing the occurrence of circulation types resulting from 73 different CTCs are available from the COST733-CAT-database of weather and circulation type classifications that has recently been developed within the framework of the COST733 Action (see Philipp et al., 2010 for a detailed description of the data set). The circulation type data cover the period from September 1957 to August 2002 and are provided – for each classification – for 12 spatial domains of varying size and position embedded into the greater North Atlantic European region (Fig. 1). A wide variety

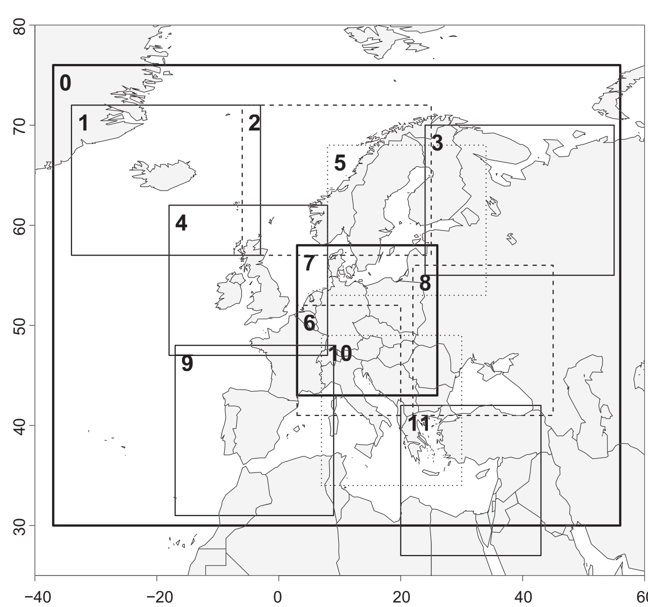

Fig. 1. Location and extension of the 12 spatial domains (numbers 0–11) for which circulation type classifications from the COST733CAT-database are available.

of classification approaches is covered by this comprehensive data set including classification methods based on optimization algorithms (mainly non-hierarchical cluster analysis), principal component analysis (e.g. t-mode principal component analysis) or so called leader algorithms (e.g. pattern correlation), methods that involve the definition of thresholds (e.g. concerning main wind directions) and finally subjective (or manual) classifications like for example the "Hess–Brezowski Großwetterlagen" (Hess and Brezowski, 1952).

The inclusion of such a wide variety of classifications based on different general approaches, the availability of all classifications for different spatial domains and the fact that all automated classifications have been determined using the same database (ERA-40 reanalysis data, see Uppala et al., 2005) and are all available for fixed numbers of classes (9, 18 and 27 classes) makes the COST733CAT-database a unique starting basis for the evaluation and subsequent comparison of individual CTCs and basic classification concepts.

2.2. Data used for evaluations

For evaluation purposes gridded daily 12 UTC data for mean sea level pressure (MSLP), 2 m temperature (2mT), convective precipitation (CP) and stratiform (large scale) precipitation (LSP) have been extracted from the ERA-40 reanalysis data set (Uppala et al., 2005) covering the period from September 1957 to August 2002. Reanalysis data are used with spatial resolutions of 2° (latitudinal) by 3° (longitudinal) for the large domain and 1° by 1° for the smaller sub-domains. Despite some deficiencies especially of ERA-40 precipitation data on different spatial and temporal scales (e.g. Bosilovich et al., 2008; Omstedt et al., 2005; Uppala et al., 2005) the reanalysis data have been shown to provide reasonable estimates for temperature and precipitation in different parts of the North Atlantic European region (e.g. Martin, 2004; Crochet, 2007). Moreover with respect to the analyses presented in this contribution the ERA-40 data provides the opportunity to perform evaluation and comparison studies for all spatial domains and all variables using only one consistent data set.

3. Methods

Separability among circulation types and within-type variability of types are quantified through the calculation of a selection of statistical metrics on the basis of daily gridded ERA-40 reanalysis data for the variable used for classification (MSLP) and as well for the associated surface climate variables 2 m temperature (2mT) and CP and LSP. Precipitation totals (PREC) for 12 UTC have been determined as the sum of convective (CP) and stratiform (LSP) precipitation. Prior to the calculation of evaluation metrics all data (circulation and surface climate variables) have been rescaled to daily anomalies (daily value minus the long-term daily mean) in order to remove the seasonal cycle. Eliminating the seasonal cycle prevents the undesirable influence of pronounced seasonal frequency variations of certain circulation types.

The task of evaluating the quality of CTCs by estimating indices for separability and within-type variability resembles the effort towards determining the optimum number of types (usually defined as the solution with lowest/highest values of within-type variability/separability of classes respectively) when performing a classification. In fact the main difference is that the first approach analyses individual partitions provided by varying methods while the latter focuses on varying partitions resulting from one single method. Thus several of the metrics used for evaluating classifications in the context of this contribution are as well commonly used for estimating the most appropriate number of types especially for classifications based on cluster analysis (e.g. Milligan and Cooper, 1985).

In detail the following metrics for determining separability and within-type variability of CTCs are calculated.

The explained variation (EV) is estimated on the basis of the ratio of the sum of squares within classes (circulation types) (WSS) and the total sum of squares (TSS),

$$EV = 1 - \frac{WSS}{TSS} \tag{1}$$

According to Calinski and Harabsz (1974) the so called Pseudo-F statistic (PF) is calculated as the ratio of the sum of squares between classes (BSS) and the sum of squares within classes (WSS) additionally taking into consideration the number of cases (n) and classes (k).

$$PF = 1 - \frac{BSS/(k-1)}{WSS/(n-k)} \tag{2}$$

A measure for the separability among circulation types that has been proposed by Huth (1996) is the pattern correlation ratio (PCR). Here PCR is defined as the ratio of the mean pattern correlation (Pearson's r) between days (daily gridded fields) assigned to the same circulation type (PCI) and the mean pattern correlation between days assigned to different circulation types (PCO).

$$PCR = \frac{PCI}{PCO} \tag{3}$$

Kalkstein et al. (1987) utilized the simple arithmetic mean of the class-specific standard deviations (SDI) as an estimator of the within-type standard deviation (WSD) for the comparison of synoptic climatological classifications. Here WSD is calculated as the pooled standard deviation in order to take into account differing sizes (n) of classes (k).

$$WSD = \sqrt{\frac{\sum_{k=1}^{K}(n_k - 1)\cdot SDI_k^2}{\sum_{k=1}^{K}(n_k - 1)}} \tag{4}$$

To get an estimate of the range of uncertainty of the variable's mean values associated to each classification, the confidence interval of the mean (CIM) has been calculated as the weighted mean (utilizing class sizes as weights) of the type specific confidence intervals of the mean for $\alpha = 0.05$ (the choice of α is arbitrary for comparison purposes but allows for the use-oriented characterization of individual circulation types).

$$CIM = \frac{\sum_{k=1}^{K} z_{1-\alpha/2}\cdot\frac{SDI_k}{\sqrt{n_k}}\cdot n_k}{\sum_{k=1}^{K} n_k} \tag{5}$$

For evaluating the quality of cluster separation Rousseeuw (1987) proposed the calculation of the so called Silhouette index (SIL). However the calculation of the Silhouette index according to Rousseeuw (1987) is rather CPU-intensive when applied to large data sets. Thus a modified approach has been used for estimating a "faster Silhouette Index" (FSIL). The difference between FSIL and SIL is that for FSIL for any case (day, i) the distances to its own class (fa_i) and its nearest neighboring class (fb_i) are estimated as the euclidean distances to the respective class centroids and not as for SIL as the average distance between the case and all cases in its own class and its closest class respectively.

$$FSIL = \frac{1}{n}\sum_{i=1}^{n}\frac{fb_i - fa_i}{\max(fa_i, fb_i)} \tag{6}$$

Several test runs have shown that FSIL results are closely related to results achieved by applying the original SIL.

With the exception of WSD and CIM all metrics combine the two basic properties – separability among classes and within-type variability – in their estimates of the quality of CTCs. All six metrics are calculated for all 73 CTCs included in the COST733CAT-database, for each of the 12 spatial domains, for each of the five variables (MSLP, 2mT, CP, LSP, PREC) and separately for the 12 months of the year, the four 3-month seasons (DJF, MAM, JJA, SON) and for the whole year. Thus all in all 6120 evaluation index samples each consisting of 73 values (according to 73 CTCs) are generated. Standardization of each index sample allows for the comparison and the unbiased combination of index values derived from different samples. If necessary the sign of the standardized index values has been changed in order to attain maximum positive values for those CTCs with minimal within-type variability and maximum separability among types respectively.

This set of evaluation indices provides information on the performance of each classification in terms of within-type variability and separability of circulation types integrating over the whole spatial field (all grid points of the considered domain). Thus spatial variations in evaluation criteria are so far only reflected by respective differences between the 12 domains. In order to gain a more detailed picture of the spatial variations within domains selected evaluation criteria for which estimates at specific locations are possible (EV, PF, WSD, CIM) are also calculated for individual grid points. In addition, a characterization of superordinate groups of basic approaches is attained in terms of evaluation criteria by averaging the evaluation metrics over the individual members of the five groups of basic classification approaches (subjective, threshold based (THRES), principal component analysis (PCA), leader algorithms (LEAD), optimization algorithms (OPT)). Finally rankings of CTCs and superordinate groups are achieved for varying summarizations over domains, seasonal subsets, variables and evaluation criteria by averaging the respective ranks estimated separately for each index sample. All analyses have been performed for the whole period of available data (1957–2002) and as well for three shorter subperiods (1957–1977, 1970–1990, 1982–2002) in order to check the results for their robustness in time.

4. Results

This section presents results of the application of the evaluation criteria explained above to the data sets described in Section 2. Due to limited space it is not intended to give a comprehensive overview of all results achieved within the COST733 Action but to focus on a few selected topics pointing out some main findings mainly based on evaluation results achieved by combining results from all criteria and specific results for EV.

4.1. Sensitivity of evaluation metrics against the number of circulation types

A crucial point concerning the informative value of the applied evaluation criteria for comparative studies is their sensitivity against the number of classes (circulation types) included in a CTC. Fig. 2 shows boxplots of standardized values of evaluation indices integrating over all evaluation criteria, all variables, all spatial domains and seasonal subsets grouped according to the number of circulation types. Obviously values of all evaluation metrics feature distinct dependency on the number of circulation types. Values of EV and WSD increase with higher numbers of circulation types while PF, FSIL and CIM exhibit relationships of the opposite sign. According to Fig. 2 PCR is the only evaluation metric showing no clear cut dependency on the number of circulation types. Hence the comparison of evaluation results for classifications with distinctly differing numbers of circulation types appears to be rather a comparison of the impact of the number of types than the effect of different classification approaches. Therefore in the following the focus is on a subset of CTCs having similar numbers of circulation types. The selected 16 CTCs are listed in Table 1 with their abbreviations and information on the number of types included, the variable used for classification, the underlying basic approach and the main reference. All classifications comprise approximately 18 circulation types and with the exception of WLKCC18 all classifications utilize solely MSLP data for classification (see Philipp et al., 2009 and references in Table 1 for detailed descriptions of the classification methods). Four classifications (THRES) utilize predefined thresholds (mostly applied to indices for flow direction and vorticity derived from gridded MSLP data) for assigning individual cases into classes. Three classifications are based on s-mode or t-mode principal component analysis (PCA) and three methods use variants of the so called leader algorithm (LEAD) for the classification of circulation types. Finally the largest group of six classifications is based on optimization algorithms (OPT – variants or modifications of the *k*-means cluster algorithm, partly in combination with preceding S-mode PCA). Hence representatives of the most commonly used approaches for objective CTC are included in this 18 classes subsample of the COST733CAT-database (Philipp et al., 2009).

4.2. Some general aspects of the performance of circulation type classifications

First of all in order to get an impression of the general performance of CTCs Fig. 3 illustrates evaluation results in terms of EV for the classification ensemble from Table 1 depending on the month of the year, the spatial domain and the target variable (note that evaluation criteria have been standardized over all months, all domains or all variables respectively). Kruskal–Wallis (K–W) tests have been applied to each of the three categorizations of EV values from Fig. 3 to test for respective significant differences. Significant ($p < 0.001$) differences in median of EV values have been determined for all three categorizations. Additionally multiple Mann–Whitney (M–W) tests with Bonferroni correction of critical values (to account for increased probability of Type I errors) have been used in order to detect significant ($p < 0.05$) differences between categories.

Fig. 3a depicts overall best performance for winter, distinctly lower values of EV during summer and intermediate EV values during the transition seasons spring and autumn. Referring to pairwise M–W tests three broad groups of categories (months) may be identified according to mainly non-significant/significant differences within/among these groups: November–February; May–August; March, April, September, October. These differences are likely to be related to: on the one hand to the intensive coupling between large-scale circulation and surface climate in winter and on the other hand to the increasing climatic importance of local-scale processes during the summer half-year that are not captured by CTCs. Concerning differences in EV between spatial domains (see Fig. 1) it can be seen from Fig. 3b that EV values are considerably lower for the large North Atlantic-European domain 0 than for the smaller sub-domains (domains 1–11). However EV differences are as well obvious among sub-domains. Better performance in terms of EV can be seen for the more westerly and maritime domains (Iceland, W Scandinavia, British Isles, Baltic, W Mediterranean) while EV values are somewhat lower for the eastern and more continental domains (Central Europe, NE Europe, E and SE Europe, E Mediterranean) due to a stronger influence of orography and land surface characteristics. Comparing domain 7 (Central Europe) and domain 6 (Alps) – two domains with comparable location but differing in size – higher EV values can be detected for the smaller domain. The arrangement of domains 1–11 into two groups according to differing EV values is also confirmed by the results of pairwise M–W tests showing significant differences for all pairwise comparisons between groups while the majority of pairwise

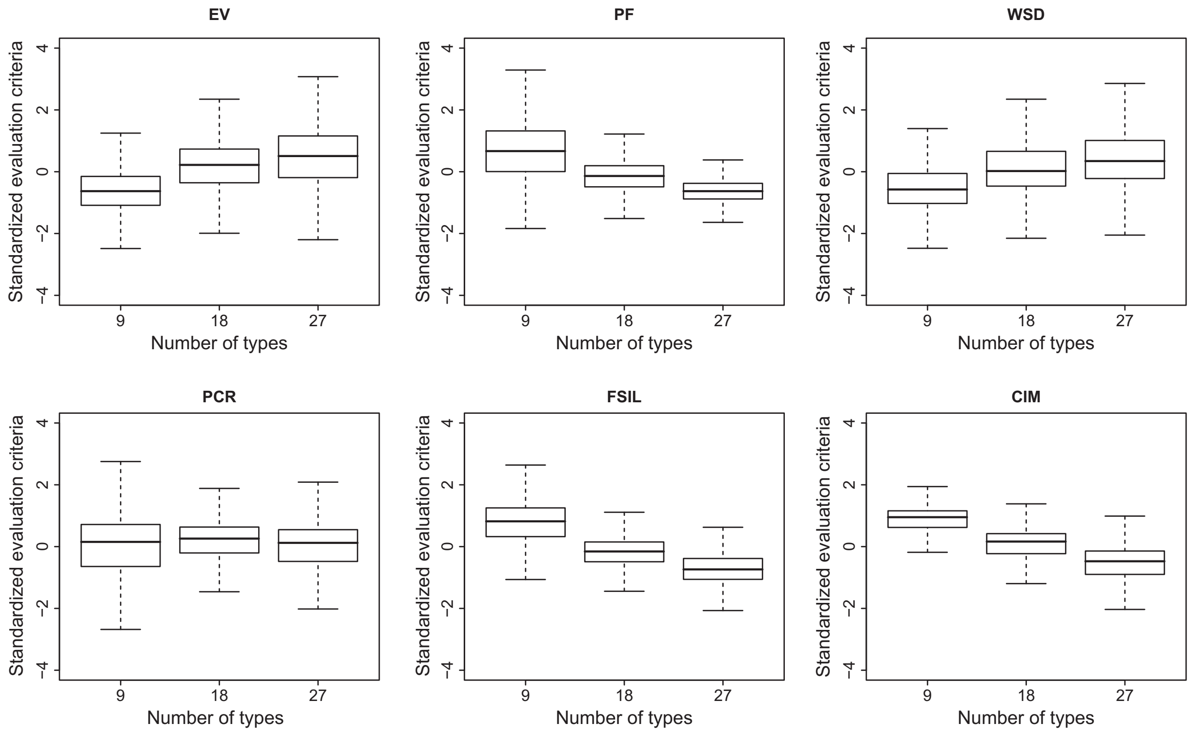

Fig. 2. Boxplots of standardized evaluation metrics (EV, PF, WSD, PCR, FSIL, CIM, see text for explanation) for circulation type classifications from the COST733CAT-database, grouped according to the number of circulation types (within the range of ±3 the numbers indicated at the *x*-axis). Boxplots include respective samples of evaluation indices estimated for 12 spatial domains, 17 seasonal subsets and 5 variables for the period 1957–2002. Upper/lower whiskers indicate the 1.5 interquartile range (IQR) from the upper/lower quartile.

Table 1
Selected circulation type classifications from the COST733CAT-database (Philipp et al., 2010).

Abbreviation	Number of types	Variable used for classification	Basic classification approach	References
GWTC18	18	MSLP	THRESHold based	Beck et al. (2007)
LITC18	18	MSLP	THRESHold based	Litynski (1969)
LWT2C18	18	MSLP	THRESHold based	James (2006)
WLKCC18	18	U700, V700, Z925	THRESHold based	Dittmann et al. (1995)
P27C16	18	MSLP	Principal Component Analysis	Kruizinga (1979)
PCAXTRC18	15–18	MSLP	Principal Component Analysis	Esteban et al. (2005)
TPCAC18	18	MSLP	Principal Component Analysis	Huth (1993)
ESLPC18	18	MSLP	LEADer algorithm	Erpicum et al. (2008)
KHC18	18	MSLP	LEADer algorithm	Kirchhofer (1973)
LUNDC18	18	MSLP	LEADer algorithm	Lund (1963)
CKMEANSC18	18	MSLP	OPTimization algorithm	Enke and Spekat (1997)
PCACAC18	18	MSLP	OPTimization algorithm	Yarnal (1993)
PCAXTRKMC18	15–18	MSLP	OPTimization algorithm	Esteban et al. (2006)
PETISCOC18	18	MSLP	OPTimization algorithm	Yarnal (1993)
SANDRAC18	18	MSLP	OPTimization algorithm	Philipp et al. (2007)
SANDRASC18	18	MSLP	OPTimization algorithm	Philipp (2008)

comparisons within groups reveals no significant differences. As MSLP is used for classification it is not surprising that best overall performance is reached for this variable (Fig. 3c). EV values for the associated surface climate variables 2mT and LSP, CP and PREC are considerably lower with CP featuring minimum values. Pairwise M–W tests yield significant differences between all categories. Conducting the analyses for the three subperiods 1957–1977, 1970–1990 and 1982–2002 (not shown) leads to very similar results pointing to the temporal robustness of the characteristics described above. In general evaluation results for CP, LSP and PREC exhibit highly significant correlations (not shown) pointing to the fact that classifications performing well with respect to one of the precipitation variables do so for the others as well. Thus from the three precipitation variables only PREC is further considered in the following.

4.3. Relationships between evaluation results for MSLP and surface climate variables

With respect to the use of CTCs within synoptic climatological analyses it is important to find out how far evaluation results for the classified variable (MSLP) are related to evaluation results for surface climate variables (2mT, PREC). Scatterplots in Fig. 4 illustrate these relationships between MSLP on the one hand and 2mT and PREC on the other hand based on evaluation results for the 16 selected CTCs for selected spatial domains and seasonal subsets. Performance for 2mT and PREC is expressed in terms of EV while EV and PCR have been chosen as evaluation metrics for MSLP in order to find out if there is one property of MSLP classification that is more closely connected to surface climate related classification performance. This may provide a hint on what property should preferably

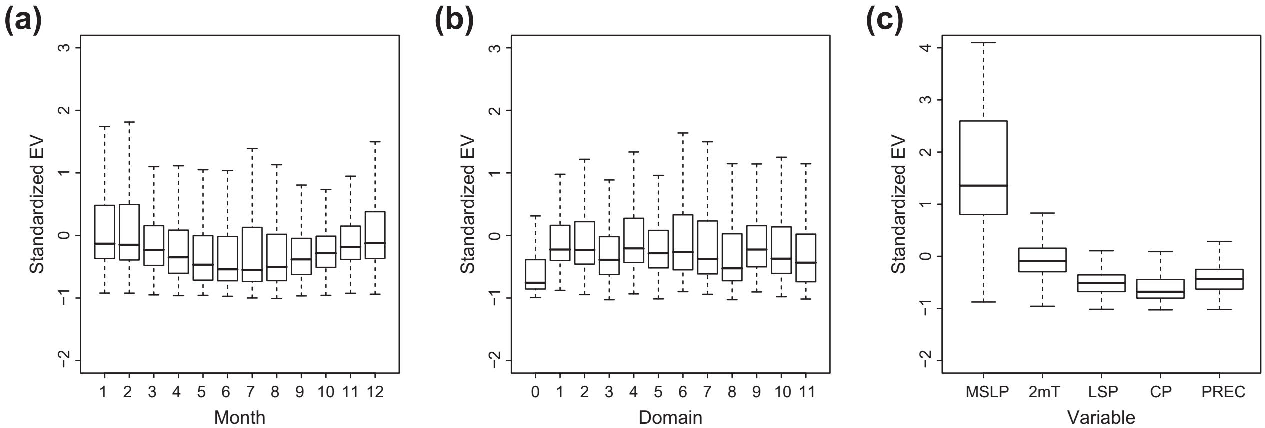

Fig. 3. Boxplots of standardized evaluation metrics (EV, determined for the period 1957–2002) for the 16 circulation type classifications from Table 1, grouped according to (a) months, (b) spatial domains, (c) evaluated variables. Upper/lower whiskers indicate the 1.5 interquartile range (IQR) from the upper/lower quartile. Note that EV has been standardized over all months, all domains or all variables respectively.

be optimized by automatic CTCs (e.g. by utilizing an adequate distance metric for assigning cases). From scatterplots for domain 0 based on all seasonal subsets (1st row in Fig. 4) a distinct positive correlation between MSLP performance (EV and as well PCR) and 2mT/PREC performance can be deduced (Spearman correlation coefficients and *p*-values are indicated in the bottom-left corner of each scatterplot). However such a clear cut relation between MSLP and 2mT/PREC cannot be seen for the smaller Central European domain 7 (2nd row in Fig. 4). Spearman correlation coefficients indicate significant correlations ($p < 0.05$) between MSLP-EV and 2mT and PREC however much lower than for the large domain 0. Related scatterplots reveal a marked bimodality of MSLP-EV values whereas scatterplots including PCR as evaluation metric for MSLP appear rather unstructured – resulting in Spearman correlation coefficients close to zero. Statistically significant correlations between MSLP-EV and 2mT and PREC detected above for domain 7 are solely due to corresponding distinct relationships during winter (see row 3 in Fig. 4). Non-significant correlations of opposite sign show up for the summer months (row 4 in Fig. 4). Respective scatterplots and Spearman correlation analyses for the other spatial domains (not shown here) reveal varying results, however some common characteristics can be summarized: For all smaller sub-domains (1–11) the relationship between classification performance for MSLP and surface climate variables is less pronounced than for the large North Atlantic-European domain 0, most probably due to the restricted consideration of important large-scale North Atlantic circulation structures associated with the advection of airmasses of varying thermal and hygric characteristics. Respective positive correlations (if present) are generally higher in winter than in summer (reflecting seasonal variations in importance of large-scale and small-scale processes for surface climate characteristics). Finally, in the majority of cases for which positive correlations are determined, correlation coefficients reach higher absolute values for the combination of EV for MSLP and EV for 2mT and PREC than when using PCR as evaluation metric for MSLP.

4.4. Ranking orders of CTCs and superordinate basic approaches

From results presented above it is clear that evaluation criteria estimated for MSLP do not necessarily reflect a CTC's performance concerning surface climate variables (2mT and PREC). This can also be deduced from Fig. 5 showing the mean ranks determined over all evaluation metrics for the selected CTCs and as well for superordinate groups of basic approaches integrated over all 12 domains and in addition for the same spatial and seasonal subsets as shown in Fig. 4. In addition to the mean ranks determined for the whole analysis period from 1957 to 2002 (grey shaded bars) mean ranks derived from evaluations for three temporal subsamples (1957–1977, 1970–1990, 1982–2002) are shown (transparent bars) in order to illustrate the temporal robustness of results.

Looking first at mean ranks determined for ensembles of CTCs combined with respect to the four basic approaches THRES, PCA, LEAD and OPT (right column barplots in Fig. 5) it is noteworthy that OPT achieves the highest mean rank (best performance) with respect to MSLP in all cases. THRES, PCA and LEAD exhibit varying ranking orders concerning performance for MSLP. While PCA is leading over THRES and LEAD in the large domain 0 a reversed ranking order appears for the Central European domain 7. For OPT and with one exception also for LEAD performance for 2mT and PREC is always lower than for MSLP whereas for THRES and PCA the opposite is true (almost always higher ranks for 2mT and PREC than for MSLP). Hence, although OPT seems to be most suitable for classifying MSLP in all cases OPT does not always perform best for the climatic target variables. This can be seen most clearly for the smaller domain 7, more distinctly for PREC than for 2mT and especially during summer. Another contrasting feature of winter and summer that refers to respective results shown in Fig. 4 is that differences between MSLP performance and 2mT/PREC performance tend to be larger during summer.

Ranks estimated for individual CTCs (left column barplots in Fig. 5) reveal differences in performance of classifications between seasons (e.g. for WLK) and spatial domains (e.g. for PCACA). Moreover partly marked differences among CTCs assigned to the same basic approach ensemble are evident.

LIT exhibits highest ranks from all members of THRES except for domain 0 and especially appears most suitable to capture 2mT while LWT2 reaches highest ranks for PREC. Remarkably WLK – the only method utilizing multiple variables for classification – features lowest ranks within this group of methods. From the group of PCA methods TPCA and PCAXTR exhibit comparable rank distributions with lower ranks for MSLP and higher ranks for 2mT and PREC, although both methods are based on different variants of principal component analysis (t-mode and s-mode respectively). In contrast the 3rd PCA method P27 features more balanced ranks for the three variables and generally reaches distinctly higher ranks for MSLP. Concerning classifications based on leader algorithms LEAD systematically lower ranks for KH compared to ESLP and LUND can be seen. Two members of OPT, CKMEANS and SANDRA – although relying on different variants of non-hierarchical cluster analysis (Philipp et al., 2010) – feature

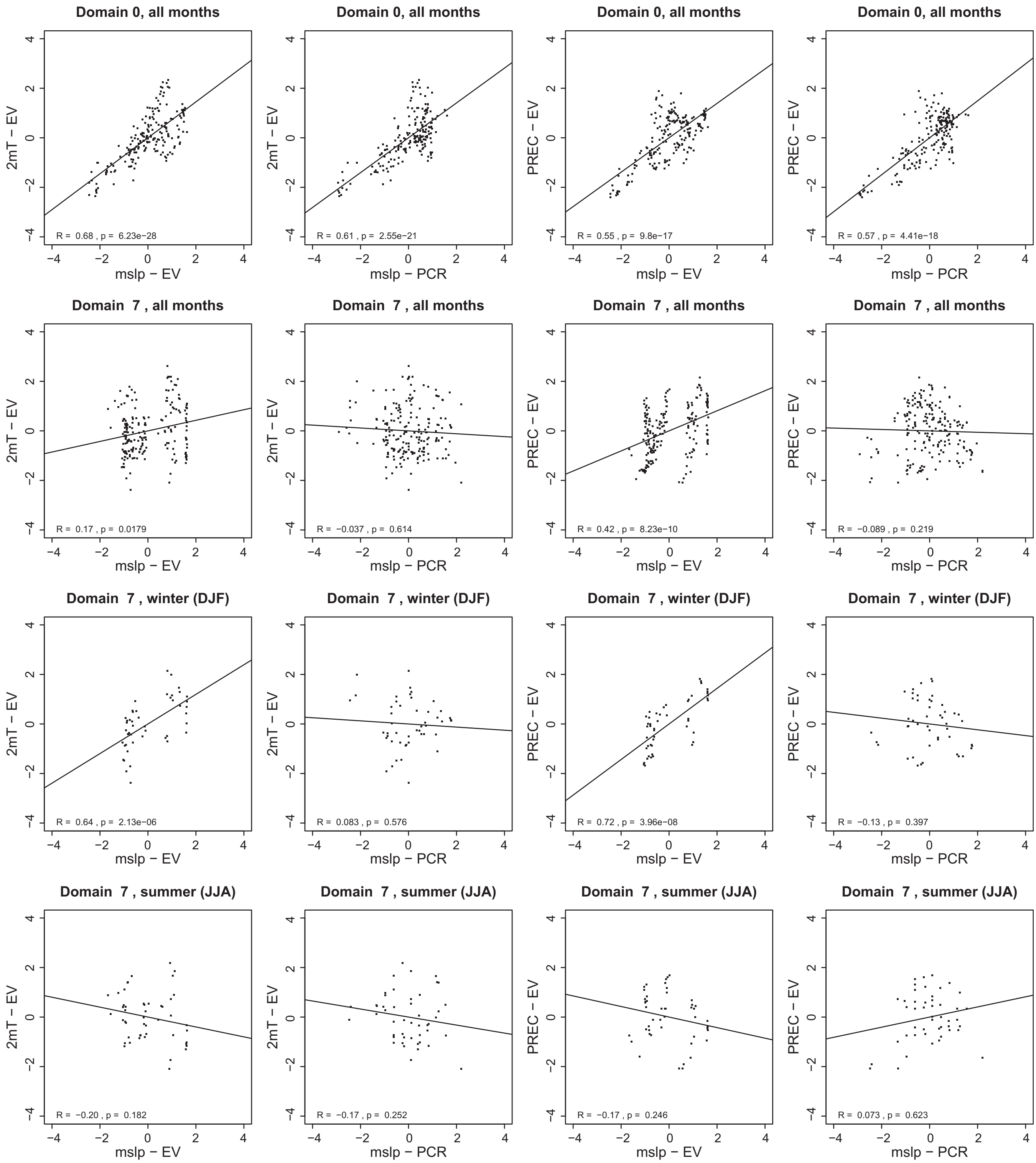

Fig. 4. Scatterplots illustrating relationships between performance of circulation type classifications for MSLP (evaluation criteria EV and PCR respectively) and 2mT and PREC (evaluation criterion EV), determined for the period 1957–2002. 1st row: for large domain 0, including all months. 2nd row: for Central European domain 7, including all months. 3rd row: for domain 7, winter months (DJF). 4th row: for domain 7, summer months (JJA). Respective Spearman correlation coefficients and *p*-values are indicated in the bottom left of each individual scatterplot.

almost identical ranks and reach overall best performance with respect to MSLP but distinctly lower ranks for 2mT and PREC. Whereas PCACA exhibits lower ranks for MSLP but mostly higher ranks for 2mT and PREC (except for domain 0). For PCAXTRKM the general characteristic of OPT – highest ranks for MSLP and lower ranks for 2mT and PREC – is reversed. SANDRAS – a successor of SANDRA – shows generally lower ranks for MSLP and PREC than SANDRA but achieves remarkably high ranks for 2mT especially during winter, most probably due to the utilization of sequences of days for classification.

Ranking orders determined on the basis of temporal subsamples (transparent bars in Fig. 8) generally indicate a high degree

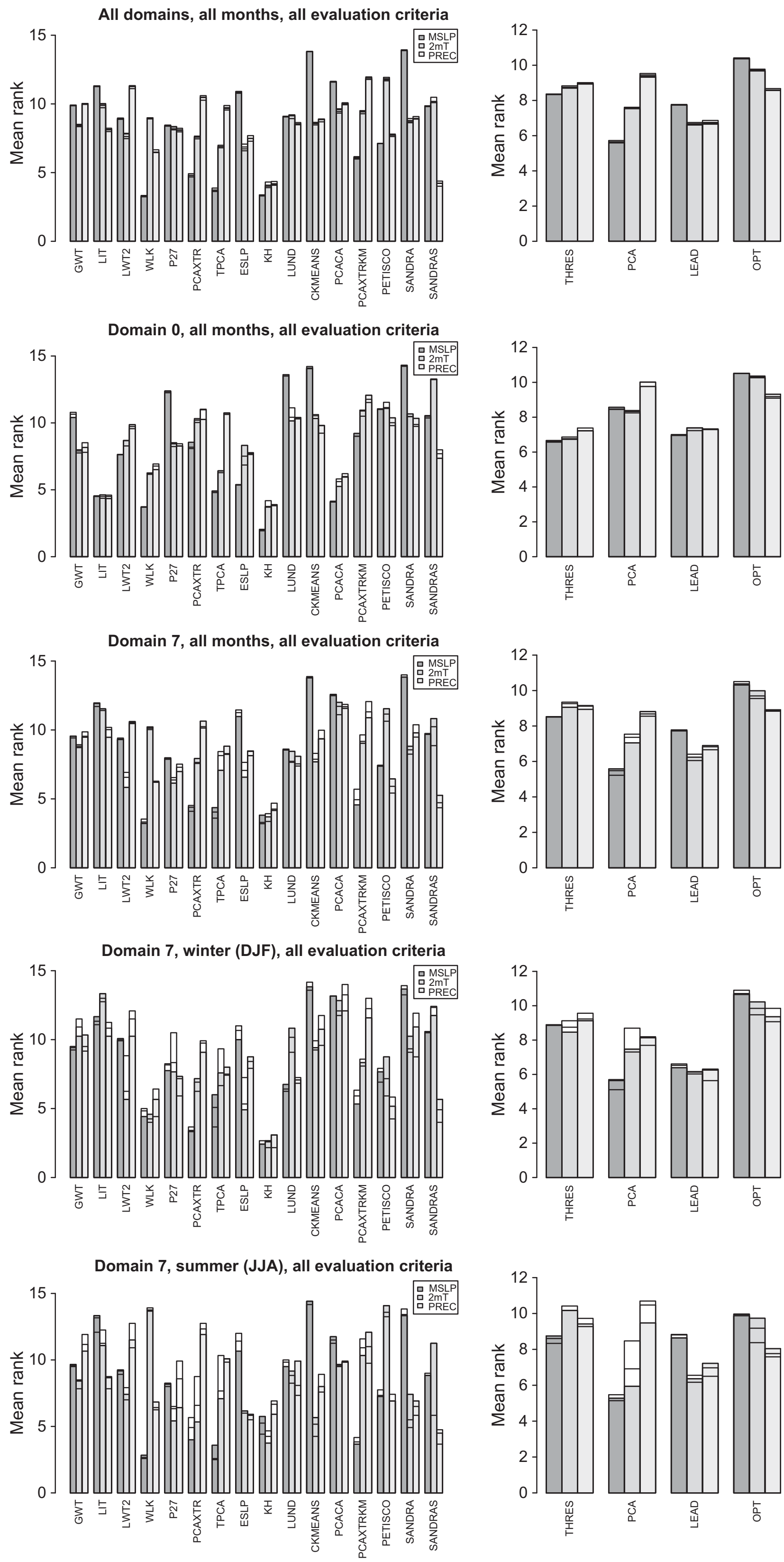

Fig. 5. Performance rankings of circulation type classifications from Table 1 (left column of barplots) and groups of basic classification methods (right column of barplots) respectively. Ranks are estimated as mean ranks, averaged over all evaluation criteria. 1st row: over all spatial domains, including all months. 2nd row: for large domain 0, including all months. 3rd row: for Central European domain 7, including all months. 4th row: for domain 7, winter months (DJF). 5th row: for domain 7, summer months (JJA). Grey shaded bars represent mean ranks determined for the whole analysis period (1957–2002), superposed transparent bars indicate mean ranks derived from evaluations for three temporal subsamples (1957–1977, 1970–1990, 1982–2002).

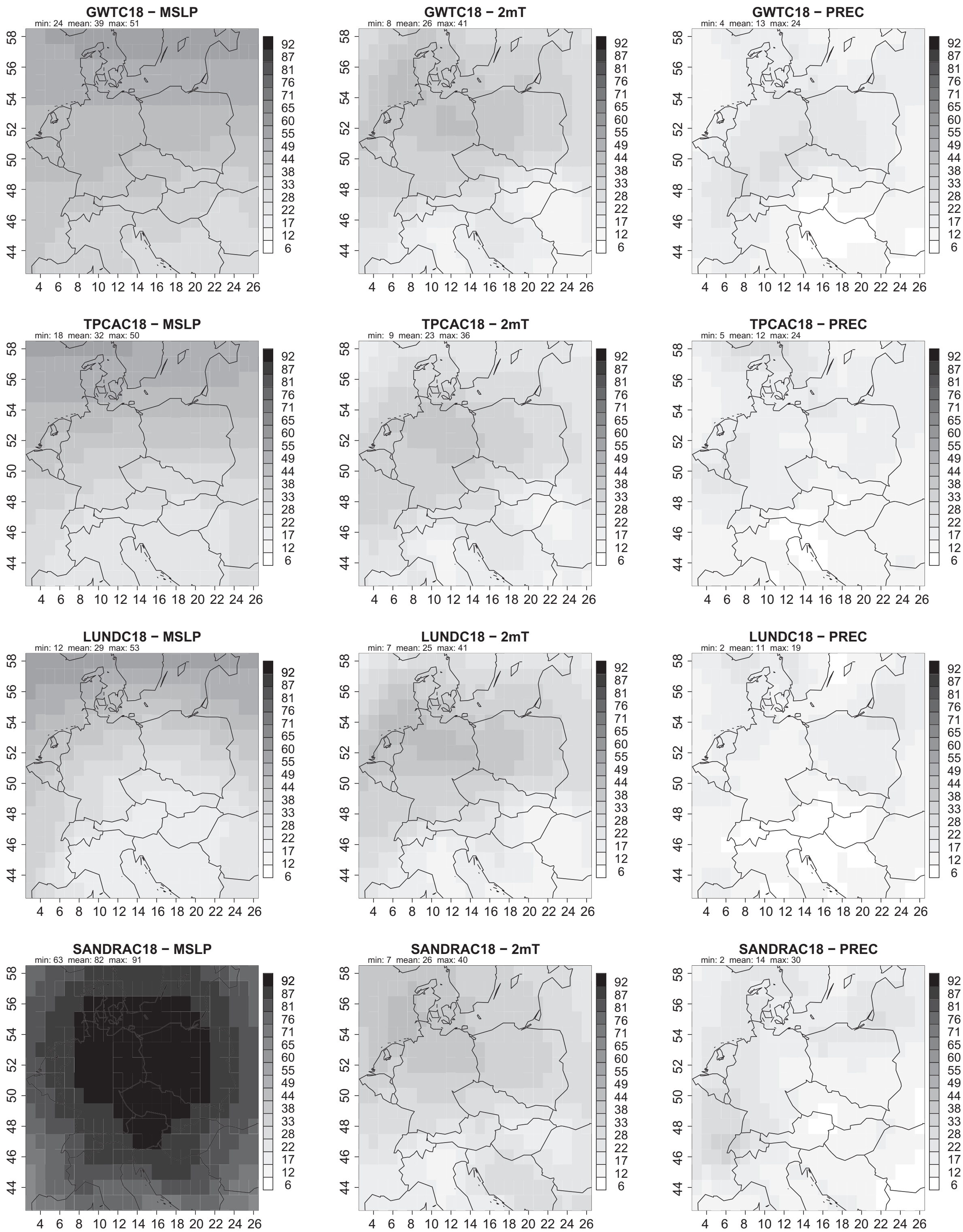

Fig. 6. Spatial distribution of EV (in%) for winter (DJF, 1957–2002) over spatial domain 7 for MSLP, 2mT and PREC respectively. For selected circulation type classifications from Table 1. Minimum, mean and maximum of EV (in%) are indicated on the top left of each map.

of temporal robustness of results. More distinct differences between rankings for different periods appear mainly for seasonal analyses and for PCA based and optimization methods (OPT).

Rankings for the other domains (not shown here) show partly varying results. However, although different domains feature modified ranking orders the main characteristics of the basic classification approaches and as well the individual CTCs as described above appear to remain more or less stable over all spatial domains.

4.5. Spatial variations in performance of CTCs

So far only evaluation results integrated over whole spatial domains have been presented. However applications of CTCs

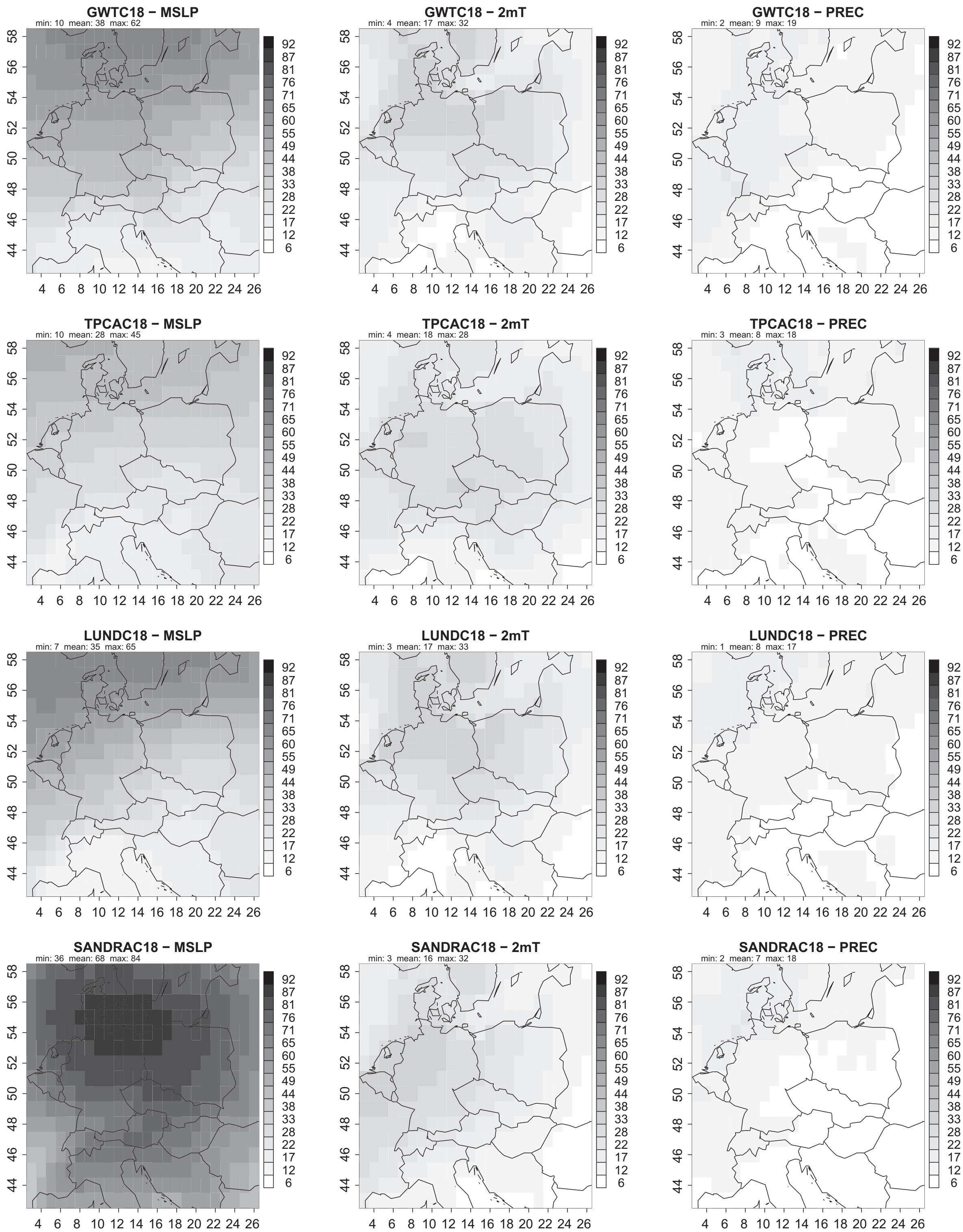

Fig. 7. Same as Fig. 6 but for summer (JJA, 1957–2002).

within synoptic climatological analyses mostly utilize regional or even local surface climate data as target variables. Thus it is appropriate to address the spatial variations of evaluation metrics within individual spatial domains. For this purpose four CTCs – each representing one basic classification approach – have been selected and respective fields of EV for MSLP, 2mT and PREC over the Central European domain 7 are presented for winter (Fig. 6) and summer (Fig. 7) respectively. Minimum, mean and maximum of EV (in%) are indicated on the top left of each map.

For MSLP in both seasons and for all CTCs highest EV values can be observed in the northern part of the domain with slightly higher maximum values for GWT and LUND compared to TPCA. However, in agreement with results from Section 4.4 the cluster based SANDRA method (and as well CKMEANS; not shown) does not only reach by far the highest EV values (maximum value of 92%) but

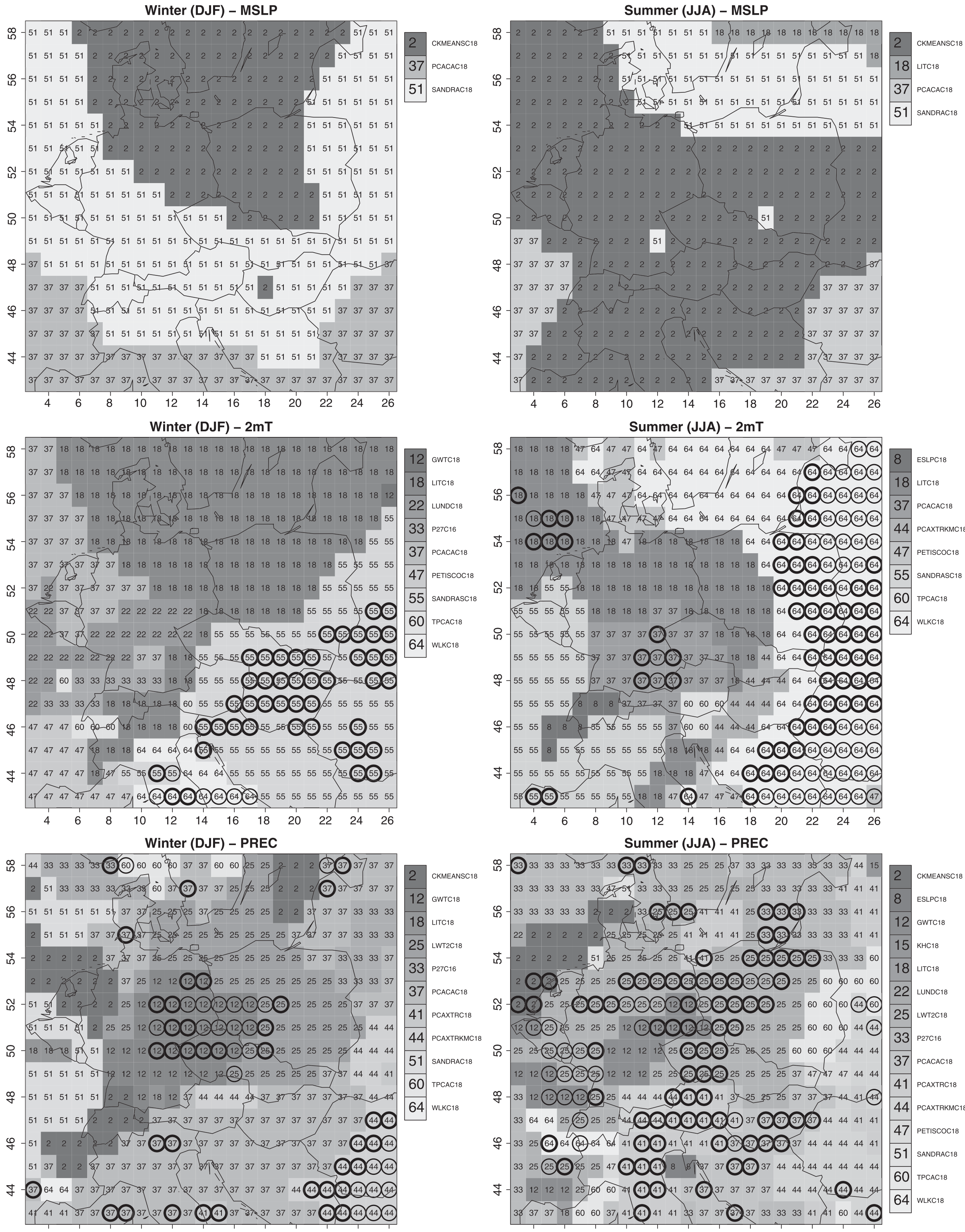

Fig. 8. CTCs with highest performance (in terms of EV) at individual grid points in spatial domain 7. For MSLP, 2mT and PREC, in winter (DJF, 1957–2002) and summer (JJA, 1957–2002). Circles indicate gridpoints for which EV of the leading CTC is higher than 1.5 times the interquartile range (IQR) above the third quartile (thin outline circle) and the EV of the following CTC is lower than 1.0 times the IQR above the third quartile (bold outline circle).

in addition features an extension of the region with distinctly high EV values over the whole domain. However this superior performance of SANDRA for MSLP does not equally apply to the surface climate variables 2mT and PREC.

EV patterns for 2mT and PREC reveal on the one hand characteristic spatial structures that apply to all four selected CTCs and on the other hand some interesting differences between CTCs. In all cases highest performance in terms of maximum values of EV is reached in winter and – concerning their spatial distribution – in western and northwestern parts of the domain. These general seasonal and spatial variations are not solely a result of performance characteristics of individual CTCs but also reflect general climate characteristics of the Central European domain. Western and northwestern regions that are more exposed to more frequently occurring western and northwestern circulation types feature a more distinct relationship between surface climate variables and large-scale atmospheric circulation dynamics. Whereas surface climate in winter is to a large extent controlled by macro-scale circulation, in summer – and especially with respect to precipitation and for the more continental parts of Europe – small-scale processes (e.g. local convection) are responsible for large fractions of temperature and precipitation variability on the synoptic time-scale. Focusing on EV patterns for 2mT in winter all four selected CTCs show rather similar spatial patterns with GWT exhibiting the largest extension of EV values over approximately 30% and reaching (together with LUND) highest maximum EV values. In summer more distinct differences between the four CTCs are evident. While LUND and GWT show best performance in the northern/northwestern parts of the domain TPCA – featuring lowest maximum EV values – performs best in the central regions and finally SANDRA exhibits a clear performance maximum in the western/northwestern Central European regions.

Spatial performance patterns for different CTCs show even more distinct differences for PREC, the variable that in general features lowest overall EV values. GWT and SANDRA, featuring highest maximum EV values in winter and summer respectively, exhibit largest extensions of regions with EV over 17% in winter and over 12% in summer respectively. Whereas LUND and TPCA reach comparably high performance only in smaller fragmented regions. However core regions of relatively high EV differ between CTCs and between seasons. Western and northwestern parts of Central Europe covered quite well by SANDRA in winter and GWT in summer whereas highest performance is shifted to the central parts of the domain for GWT in winter and is restricted to the Northwest for SANDRA in summer and also for TPCA in both seasons and for LUND in summer.

Hence in contrast to results for MSLP, spatial patterns of EV for 2mT and PREC do not clearly indicate a generally superior performance of one CTC but depict rather small differences between the four selected CTCs concerning 2mT and in addition point to some specific "core regions" of higher performance related to individual CTCs especially for PREC.

As it is impossible to show respective maps for all CTCs from Table 1 Fig. 8 indicates for each grid point in domain 7 and for the three variables MSLP, 2mT and PREC the one CTC that performs best in terms of EV at the respective location. However the best performing CTC is not necessarily clearly distinguished from the mean performance of the CTC ensemble or maybe is only slightly better than the second best classification. Thus for each gridpoint it is checked if (i) the EV value of the leading CTC is higher than 1.5 times the interquartile range (IQR) above the third quartile (thin outline circles in Fig. 8) and (ii) if at the same time the EV value of the second best CTC is lower than 1.0 times the IQR above the third quartile (bold outline circles in Fig. 8). Hence it is possible to indicate those grid points for which one CTC (i) features distinctly better performance than the median of all CTCs and (ii) the leading CTC is clearly separated from the second best CTC.

A main result that can be deduced from Fig. 8 is that only a few CTCs are selected as the "best" methods with respect to MSLP. Increasing numbers of leading CTCs appear on respective maps for 2mT and PREC leading to more patchy spatial patterns for these two variables compared to the clearly defined spatial distributions for MSLP. Mainly three members from the OPT group of basic approaches reach highest performance for MSLP in winter and as well in summer. However for none of the grid points the respective leading CTC reaches EV values distinctly different from the median. With respect to 2mT LIT, SANDRAS and PCACA appear in both seasons as superior CTCs for remarkable fractions of grid points, however covering distinctly different regions in winter and in summer and – with the exception of SANDRAS during winter in the south-eastern part of the domain – not featuring clearly outstanding EV values. WLKC shows distinctly superior performance for large parts of the eastern part of the domain only in summer pointing to the marked seasonal differences in performance of this CTC (see also Fig. 5). Concerning PREC it is noteworthy that two CTCs from the THRES group of methods (LWT2, GWTC) feature highest and clearly separated EV values in the central parts of the domain. This is most probably due to the inclusion of a vorticity estimate (derived from MSLP) in both CTCs. Again these results show clearly that superior performance of a specific CTC or a group of methods for MSLP is not necessarily associated to a leading position concerning performance for 2mT and/or PREC. Moreover spatial patterns of superior performance concerning one variable do also differ distinctly between seasons. Performing the same analyses for three subperiods (1957–1977, 1970–1990 and 1982–2002; not shown) leads to very similar results; especially with regard to the regions exhibiting distinctly superior performance of individual CTCs. Respective maps for other spatial domains (not shown) confirm the leading performance of OPT methods for MSLP in all domains and all seasons, whereas spatial performance patterns for 2mT and PREC reveal respective substantial seasonal and spatial variations.

5. Discussion and conclusions

This study presented investigations on the evaluation and comparison of CTCs that are currently carried out within the framework of the COST733 Action "Harmonisation and Applications of Weather Types Classifications for European Regions". Analyses have been performed for a comprehensive data set of daily CTCs for the larger European region and several smaller European subdomains, covering the period from 1957 to 2002 and comprising variants of representatives of all commonly used basic classification methods (Philipp et al., 2010). The performance of CTCs in terms of the separability and within-type variability of circulation types as basic properties of CTCs has been estimated by calculating several statistical metrics for the variable used for classification (MSLP) and as well for associated surface climate variables (2mT, PREC) using ERA-40 reanalysis data (Uppala et al., 2005) as one consistent data set for all analyses.

Earlier comparison studies (Huth, 1996) included several other properties for evaluating the performance of CTCs. However evaluation criteria like "stability in space and time" (sensitivity of CTCs to variations in spatial resolution or sample size) and "consistency" (sensitivity to varying numbers of types) are beyond the scope of this study. Other desirable features (according to Huth, 1996) like equally sized classes and the ability to reproduce predefined types (estimated from subjective classifications) are, although easy to determine, hard to assess as both features implicitly assume the a priori knowledge of the respective natural characteristics.

Compared to these earlier analyses the present study expands the comparative analyses of CTCs by (i) including a wider variety of classification methods (e.g. the whole group of threshold based methods), (ii) performing evaluations for varying spatial (spatial domains) and (iii) temporal (monthly to annual) subsets and (iv) evaluating the separability and within-type variability of CTCs not only with respect to circulation data (MSLP) but as well for climatic target variables (TEMP, PREC).

From evaluation analyses applied to the whole set of available CTCs it became clear that evaluation criteria exhibit distinct sensitivity to the number of circulation types. Thus the analyses presented in this paper focus on a selection of 16 different automatic CTCs, each of them comprising 18 circulation types, in order to ensure that results of evaluations and comparisons reflect differences between individual CTCs and basic classification methods and not differences due to varying numbers of circulation types. Evaluation results averaged over the whole ensemble of selected CTCs indicate generally higher performance of CTCs for winter months, for the smaller and more westerly spatial domains, and for MSLP compared to 2mT and PREC. Whereas higher performance for MSLP is evidently due to the fact that only this variable is used for classification by all CTCs (with the exception of one CTC), systematic differences in general performance levels between seasons and spatial domains can be explained by respective spatiotemporal variations in the intensity of the links between large-scale circulation and surface climate. Beside these general variations in evaluation results that appear with varying intensity for all CTCs the comparison of evaluation indices reveals distinct differences in performance among individual CTCs and between superordinate groups of basic methods (THRES, PCA, LEAD, OPT) as well. However, ascribing such performance variations to specific methodological features of individual CTCs or superordinate approaches is possible only to some extent (e.g. better performance of threshold based CTCs with respect to precipitation is most probably due to the consideration of simple vorticity estimates).

In this context it's worth to mention that the arrangement of individual CTCs into groups, defined with respect to the common use of basic classification methods (e.g. principal component analysis) is not in all respects convincing. Evaluation results have shown distinct variations in performance characteristics among the members of some of these groups of basic methods. Partly this may be due to questionable assignments of CTCs to a specific group of methods. For example, P27 (Kruizinga, 1979) is assigned to the PCA group because it utilizes S-mode PCA during initial steps of the classification. However taking into account the conceptual layout P27 is maybe more closely related to some of the CTCs that are combined in the THRES group of methods. Accordingly evaluation results for P27 are more similar to some members of the THRES groups of methods than to any member from the PCA group. Hence, findings concerning the performance of superordinate groups of basic methods may be disturbed by the fact that individual methods cannot always be assigned unambiguously to one superordinate group of methods.

Nevertheless, one main finding from this study is that CTCs utilizing optimization algorithms (in fact variants of non-hierarchical cluster analysis) for classification reach highest performance for MSLP in most cases. However, it is shown that superior performance concerning MSLP is not necessarily related to comparable high performance levels for associated surface climate variables. In many cases CTCs reaching lower performance for MSLP exhibit superior skill for surface climate variables. With respect to the aspired development of optimized CTCs featuring superior performance for varying environmental target variables this implies that it is most probably not feasible to create such an overall best CTC. Instead, results from the presented analyses can provide recommendations for the use of existing CTCs for specific applications and the development of "custom-designed" CTCs. At this stage of our evaluation studies it is possible to provide the following recommendations. CTCs using optimization algorithms (OPT) should preferably be used for analyses that strongly rely on the availability of well defined circulation patterns (e.g. long-term frequency variations of circulation types) as these classification methods provide circulation types with highest/lowest separability/within-type variability. For relating circulation types to surface climate variables (e.g. for statistical downscaling purposes) the use of PCA based or threshold based CTCs – in some cases – appears to be more appropriate. In order to provide meaningful suggestions for the development of new or the improvement of existing CTCs future work on the evaluation and comparison of CTCs should strongly focus on the attribution of detected performance characteristics to specific methodological properties of CTCs. Based on the presented evaluation studies we recommend that methodological features of the THRES methods (inclusion of air flow direction and/or vorticity estimates derived from MSLP data) apparently leading to a better discrimination of surface climate variables should be taken into consideration in the development of improved CTCs. Moreover the inclusion of additional variables (e.g. atmospheric layer thickness, wind components from different levels) – that is intended for future steps in method development within the framework of COST733 – may further increase the performance of classifications. However, according to results presented in this study, such "custom-designed" CTCs should be developed not only with regard to specific target variables but should also take into consideration the envisaged seasonal and regional priorities.

Acknowledgement

The work presented in this paper was performed in the context of the COST733 Action "Harmonisation and Applications of Weather Types Classifications for European Regions" that is funded by the EU within the framework of COST (European Cooperation in Science and Technology).

References

Beck, C., Jacobeit, J., Jones, P.D., 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.

Bosilovich, M.G., Chen, J., Robertson, F.R., Adler, R.F., 2008. Evaluation of global precipitation in reanalyses. J. Appl. Meteorol. Climatol. 47, 2279–2299.

Brinkmann, W.A.R., 1999. Within-type variability of 700 hPa winter circulation patterns over the Lake Superior Basin. Int. J. Climatol. 19, 41–58.

Buishand, T.A., Brandsma, T., 1997. Comparison of circulation classification schemes for predicting temperature and precipitation in the Netherlands. Int. J. Climatol. 17, 875–889.

Calinski, T., Harabsz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–2.

Crochet, P., 2007. A study of regional precipitation trends in Iceland using a high-quality gauge network and ERA-40. J. Clim. 20, 4659–4677.

Demuzere, M., Trigo, R.M., Vila-Guerau de Arellano, J., Van Lipzig, N.P.M., 2008. The impact of weather and atmospheric circulation on O_3 and PM10 levels at a mid-latitude site. Atmos. Chem. Phys. Discuss. 8, 21037–21088.

Dittmann, E., Barth, S., Lang, J., Müller-Westermeier, G., 1995. Objektive Wetterlagenklassifikation. Ber. Dt. Wetterd., 197, Offenbach a. M., Germany.

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: Book of Abstracts COST 733 Mid-term Conference, Advances in Weather and Circulation Type Classifications and Applications, 22–25 October 2008 Krakow, Poland, p. 33.

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

Esteban, P., Martin-Vide, J., Mases, M., 2006. Daily atmospheric circulation catalogue for Western Europe using multivariate techniques. Int. J. Climatol. 26, 1501–1515.

Hess, P., Brezowski, H., 1952. Katalog der Großwetterlagen Europas. Ber. Dt. Wetterd. in der US-Zone 33, Bad Kissingen, Germany.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 63–76.

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., Untstrul, Z., Cahynova, M., Kysely, J., Tveito, O.E., 2008. Classifications of atmospheric circulation patterns: recent advances and applications. Ann. NY Acad. Sci. 1146, 105–152.

Jacobeit, J., Wanner, H., Luterbacher, J., Beck, C., Philipp, A., Sturm, K., 2003. Atmospheric circulation variability in the North-Atlantic-European Area since the mid-seventeenth century. Clim. Dynam. 20, 341–352.

James, P.M., 2006. Second generation lamb weather types – a new generic classification with evenly tempered type frequencies. In: 6th Annual Meeting of the EMS/6th ECAC, EMS2006A00441, Ljubljana, Slovenia.

James, P.M., 2007. An objective classification method for Hess and Brezowsky Grosswetterlagen over Europe. Theor. Appl. Climatol. 88, 17–42.

Kalkstein, L.S., Tan, G., Skindlov, J.A., 1987. An evaluation of three clustering procedures for use in synoptic climatological classification. J. Appl. Meteorol. 26, 717–730.

Kirchhofer, W., 1973. Classification of European 500 mb patterns. Arbeitsbericht der Schweizerischen Meteorologischen Zentralanstalt, 43, Geneva, Switzerland.
Kruizinga, S., 1979. Objective classification of daily 500 mbar patterns. In: Preprints Sixth Conference on Probability and Statistics in Atmospheric Sciences, 9–12 October 1979 Banff, Alberta, American Meteorological Society, Boston, MA, pp. 126–129.
Kysely, J., Huth, R., 2006. Changes in atmospheric circulation over Europe detected by objective and subjective methods. Theor. Appl. Climatol. 85, 19–36.
Litynski, J., 1969. A numerical classification of circulation patterns and weather types in Poland. PIHM 97, 3–15 (in Polish).
Lund, I.A., 1963. Map-pattern classification by statistical methods. J. Appl. Meteorol. 2, 56–65.
Martin, E., 2004. Validation of Alpine Snow in ERA-40. ERA-40 Project Report Series 14. European Centre for Medium-Range Weather Forecasts, Shinfield, Reading, UK.
Milligan, G., Cooper, M., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179.
Milligan, G., Cooper, M., 1986. A study of the comparability of external criteria for hierarchical cluster analysis. Multivar. Behav. Res. 21, 441–458.
Omstedt, A., Chen, Y., Wesslander, K., 2005. A comparison between the ERA40 and the SMHI gridded meteorological databases as applied to Baltic Sea modelling. Nord. Hydrol. 36, 369–380.
Philipp, A., Della-Marta, P.M., Jacobeit, J., Fereday, D.R., Jones, P.D., 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., 2008. Comparison of principal component and cluster analysis for classifying circulation pattern sequences for the European domain. Theor. Appl. Climatol.. doi:10.1007/s00704-008-0037-1.
Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Huth, R., James, P., Jourdain, S., Krennert, T., Lykoudis, S., Michaelides, S., Pianko, K., Post, P., Rasilla Álvarez, D., Spekat, A., Tymvios, F.S., 2010. COST733CAT – a database of weather and circulation type classifications. Phys. Chem. Earth 35, 360–373.
Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66, 846–850.
Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Schmutz, C., Wanner, H., 1998. Low frequency variability of atmospheric circulation over Europe between 1785 and 1994. Erdkunde 52, 81–94.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. 623–656.
Stehlik, J., Bardossy, A., 2003. Statistical comparison of European circulation patterns and development of a continental scale classification. Theor. Appl. Climatol. 76, 31–46.
Strehl, A., Gosh, J., 2002. Cluster ensembles – a knowledge reuse framework for combining partitions. J. Mach. Learn. Res. 3, 583–617.
Uppala, S.M., Kallberg, P.W., Simmons, A.J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., Mcnally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J., 2005. The ERA-40 re-analysis. Q. J. R. Meteorol. Soc. 131, 2961–3012.
Yarnal, B., 1993. Synoptic Climatology in Environmental Analysis. Belhaven Press, London, UK. 195 pp.