Synoptic-climatological evaluation of the classifications of atmospheric circulation patterns over Europe

Radan Huth,[a,b,c]* Christoph Beck[d] and Monika Kučerová[b]

[a] *Department of Physical Geography and Geoecology, Faculty of Science, Charles University, Prague, Czech Republic*
[b] *Institute of Atmospheric Physics, Czech Academy of Sciences, Prague, Czech Republic*
[c] *Global Change Research Centre, Czech Academy of Sciences, Brno, Czech Republic*
[d] *Department of Geography, University of Augsburg, Germany*

ABSTRACT: This study evaluates the classifications of atmospheric circulation patterns collected in the COST733 database (COST733cat) in terms of their ability to stratify daily surface temperature and precipitation in 12 domains covering the whole of Europe. The classifications differ in the classification methods used, in the number of types, the variable(s) classified, the number of days in a sequence that are classified and in whether the classification is based on year-round or seasonal data. Several classification methods that perform fairly well are identified; they include a simple k-means clustering, a k-means clustering preceded by hierarchical cluster analysis, Litynski's method, and a classification based on circulation prototypes. On the other hand, there are a couple of classification methods that do not provide a good stratification of temperature and precipitation: orthogonally and obliquely rotated principal component analysis in a T-mode, Lund's correlation method, Kirchhofer's sums-of-squares method, and Erpicum's method. Some methods tend to perform better on large domains, while others tend to perform better on smaller domains; however, the sensitivity of most classification methods to the domain size appears to be small. Several methods exhibit a geographical dependence of their performance, e.g. the method based on circulation prototypes tends to perform better in the northern domains, while Jenkinson–Collison and Erpicum's methods perform better in the southern domains. Classifications of 4-day sequences are usually better in stratifying surface temperature than ordinary instantaneous classifications; the opposite is true for precipitation. Adding a mid-tropospheric variable (500 hPa heights or 1000/500 hPa thickness) to sea level pressure as a classified variable improves the skill of classifications in stratifying temperature.

1. Introduction

1.1. General issues

Classifications of atmospheric circulation patterns (also referred to as synoptic classifications) are useful tools for handling an immense and boundless continuum of individual instantaneous circulation patterns. Classifications simplify the physical reality by identifying a small number of representative patterns (types) to which the instantaneous patterns are assigned (Huth *et al.*, 2008). One of the goals of synoptic classifications is to aid in the description of effects the atmospheric circulation has on surface climate, which is considered the main task of synoptic climatology (e.g. Yarnal, 1993; Barry and Carleton, 2001; Yarnal *et al.*, 2001). Therefore, it is important to evaluate if, and to what extent, the classifications are able to describe surface weather, climate and environmental variables. Numerous studies on the relationships between circulation types and surface climate and environment have been published, from those published several decades ago (e.g. Paegle, 1974; Yarnal, 1985; Yarnal *et al.*, 1988; O'Hare and Sweeney, 1993) to recent ones, which are further referenced in this study. The range of climate and environmental variables, for which the classifications are evaluated, is quite wide although most studies deal with temperature and precipitation (e.g. Kostopoulou and Jones, 2007; Nishiyama *et al.*, 2007; Lorenzo *et al.*, 2008; Jones and Lister, 2009; Schuenemann *et al.*, 2009; Casado *et al.*, 2010; Tveito, 2010; Brisson *et al.*, 2011; Küttel *et al.*, 2011; Raziei *et al.*, 2012 to name just a few recent examples) including extremes (Cassano *et al.*, 2006; Jacobeit *et al.*, 2009; Maraun *et al.*, 2011) and diurnal cycles (Twardosz, 2010). Many other climatic and environmental characteristics have, nevertheless, been investigated as well: wind and wind storms (Cassano *et al.*, 2006; Leckebusch *et al.*, 2008; Donat *et al.*, 2010), droughts (Vicente-Serrano and López-Moreno, 2006; Fleig *et al.*, 2010), floods (Dayan *et al.*, 2012), snowfall and avalanches (Esteban *et al.*, 2005; Bednorz, 2008; García *et al.*, 2009), lightning activity (Pineda *et al.*, 2010), atmospheric pollutants and dust transport (Demuzere *et al.*, 2009, 2011; Lykoudis *et al.*, 2010; Stefan *et al.*,

* Correspondence to: R. Huth, Department of Physical Geography and Geoecology, Faculty of Science, Charles University, Albertov 6, 128 43 Praha 2, Prague, Czech Republic. E-mail: huth@ufa.cas.cz

2010; Dayan *et al.*, 2012; Gaetani *et al.*, 2016; Beck *et al.*, 2014), forest fires (Kassomenos, 2010; Rasilla *et al.*, 2010), storm surges (Ullmann and Monbaliu, 2010; Rasilla and García-Codrón, 2016), potato yields (Sepp and Saue, 2012) and many others.

The strength of the link between the circulation types and surface environment can be quantified by a variety of criteria and statistical approaches. They include testing for the equality of the mean values between individual types (Kostopoulou and Jones, 2007), correlations between seasonal frequencies of types and seasonal means of climate variables (Anagnostopoulou *et al.*, 2008) and approaches quantifying and testing the (dis)similarity of statistical distributions conditioned by circulation types, such as the analysis of variance (Jiang *et al.*, 2005), Brier skill score (Schiemann and Frei, 2010), Kruskal–Wallis test (Fernau and Samson, 1990), χ^2-test (Makra *et al.*, 2006), the Mann–Whitney test (Fernau and Samson, 1990), and the Kolmogorov–Smirnov test (Huth, 2010; Tveito, 2010). Several statistical criteria have been designed directly to quantify the separability between individual types; some of them are listed in Beck and Philipp (2010).

It is necessary to distinguish between the terms 'circulation pattern' and 'circulation type', which are sometimes confused. A circulation pattern describes the atmospheric circulation at any given moment in time (e.g. at a time of a synoptic measurement, a daily mean, a monthly mean); it is essentially a synoptic map. On the other hand, a circulation type is a product of classification; it is a group (cluster) of individual instantaneous patterns, and it is also a pattern (map) that represents all the individual patterns (maps) classified with this type (usually as their mean, median or centroid).

In this article, we also make a clear distinction between the terms 'classification' and 'classification method'. While a classification method is a mathematical (or another) algorithm to identify types and to assign individual circulation patterns to the types, a classification is the resulting grouping of individual patterns into types. As such, it can be seen as a specific realization of a classification method for a particular data set and a particular setting of the parameters of the method. Therefore, a single classification method may produce many individual classifications, depending on the input data used and method's parameters chosen.

It is important to know how the choice of the classification method and its parameters affects the ability of the resulting classification to stratify (sort) the surface climate variable of interest, such as temperature and precipitation. No such systematic study has been performed; the few studies known to the authors have only examined single aspects of classifications. For example, Kidson (1997) found that the choice of pressure level affects the stratification of surface climate elements in New Zealand only weakly and that classifications based on multiple levels do not outperform those based on a single level. Beck and Philipp (2010), Huth (2010), Schiemann and Frei (2010), and Tveito (2010) examined the effect of the method choice and of the number of types on various aspects of stratification of the surface temperature and precipitation for a previous version of the COST733 classification database (COST733cat v1.2).

1.2. Context within the COST733 Action

This study is a part of a special issue devoted to the results of the international research activity conducted under the umbrella of the COST (European Cooperation in Science and Technology) Action 733 'Harmonisation and Applications of Weather Types Classifications for European Regions'. More information on the COST733 Action can be found in the review paper by Huth *et al.* (2008) and the preface to the special issue in *Physics and Chemistry of the Earth* by Huth *et al.* (2010). Central to this activity, and described elsewhere in this special issue (Philipp *et al.*, 2016) as well as previous studies (Philipp *et al.*, 2010), is a database of a large number of circulation (synoptic) classifications for 12 domains covering Europe. The classifications have been created by 18 different classification methods, following a unified methodology, and are based on a unified data set. This provides a unique opportunity for intercomparisons of various classification methods as well as of various settings of the methods. This study concentrates on one particular aspect of synoptic classifications from the COST733 database (referred to as 'COST733cat' below) that is essential in synoptic climatological studies: the ability of classifications to stratify the values of surface climate variables, namely temperature and precipitation. The approach adopted in COST733 was that classifications have not been tailored to specific applications; rather, their utility in different applications is being evaluated a posteriori. For examples of such evaluations, an interested reader may consult other papers in this special issue or Huth *et al.* (2010). Specifically, the utility of classifications from COST733 in various environmental applications was examined and compared by Fleig *et al.* (2010), Kassomenos (2010), Lykoudis *et al.* (2010), Stefan *et al.* (2010), and Trigo *et al.* (2016).

This study evaluates classifications from version 2.0 of the COST733cat database (Philipp *et al.*, 2016). The individual classifications differ from each other in the methods used, the variable(s) used for classification, the number of types, the seasonality of definition (i.e. whether the types are defined for the whole year or separately for individual seasons), and sequencing (i.e. how many sequential daily patterns are used as input for the classification). The database therefore provides an outstanding opportunity to study the consequences of the choice of classification methods and their settings on the properties of classifications and, in the case of this study, on stratification of surface climate elements. Thanks to its geographical extent, the high number of methods included and a range of settings covered, results found are easy to generalize.

We note that the COST733cat database does not allow us to fully assess one of the important aspects of classifications, viz. the size of the domain over which the circulation patterns are classified, mainly because the domains defined in COST733 have similar sizes and have little, if any, spatial overlaps. It is therefore covered only marginally in this

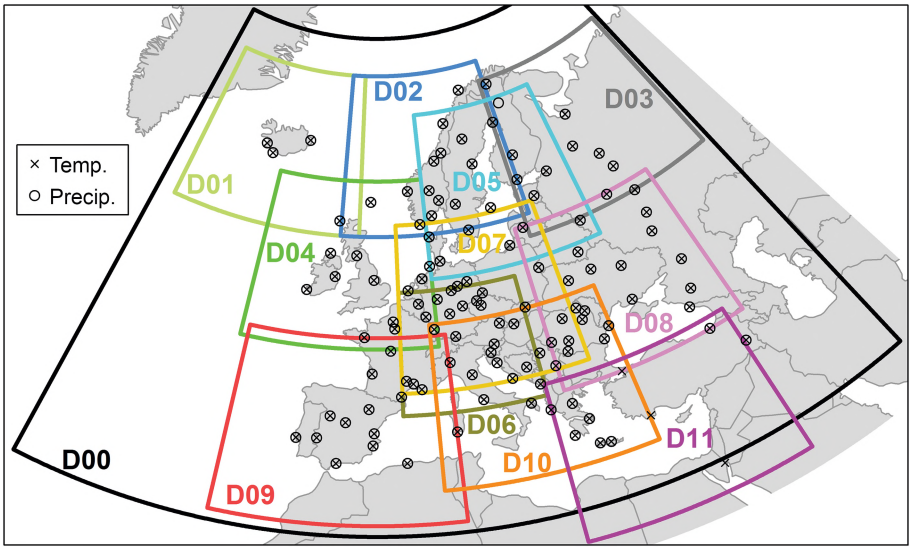

Figure 1. Domains (D00–D11) and stations used in the analysis.

article. A deeper insight into the effects of the domain size on the quality of classifications is provided elsewhere in this issue by Beck *et al.* (2016).

In short, the current study aims to assess the synoptic-climatological applicability of the circulation classifications assembled in version 2.0 of the COST733cat database, i.e. to determine how well they stratify surface weather (climate) conditions in Europe. Specifically, the effect of the selection of the classification method and its parameters is examined. To this end, several statistical criteria and climate data sets are employed.

2. Data and methods

The analysis covers the period of 1961–2000. Winter (December–February) and summer (June–August) seasons are analysed separately.

2.1. Classifications

The COST733cat database, version 2.0 (hereafter referred to simply as 'COST733cat'), contains 423 classifications for each of the 12 domains, which are displayed in Figure 1. One domain (D00) covers almost the entire area of Europe together with adjacent parts of the North-Eastern Atlantic Ocean, while the other domains (D01–D11) cover individual European regions, usually with mutual overlaps. All the classifications (except the subjective ones) are based on data from the ERA-40 reanalysis (Uppala *et al.*, 2005). The methods are grouped into families according to the basic features of the classification algorithm: subjective (not included in our evaluation; see below), threshold-based, principal component analysis-based, leader algorithm-based, and optimization. A method based on the random partition of data is also included for comparison. The classifications differ from each other in (1) the classification method, (2) input variables [whether only sea level pressure (SLP) is used for classification or SLP plus additional variable(s)], (3) whether individual daily patterns or 4-day sequences of daily patterns are classified, (4) seasonality (whether one classification is defined for the whole year or separate classifications are defined for individual seasons), and (5) the number of types (classifications with the numbers of types as close to 9, 18, and 27 as possible were formed). For more details on the classifications and the software used to calculate them, please refer to Philipp *et al.* (2016).

All the 423 classifications available in the database are evaluated, but only a subset is further analysed. We omit four groups of classifications from further consideration. (1) *Subjective classifications and their objectivised versions* are omitted because they have just one realization, which precludes them from comparisons about the number of types, sequencing, additional variables, seasonality of the definition, etc. (altogether seven classifications). (2) *Classifications not calculated by the COST733 software* (which were typically produced by individual authors before the COST733 software was made available) are omitted because they are either identical to one of classifications produced by the COST733 software or differ marginally from the latter because of rounding errors; also, they cannot be used for assessing the sensitivity to sequencing, additional classified variables, etc. (altogether 41 classifications). (3) *The objective weather classification method* after Dittmann *et al.* (1995) is omitted because it uses, as the only method in the database, many input variables for its definition even in its basic version; this precludes a fair comparison with all other methods, which are based on SLP only in their basic version (altogether seven classifications). (4) *The self-organizing maps method* is omitted because its outputs were identical to the Sandra method (SAN) (one classification).

Hence, 367 classifications enter the 'competition'. A brief summary of the methods evaluated in this study

Table 1. Overview of classifications used.

Family	Abbreviation	Method	Number of classifications	Numbers of types	Sequences of days	Input variables	Annual/ seasonal definition
Threshold	GWT	Grosswetter types (prototype classification)	3	10, 18, 26	1	SLP	A
	JCT	Jenkinson–Collison (Lamb)	3	10, 18, 26	1	SLP	A
	LIT	Litynski	3	9, 18, 27	1	SLP	A
PCA	KRZ	Kruizinga	30	8, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
	PXE	Principal components' extreme score	30	8, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
	PCT	Obliquely rotated PCA in a T-mode	30	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
	PTT	Orthogonally rotated PCA in a T-mode	30	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
Leader	LND	Lund (correlation)	30	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
	KIR	Kirchhofer (sums-of-squares)	3	9, 18, 27	1	SLP	A
	ERP	Erpicum	35	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A, S
Optimization	CKM	k-means clustering	35	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A, S
	CAP	k-means clustering preceded by hierarchical clustering	35	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A, S
	PXK	k-means clustering with extreme principal component scores as seeds	30	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A
	SAN	Simulated annealing (SANDRA)	35	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A, S
Random	RAC	Random medoid classification	35	9, 18, 27	1, 4	SLP, SLP + Z5, SLP + TH, SLP + V5, SLP + Z5 + TH + V5	A, S

More information on methods and individual classifications can be found in Philipp *et al.* (2016). Input variables: sea level pressure (SLP), 500 hPa heights (Z5), 1000/500 hPa thickness (TH), 500 hPa vorticity (V5).

is provided in Table 1 along with their abbreviations, brief descriptions, and available settings. Four methods have three realizations (i.e. result in three different classifications) for three different numbers of types (GWT, JCT, LIT, KIR). Six methods have 30 realizations (for all possible combinations of three numbers of types, two sequence lengths, and five sets of variables: KRZ, PXE, PCT, PTT, LND, PXK), while five other methods have 35 realizations (same as for the previous group plus a seasonal definition for all five sets of variables with a single number of types and without sequencing: ERP, CKM, CAP, SAN, RAC).

The presence of rare (or even not occurring at all, i.e. empty) types may potentially adversely affect the results of the analysis. Also, the infrequent types cannot be considered typical representative patterns, going, therefore, against the goal of synoptic classifications, viz. to find representative, typical, recurring patterns. We therefore omit classes with ten and fewer days in a particular season (the term 'infrequent types' is used for the omitted types in the text below). As a result, the numbers of types in a classification that we analyse may differ from the real numbers of types. We illustrate this issue using the numbers of omitted types for 15 classification methods with

Table 2. Numbers of types with population of ten or fewer days for classifications with SLP as the input field, 27 theoretical types, without sequencing, and using the annual definition.

	No. of types	D00	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11
							Winter						
Method													
GWT	26	1	0	0	0	0	0	0	0	0	0	0	0
JCT	26	1	0	7	1	1	0	7	1	0	0	1	6
LIT	27	0	0	0	0	0	0	0	0	0	0	0	0
KRZ	27	0	0	0	0	0	0	0	0	0	0	0	3
PXE	27	9	16	12	13	9	12	12	11	14	7	11	6
PCT	27	4	3	5	8	7	9	9	10	11	9	9	15
PTT	27	16	19	19	16	17	21	21	19	18	19	20	20
LND	27	3	1	1	1	0	0	1	0	1	2	1	1
KIR	27	0	0	0	0	0	0	0	0	0	0	0	0
ERP	27	11	7	9	9	5	7	8	7	7	7	7	11
CKM	27	0	0	0	2	0	2	5	1	1	0	2	4
CAP	27	0	0	0	0	0	0	0	0	0	0	0	4
PXK	27	10	16	13	13	9	13	12	11	14	8	12	5
SAN	27	1	0	0	0	0	0	0	0	0	0	0	5
RAC	27	1	0	1	0	0	0	0	0	0	1	1	4
							Summer						
Method													
GWT	26	5	0	0	0	0	0	0	0	0	1	0	16
JCT	26	0	0	2	0	0	1	17	0	1	7	15	19
LIT	27	0	0	0	0	0	0	0	0	0	0	0	0
KRZ	27	2	0	0	0	0	0	1	0	0	1	1	7
PXE	27	9	16	11	13	9	12	12	11	14	7	11	10
PCT	27	17	3	10	14	6	14	13	14	18	18	16	20
PTT	27	19	16	13	16	19	15	16	16	14	21	19	24
LND	27	5	0	0	0	0	0	1	0	1	7	2	19
KIR	27	1	0	0	0	0	0	0	0	0	3	2	9
ERP	27	16	15	19	16	13	12	17	16	16	12	14	15
CKM	27	15	11	12	14	8	13	17	15	14	15	16	20
CAP	27	13	5	7	6	4	5	5	8	10	12	11	17
PXK	27	8	16	11	13	9	12	12	11	14	9	11	12
SAN	27	14	6	5	7	4	7	6	8	9	12	12	18
RAC	27	13	5	4	6	5	7	6	9	12	15	11	14

The original number of types is also shown in the second column.

about 27 types, SLP as the only input variable, without sequencing, and using the annual definition, both for winter and summer (Table 2). One can see that (1) several methods are prone to producing fairly high numbers of infrequent types (e.g. PTT, PCT, PXE, PXK, ERP), i.e. they either exhibit a strong seasonality in the frequency of types or produce very small or even empty types; (2) the tendency to a larger number of empty types (i.e. to fewer types that occur with a non-zero frequency) is stronger in summer; and (3) the tendency to a larger number of empty types is particularly strong in domain D11 (eastern Mediterranean).

2.2. Temperature and precipitation data

Two databases of surface temperature and precipitation were used: station data taken from the database of the European Climate Assessment and Dataset (ECA&D) project (Klok and Klein Tank, 2009) and gridded data from the ERA-40 reanalysis (Uppala *et al.*, 2005). These databases allow daily precipitation totals to be evaluated in both ECA&D and ERA-40. Daily maximum and minimum temperatures are analysed in ECA&D while daily mean temperatures are analysed in ERA-40.

We retrieved data from 122 stations with daily surface maximum and minimum temperature series and 120 stations with daily precipitation series from the ECA&D database. Stations with complete time series were selected for the analysis. The locations of the stations are shown in Figure 1. For the evaluation of each domain, only data from stations within that domain are used. The number of stations in each domain is given in Table 3. Note that the domains overlap, and thus, some stations are included in more than one domain.

Daily 1200 UTC data for 2 m temperature and precipitation were extracted from the gridded ERA-40 reanalysis data set (Uppala *et al.*, 2005). For the large domain (D00), a spatial resolution of 2° (latitudinal) by 3° (longitudinal) was chosen whereas a 1° × 1° resolution was selected for the smaller domains (D01–D11). Only gridpoints over land were used in the analyses (see last column in Table 3 for the number of gridpoints available in each domain).

Table 3. Numbers of stations and gridpoints (over land) within each domain.

Domain number	Region	No. of stations	No. of gridpoints
00	Europe	121 (120)	330
01	Iceland	4	50
02	Western Scandinavia	16	175
03	Northeastern Europe	11 (12)	469
04	British Isles	18	100
05	Baltic Sea	21 (22)	269
06	Alps	25	138
07	Central Europe	42	256
08	Eastern Europe	24	319
09	Western Mediterranean	15	212
10	Central Mediterranean	33 (32)	206
11	Eastern Mediterranean	9 (7)	275

Numbers in parentheses indicate stations with precipitation data if different from the numbers of stations with temperature data.

2.3. Evaluation metrics

The synoptic-climatological applicability of circulation classifications is assessed by three different metrics, which were selected from a wide range of possibilities: explained variance (EV), Kolmogorov–Smirnov statistic (KS), and pseudo-F statistic (PSF). The station data were subjected to the evaluation using the EV and KS metrics, while the EV and PSF metrics were applied to reanalysis data.

EV is calculated as

$$\mathrm{EV} = 1 - \frac{\sum_{j=1}^{K}\sum_{i=1}^{n}\left(y_{ji} - \overline{y}_j\right)^2}{\sum_{i=1}^{n}\left(\overline{y}_i - \overline{y}\right)^2} \qquad (1)$$

where K is the number of classes (circulation types), n is the number of observations, $\overline{y}$ is the overall mean, and $\overline{y}_j$ is the mean for class j. The EV index can reach values from 0 to 1; 0 means no skill to stratify climatic data into types, whereas 1 means that there is no within-type variability (the data within every circulation type are uniform). The value of the EV index depends on the number of classes; the higher the number of classes, the better the division of data into these classes.

The PSF, according to Calinski and Harabasz (1974), is calculated as

$$\mathrm{PSF} = \frac{\sum_{j=1}^{K}\left(\overline{y}_j - \overline{y}\right)^2/(K-1)}{\sum_{j=1}^{K}\sum_{i=1}^{n}\left(y_{ji} - \overline{y}_j\right)^2/(n-K)} \qquad (2)$$

where K denotes the number of circulation types and n the number of observations, $\overline{y}$ is the mean overall observations, and $\overline{y}_j$ is the mean for class j. Increasing PSF values indicate larger variability between classes and/or less variability within classes, pointing to a better separability of classes.

The evaluation by the KS statistic follows the description in Huth (2010). Here, we repeat it in brief. For each circulation type in each classification and at each station, the conditional empirical probability distribution function (PDF) of a given element (maximum or minimum temperature) is constructed. It is then compared, using the two-sample Kolmogorov–Smirnov test, with the PDF at that station for the rest of data, i.e. for all days except those classified with the given type. This ensures that the two samples for which PDFs are compared are independent, which would not be the case if we conducted the comparison against the distribution of all data. It is important to note that the KS test reflects a whole PDF, not only the mean value; thanks to it, a type connected with a narrow temperature (precipitation) distribution around the long-term mean may be seen as having temperature different from the overall conditions, which would not be possible if, e.g. the standard t-test for the equality of means was employed. We employ the 5% significance level for the KS test. The rejection of the test indicates that values of the climate element (temperature or precipitation) under a particular type are well separated from the values in the rest of data; if the test is not rejected, the climate variable under a particular type does not significantly differ from the rest of the data. The numbers of rejections of the KS test are counted over individual types for each classification at each station. The larger the number of rejections, the better the stratification of surface temperature by the particular classification at a given station. The percentage of rejections is then calculated at each station; it serves as a basis for subsequent evaluations.

2.4. Averaging and ranking

Each of the evaluation criteria (EV, KS, PSF) produces values of its metrics at every site (station or gridpoint) for each classification. Separate evaluations (i.e. rankings) are constructed for each evaluation criterion. The values of the metrics are ranked at each station. The ranks are calculated so that the best method ranks first, i.e. the classification with the lowest KS statistic, or the highest EV, or the highest value of PSF is ranked first. The ranks are then averaged over sites located inside each of the 12 domains, thereby providing a domain mean rank for each classification. Finally, the domain mean ranks are ranked.

As a result, each classification in every domain and for every evaluation metric is assigned a rank between 1 and 423. It is worth stressing here again that the ranks are attributed to all the 423 classifications in the COST733cat database, i.e. the ranks of the classifications range from 1 (best) to 423 (worst), but only 367 of them are evaluated and compared.

One could argue that areal averaging of ranks, and not directly of the values of metrics, may not be optimum because it may exaggerate real differences for the moderately performing classifications, while underrating them for the outliers (the best and worst performing ones). However, in fact, the two approaches lead only to little differences in the results (not shown), which do not affect the conclusions.

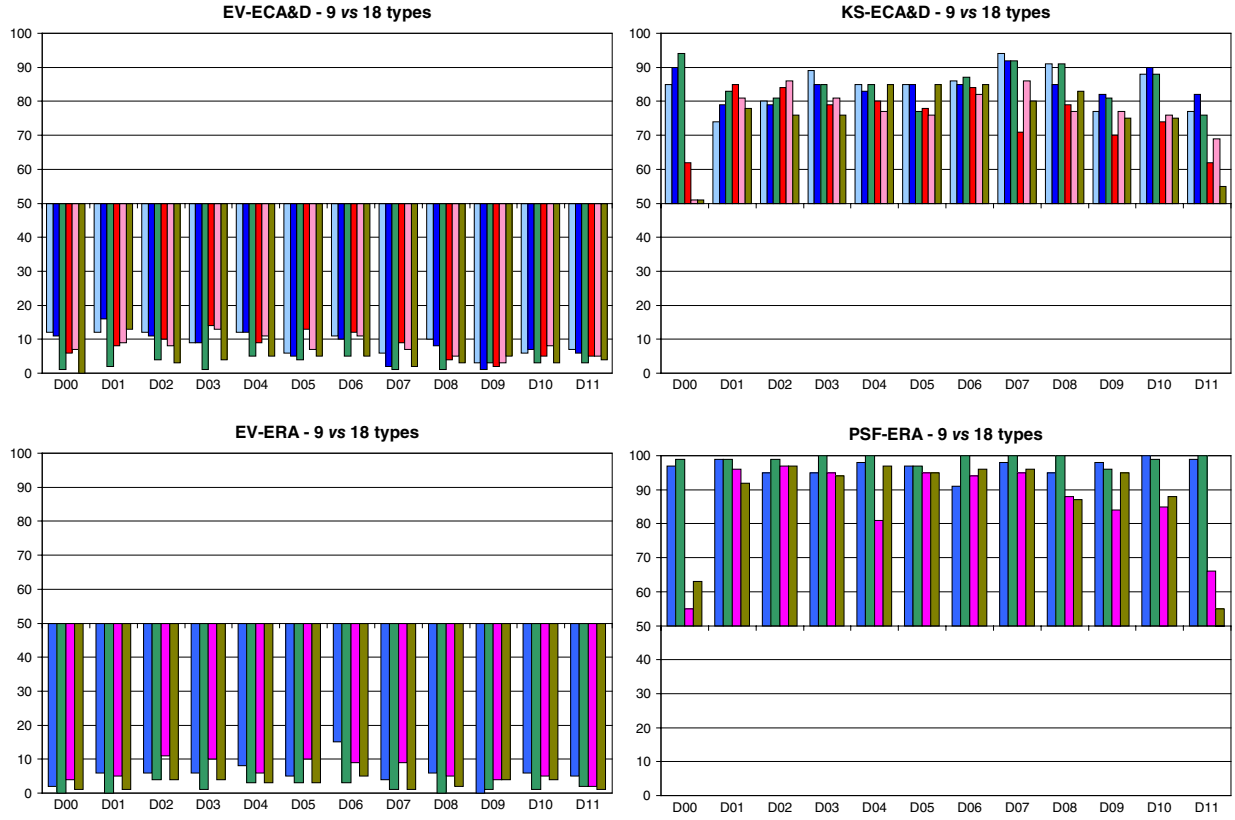

Figure 2. Comparison of performance of classifications with different number of types (approximately 9 *vs* approximately 18 in top four graphs and approximately 18 *vs* approximately 27 in bottom four graphs). Displayed are the percentages of classifications for which the lower number of types has the lower rank, i.e. values over (below) 50 indicate that the classifications with a lower (higher) number of types are more frequently better according to a given criterion. Mean differences in ranks (for details see text) between the classifications with approximately 9 and 18 types (left four graphs) and approximately 18 and 27 types (right four graphs). Separate graphs are shown for the EV criterion on ECA&D data, EV criterion on ERA-40 data, KS criterion on ECA&D data, and PSF criterion on ERA-40 data. Each cluster of bars corresponds to one spatial domain (D00–D11). The bars describe (from left to right, for ECA&D data) maximum temperature in winter (light blue), minimum temperature in winter (dark blue), precipitation in winter (green), maximum temperature in summer (red), minimum temperature in summer (light pink), and precipitation in summer (olive). For the ERA-40 data, the bars describe mean temperature in winter (blue) and mean temperature in summer (pink) instead of maximum and minimum temperatures.

3. Effect of the number of types

Before evaluating classification methods themselves, we concentrate on the effect the number of types has on the synoptic-climatological quality of classifications. Unlike effects of all other options, the effect of the number of types (and, consequently, of sample sizes in the types) is inherent to the individual evaluation criteria. The criteria are constructed so that they prefer either small numbers of types (implying large sample sizes) or large numbers of types (implying small sample sizes). For example, in the KS test, the difference between the two PDFs necessary to become statistically significant gets smaller with increasing sizes of the samples compared. As a consequence, the KS criterion favours classifications with small numbers of types; this behaviour was documented and discussed for the previous version of the COST733cat database by Huth (2010). The PSF criterion behaves in the same way, while the EV criterion behaves in the opposite way, preferring small type sizes to large type sizes, i.e. preferring large numbers of types to small numbers of types.

The effect of the number of types is examined on the classifications based on SLP only, without sequencing, for the annual definition. For each criterion, each domain, and each climate variable, one pair of classifications with approximately 9 and 18 types and one pair of classifications with approximately 18 and 27 types were formed. The pairs, for which the classification with a lower number of types has a lower rank, i.e. it exhibits a better separation according to the given criterion, are counted. The percentage of these counts is displayed in Figure 2. The percentage values above (below) 50 indicate a tendency towards a higher (lower) number of types being better separated, the tendency being stronger for the percentage values being closer to 100 (0).

The prevalence of the change in synoptic-climatological applicability in one direction (whether improvement or deterioration) is larger for smaller numbers of types (9 *vs* 18); this seems natural because the relative change in the number of types as well as in their average size is then also larger. The expected sign of differences prevails in all cases; only for the KS and PSF criteria in summer in domains D00 and D11, the percentage values are close to 50, indicating that the numbers of improvements and deteriorations are close to each other. We may speculate

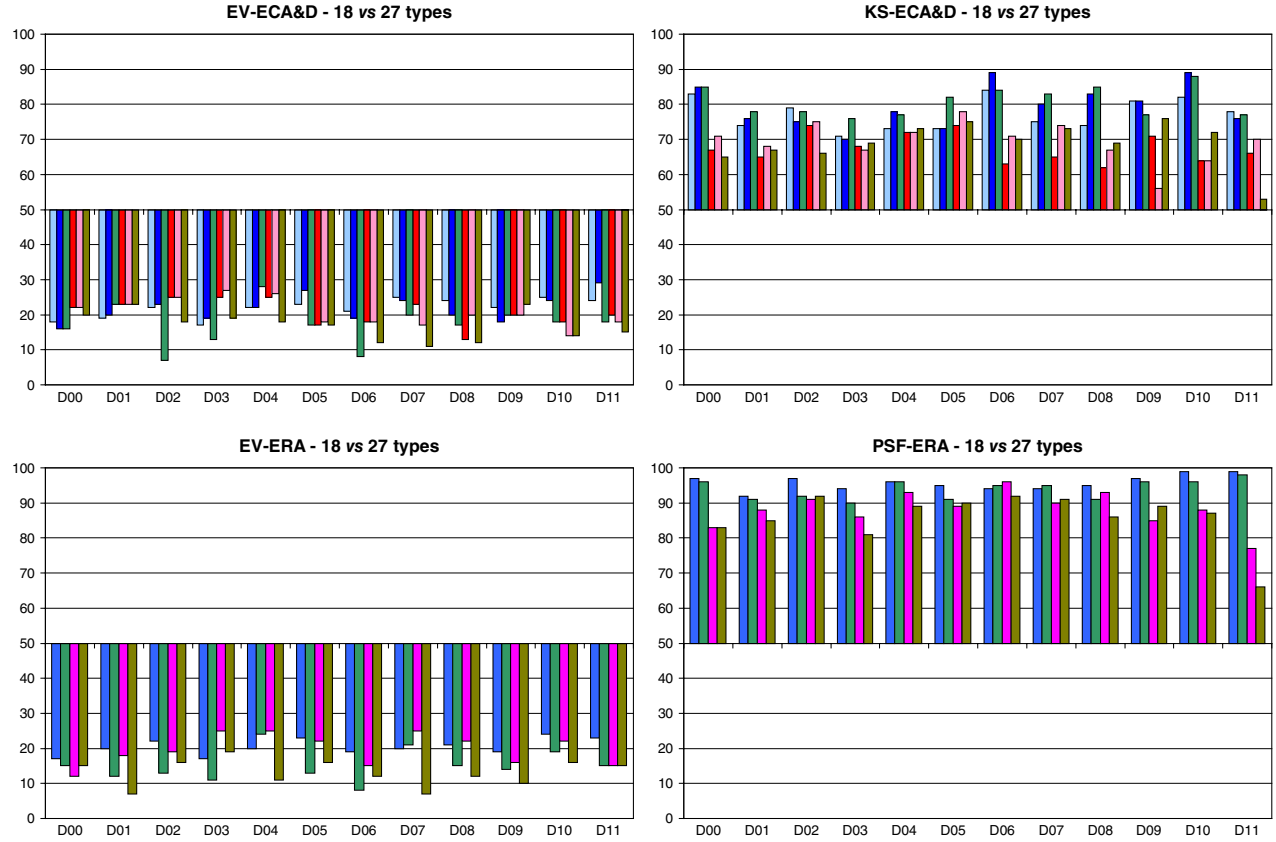

Figure 2. Continued

about the cause. In D11 (Eastern Mediterranean), many classifications have large numbers of empty or infrequent types in summer, resulting in the real numbers of types being much lower than their theoretical numbers; the differences are then frequently calculated between classifications with similar real numbers of types. Domain D00 (whole Europe) is the largest one. The results may indicate that the low number of types (about nine) may be insufficient to capture the variability of surface climate in summer. The EV criterion tells us that the smaller number of types is worse anyway and is not able to make the distinction that in D00, in summer, the classifications with small numbers of types worsen even more.

Another observation is that the KS criterion is the least sensitive to the number of types. The differences tend to be lower in summer, although not in all domains. This appears to reflect the fact that the effect of atmospheric circulation on surface climate is generally weaker in summer. There is no systematic difference between variables: the effect of the number of types is similar for both temperatures and precipitation. We observe no apparent difference between the two data sets (ECA&D *vs* ERA-40 for the EV criterion) either.

4. Comparison of classification methods

To compare the performance of individual classification methods, the effect of the number of types, as well as of other parameters of classifications, must be eliminated. For the comparison, similar to the analysis of the effect of the number of types in Section 3, we choose only 'basic' classifications, i.e. those based on SLP, without sequencing, and for the annual definition. The way the classifications are ranked is illustrated in Table 4, showing an example for the KS criterion for maximum temperature in winter in domain D00. The overall ranks of the classifications (i.e. numbers from 1–423; second to fourth row in Table 4) are ranked separately for the classifications with approximately 9, 18, and 27 types (fifth to seventh row in Table 4). The sum of these ranks is calculated over the three numbers of types (eighth row in Table 4), and the sums are eventually ranked (last row). The final rank is free of the effects of the number of types, and it represents an overall synoptic-climatological performance of the given method for the given climate variable over a given domain for a particular evaluation criterion.

The same ranking procedure is applied to all variables in all domains for both winter and summer and for all the evaluation criteria. The result for the KS criterion is displayed in 'abacus-like' diagrams in Figure 3. Obviously, the spread of ranks is large for most methods, which are among the best for some variables and/or domains, while being among the worst for other variables and/or domains. Nevertheless, some methods tend to be better overall (the symbols cluster in the left part of the graph; e.g. SAN, CAP, CKM, LIT, and GWT in winter; PXK and CKM in summer), whereas some methods tend to be worse

Table 4. An example of the ranking procedure for the evaluation of classification methods: maximum temperature, winter, KS criterion, domain D00, classifications of SLP only, without sequencing, annual definition.

Classification		GWT	JCT	LIT	KRZ	PXE	PCT	PTT	LND	KIR	ERP	CKM	CAP	PXK	SAN	RAC
Rank among all classifications	9 types	169	229	325	212	147	184	254	243	323	308	142	113	187	81	82
	18 types	260	262	259	274	302	315	398	385	408	299	277	293	336	245	338
	27 types	330	397	381	362	373	341	403	388	423	399	305	365	344	335	340
Rank	9 types	6	10	15	9	5	7	12	11	14	13	4	3	8	1	2
	18 types	3	4	2	5	9	10	14	13	15	8	6	7	11	1	12
	27 types	2	12	10	7	9	5	14	11	15	13	1	8	6	3	4
Sum of ranks		11	26	27	21	23	22	40	35	44	34	11	18	25	5	18
Final rank		2.5	10	11	6	8	7	14	13	15	12	2.5	4.5	9	1	4.5

Ranks among all classifications are shown for each classification method in the upper part of the table for classifications with approximately 9, 18, and 27 types. They are ranked for each number of types separately; these ranks are displayed in rows 5–7. Sums of the latter ranks are calculated (row 8), resulting in final ranking (last row).

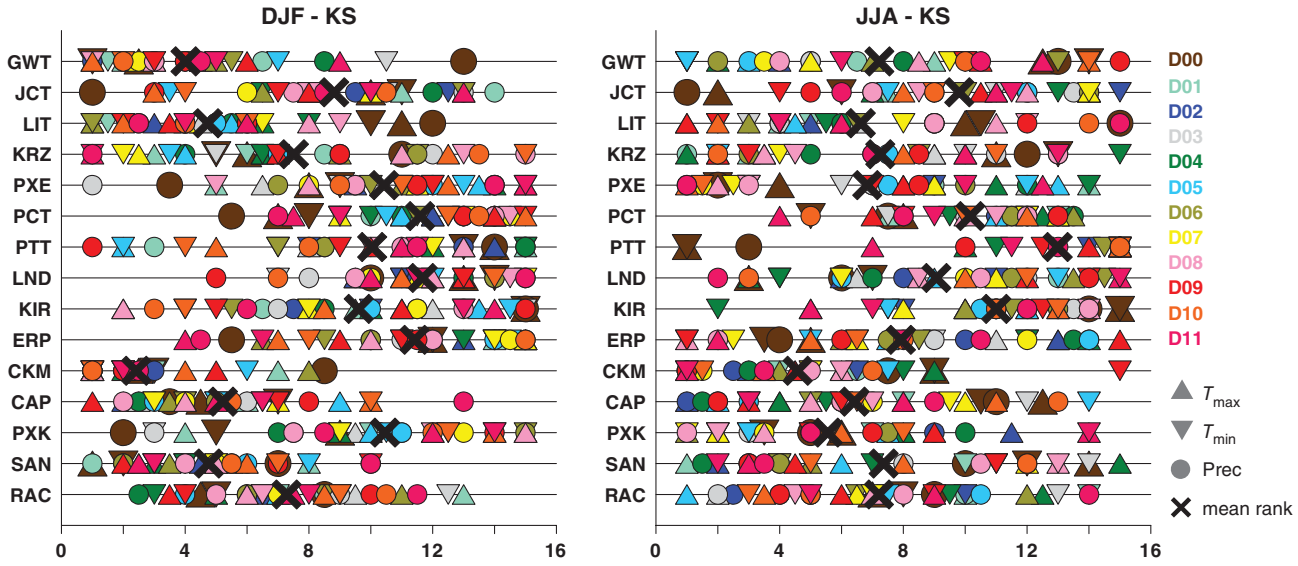

Figure 3. Ranking of classification methods (in rows), KS criterion, DJF (left), and JJA (right). Maximum temperature shown in upward triangles, minimum temperature in downward triangles, precipitation in circles. Colour coding of domains (D00–D11) is the same as in Figure 1 and is shown on the right-hand side of the figure; in general, domains with cool and wet (warm and dry) climate are denoted by cold (warm) colours. The large domain (D00) is denoted by larger symbols. The black cross denotes the average of all ranks. Note that the symbols may overlap, and hence, some of them may remain hidden.

(symbols clustering in the right part of the graph; e.g. ERP, LND, PCT in winter; KIR and PTT in summer) than others. Also, a different performance between winter and summer may be noted.

Is there a difference in the performance of the classification methods between the climate variables or from one domain to another? These questions are hard to answer based on the graphs in Figure 3; an aggregation of information is needed. Therefore, we averaged the ranks in Figure 3 over the variables to produce the dependence of the performance on domains; the result is shown in Figure 4 for all evaluation criteria and both seasons. Analogously, the ranks were averaged over the domains, resulting in the dependence of the performance on the climate variables; this is displayed in Figure 5.

There is no apparent general geographical dependence of rankings (e.g. north *vs* south, west *vs* east) for any method; the specific behaviour usually appears for one season or one criterion only and therefore cannot be generalized. For example, the GWT method in summer performs better in the cool and wet (northwestern) domains (bluish and greenish colours in Figure 4) than in warm and dry (southeastern) domains (reddish colours) according to the KS criterion, but this behaviour does not repeat for other criteria and in winter. Perhaps only the JCT method tends to perform consistently better in the southeastern domains than in the northwestern domains. However, several methods suggest they are sensitive to the size of the domain; they differ in their performance between the large domain (D00; large brown symbols) and the other smaller domains. Most striking is this difference for LIT, which is among the leading methods for small domains in both seasons for all the criteria but performs much worse on the large domain. A similar behaviour is observed for KIR and GWT. Both the LIT and GWT methods are based on the similarity of circulation patterns with structures of a domain-scale, which

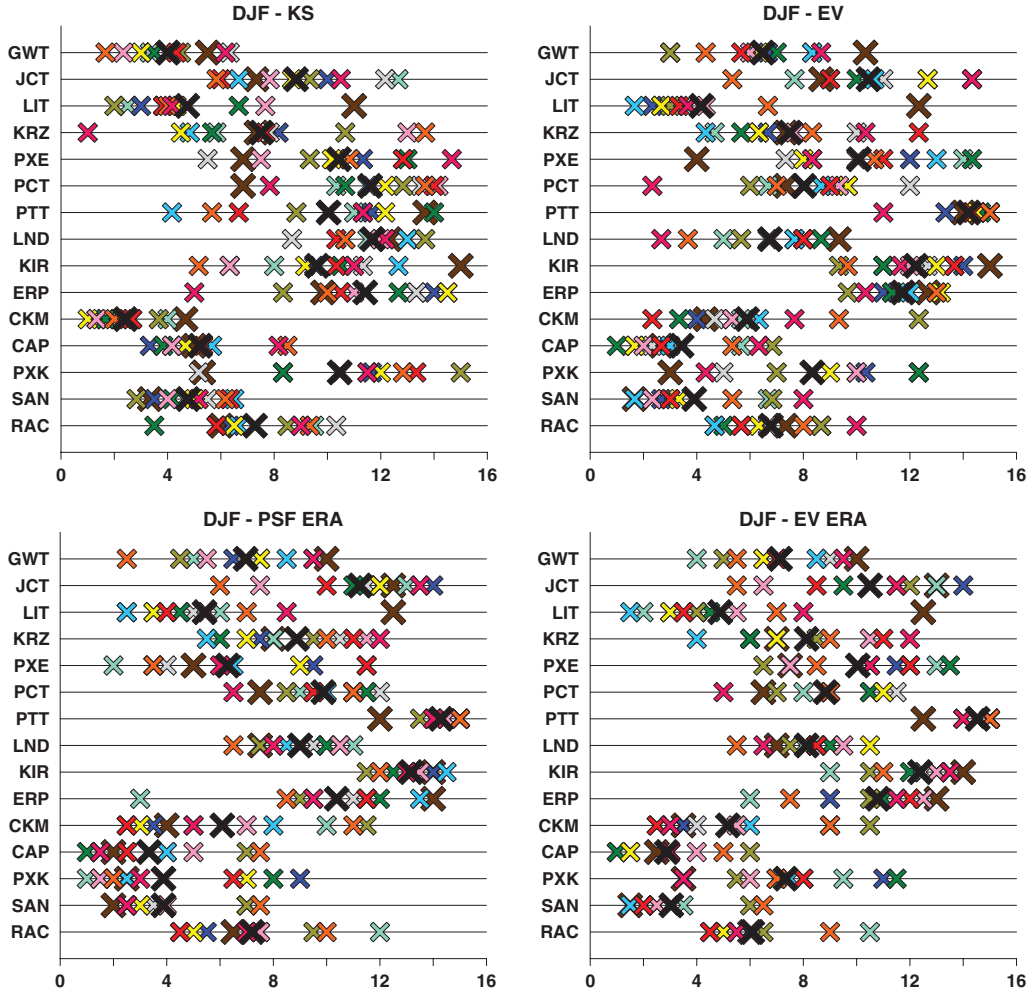

Figure 4. Dependence of the performance of the classification methods (in rows) on the domain. Averages of ranks over variables are displayed by coloured crosses (for colour coding see Figure 3; D00 denoted by a larger brown cross; the average rank for each method in black). Season, criterion, and data set indicated on the top of each graph.

are too large in the large domain to be able to account for local climate peculiarities. The opposite behaviour, i.e. a better performance for the large domain, is seen for PXE and PXK in winter and for JCT in summer.

Some of the methods differ in their performance for temperature on one side and precipitation on the other (Figure 5). LIT, KRZ, PCT, and KIR are better in characterizing temperature for most criteria in both seasons, while JCT, PXE, PXK, CKM, and RAC are more suitable for stratifying precipitation. We may hypothesize that the latter methods capture more detailed structures relevant for precipitation, whereas the former methods tend to capture larger-scale features more relevant for temperature.

The varied performance of the methods in different domains and for different climate variables suggests that no method can be identified as the best (the same holding for the worst) and that a single ranking would have a limited information value. Nevertheless, in the end, we decided to provide the final ranking of the methods as it may provide general guidelines on their synoptic-climatological applicability. Table 5 displays the ranking of the ranks averaged both over domains and climate variables (this is denoted by black crosses in Figures 3 and 4) separately for individual criteria and seasons. The methods that perform well are highlighted in red and pink, while the inferior methods are in blue. Several methods are inferior in both seasons; JCT, PTT, KIR, and ERP do not appear among the top six methods according to any criterion in either season. On the other hand, only LIT is among the top six methods in all the eight cases. A further four methods, viz. GWT, KRZ, CAP, and SAN, perform very well (they are among the top six) in one season while performing moderately in the other season. There is a considerable seasonal difference in the performance of several methods and, in addition, a marked difference among the evaluation criteria. In particular, the rankings according to the KS criterion tend to differ from the other two criteria; this is very strongly pronounced for the PCT, LND, and CKM methods in summer. A general explanation for the different performance of different criteria is that each classification method uses a specific similarity metric, and if this similarity metric corresponds well to (or is even identical to) one of the criteria, the performance of the method according to this criterion is naturally improved. For example, if we used a criterion based on the Euclidean distance, the methods using Euclidean distance to define types (such

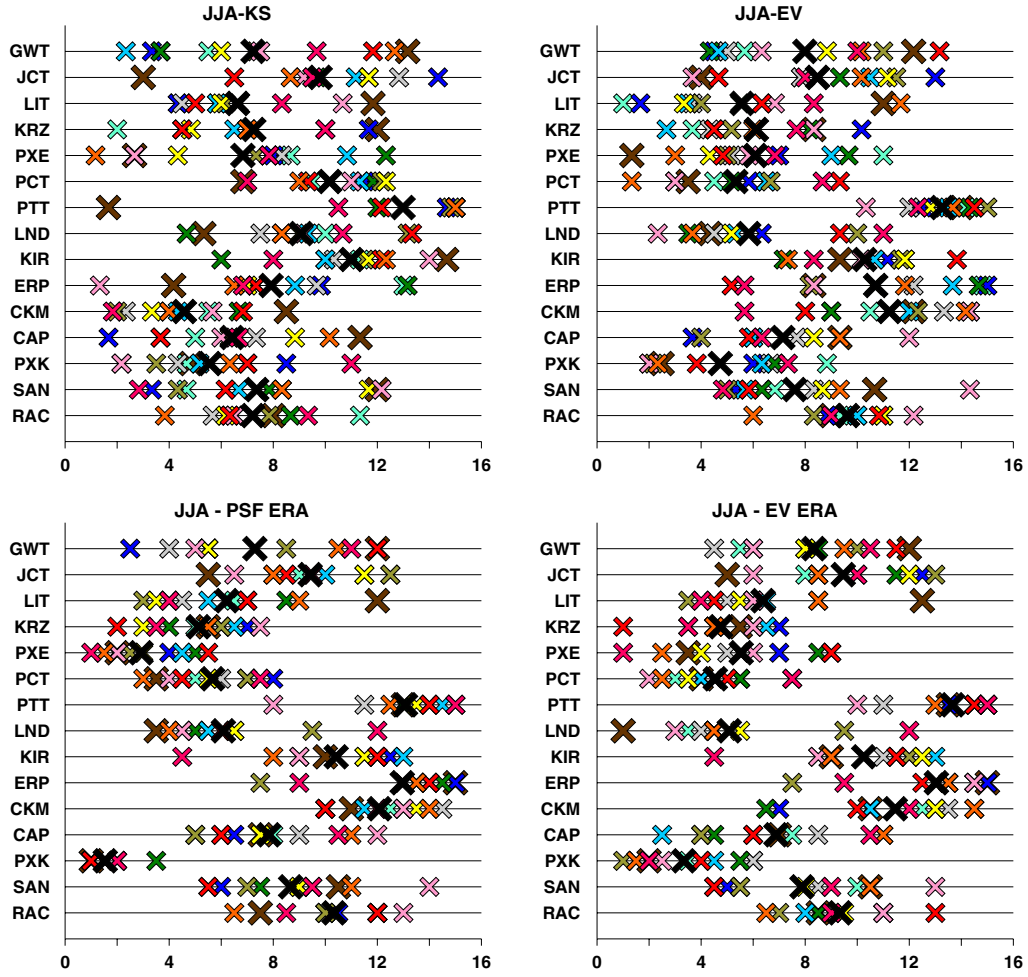

Figure 4. Continued

as cluster analysis) would be favoured. It is worth noting that, unlike the evaluation criterion, the evaluation database affects the results rather marginally: differences in the ranks between the ECA&D and ERA-40 data sets for the EV criterion are very small, not exceeding three. Another interesting feature is a different performance of the LND and KIR methods, which are very similar in their algorithm, differing in the (dis)similarity measure used: LND uses correlation, while KIR uses root-mean-square difference, the latter being further refined by also taking into account the similarity in smaller scales. This refinement appears, however, to lead paradoxically to a deterioration of the ability to stratify surface climate elements. We can also see that sophisticated methods are not a guarantee of good synoptic-climatological applicability. The simplest possible method, RAC, consisting in the assignment of datapoints to randomly selected centroids on a shortest distance basis, performs better than several other methods intentionally designed for the classification of circulation patterns.

The total rank in Table 5 should be interpreted with caution. The bottom methods (PTT, ERP, KIR, JCT, PCT) are not recommendable for synoptic-climatological analyses, although they may work well under some circumstances. On the other hand, the methods that ended up on the top in our 'competition', i.e. LIT, CAP, PXK, SAN, and GWT, are likely to work well on many occasions, but their performance varies from season to season, from one climate variable to another, and also regionally. Therefore, the presence of the method on the top of the ranking list does not secure its excellent performance in all circumstances. One must carefully select the method according to the goals of a particular study and its settings, and, ideally, one should use multiple methods and multiple criteria for their evaluation in order to obtain unbiased and generalisable results.

5. Effect of sequencing

To determine the effect of sequencing, classifications based on 4-day sequences are compared with those without sequencing. Surface variables (temperature and precipitation) are attributed to the last day of a 4-day sequence, i.e. the surface climate elements are stratified by a classification based on the circulation the same day and three preceding days. To quantify the effect of sequencing, the pairs of classifications that only differ in sequencing and have all the same other attributes (number of types, input variables, seasonality) are formed and

Figure 5. Dependence of the performance of the classification methods (in rows) on the climate variable. Averages of ranks over domains are displayed by coloured symbols: upward triangle, maximum temperature; downward triangle, minimum temperature; right-pointing triangle, mean temperature; circle, precipitation. Season, criterion, and data set indicated on the top of each graph.

those for which the 4-day sequence has a lower rank (is better) than its non-sequential counterpart are counted. The percentage of such pairs of classifications is reported in Table 6 as an indicator of an improvement because of sequencing. Values exceeding (below) 50 indicate that sequencing leads to an improvement (deterioration) for a majority of classification pairs. In addition, the effect of sequencing is quantified by subtracting the ranks of the two classifications (4-day sequence minus non-sequential) in each pair. A negative difference indicates a lower rank for the sequential classification and, hence, an improvement because of the introduction of sequencing. The differences in ranks are then averaged for the overall classification pairs for each domain, climate variable, and criterion. Whether the difference is significantly different from zero is finally tested by the non-parametric Wilcoxon signed-rank test; the significance is denoted by bold types (italics) in Table 6 for significant improvements (deteriorations) because of sequencing.

One can see that in winter: (1) sequencing improves stratification by circulation classifications for temperature (with a few exceptions) but deteriorates it for precipitation. In other words, precipitation is determined mainly by instantaneous circulation conditions, whereas for temperature, the circulation on previous days is also important. (2) The improvement is larger (without any exception) for minimum temperature than for maximum temperature. In other words, the circulation on preceding days is more important for determining nighttime (minimum) temperature. (3) The improvement for temperature is largest, and the deterioration for precipitation is smallest, for the large domain. This suggests that the

Table 5. Rankings of classification methods for the two seasons and four criteria (s – stations, i.e. ECA&D database; E – ERA-40 reanalysis).

	DJF				JJA				Total
	KS	EV-s	EV-E	PSF	KS	EV-s	EV-E	PSF	
GWT	2	5	6	7	7	9	9	7	5
JCT	8	12	12	13	12	10	11	10	12
LIT	3	3	3	4	4	3	6	6	1
KRZ	7	8	8.5	9	8	6	3	3	6
PXE	11	11	11	6	5	5	5	2	8
PCT	14	9	10	11	13	2	2	4	11
PTT	10	15	15	15	15	15	15	15	15
LND	15	6	8.5	10	11	4	4	5	9
KIR	9	14	14	14	14	12	12	12	13
ERP	13	13	13	12	10	13	14	14	14
CKM	1	4	4	5	1	14	13	13	7
CAP	5	1	1	1	3	7	7	8	2
PXK	12	10	7	2.5	2	1	1	1	3
SAN	4	2	2	2.5	9	8	8	9	4
RAC	6	7	5	8	6	11	10	1	10

Colour coding is used to emphasize the best and worst performing methods for each criterion and season (red – top three; pink – 4th–6th; blue – bottom three; pale blue – 10th–12th). The final ranking in the rightmost column was obtained by ranking the ranks averaged over the eight columns (two seasons times four criteria).

circulation features on preceding days that are relevant for determining temperature tend to be remote from the site where the temperature is recorded, and a large domain is more likely to comprehend them than smaller domains. (4) The only domain where the sequencing deteriorates results for maximum as well as mean temperature is the northwestern-most domain, D01, covering Iceland. It is the high cyclonic activity in this domain, connected with a low persistence of circulation, that is a likely reason for such behaviour.

In summer, the positive effect of sequencing on temperature tends to be weaker. For maximum temperature, the statistically significant improvements are rather scarce, whereas deteriorations are much more frequent, some of them being significant, especially for the EV criterion. The negative effect of sequencing on precipitation is strong, though not as strong as in winter according to the EV and PSF criteria. Similar to winter, the improvement for both temperature variables tends to be largest and deterioration for precipitation smallest for the large domain (D00).

6. Effect of the selection of classified variables

The next question to answer is whether the inclusion of 500 hPa geopotential height, thickness of the layer between 1000 and 500 hPa, representing mean temperature in the lower troposphere, and vorticity at 500 hPa among the classified variables in addition to SLP improves the specification of temperature and precipitation. Analogous to the previous section, the effect of including additional input variables is quantified by counting pairs of classifications differing only in the input variables (SLP plus (an) additional variable(s) minus SLP only), the other attributes (number of types, sequencing, seasonality) being equal, for which the additional input variable results in an improvement (i.e. lower rank). Also, statistical significance of the mean difference in ranks is calculated in a way analogous to the previous section.

Results are shown in Table 7 only for the KS criterion in order to save space. Results for the other criteria are qualitatively similar. The discussion starts with winter. There is a prevalent significant improvement in stratification of both temperature variables because of the addition of 500 hPa heights as well as 1000/500 hPa thickness. Smaller improvements, some of which fall below the significance level, are found in central Europe and the Baltic area (D05, D06, and D07); the British Isles (D04) is the only domain where no significant improvements appear. The addition of heights leads to the strongest improvement for temperature in southern and southeastern Europe (domains D08–D11). The response to adding height or thickness is more varied for precipitation; there is a tendency towards significant improvements in southern and southeastern Europe (D08, D10, D11, and to a smaller extent, also D09). The improvements caused by adding height tend to be larger than those caused by adding thickness. Vorticity at 500 hPa does not contribute to the synoptic-climatological applicability of classifications: only one of 36 cases shows a significant improvement by adding vorticity, while almost a half of the cases show significant deterioration.

In summer, the improvements of the stratification of temperature caused by adding height and thickness are stronger in the southern part of Europe (D08–D11) where the improvement is larger than in winter and stronger for minimum temperature than for maximum temperature. Adding heights leads to the deterioration, mostly significant, for precipitation in northern and northwestern Europe (D01–D05), whereas improvements, though mostly insignificant, appear in central, southern, and southeastern Europe (D06–D11). Adding thickness leads to less improvement and more deterioration for precipitation than adding height. Vorticity in summer brings an improvement

Table 6. Effect of sequencing for winter (top) and summer (bottom).

	TX		TAVG		TN		RR			
	KS	EV-s	EV-E	PSF	KS	EV-s	KS	EV-s	EV-E	PSF
Domain DJF										
00	**90**	**98**	**97**	**99**	**93**	**100**	*43*	*4*	*0*	*1*
01	46	*18*	*26*	*25*	**65**	**75**	*22*	*1*	*0*	*0*
02	**72**	**79**	**75**	**72**	**85**	**93**	*8*	*0*	*0*	*1*
03	**68**	**68**	**90**	**86**	**75**	**90**	*8*	*0*	*0*	*0*
04	50	*45*	**70**	**67**	**81**	**96**	*13*	*0*	*0*	*0*
05	**58**	*38*	58	56	**74**	**74**	*13*	*0*	*0*	*0*
06	**69**	*38*	**61**	**59**	**75**	**79**	*10*	*1*	*0*	*0*
07	**68**	61	**79**	**76**	**72**	**92**	*15*	*0*	*0*	*0*
08	**70**	**61**	**58**	58	**81**	**87**	*15*	*3*	*0*	*0*
09	55	*31*	**96**	**96**	**85**	**99**	*19*	*2*	*2*	*2*
10	**62**	45	**75**	**74**	**81**	**93**	*13*	*1*	*0*	*0*
11	56	*12*	**86**	**85**	**67**	41	*17*	*1*	*0*	*1*
Domain DJF										
00	**85**	**81**	**79**	**84**	**90**	**90**	*31*	*12*	*3*	*4*
01	*35*	*30*	*31*	*32*	**64**	**88**	*11*	*0*	*0*	*0*
02	**61**	47	**61**	**61**	**82**	**98**	*9*	*0*	*0*	*0*
03	49	39	46	45	56	**60**	*13*	*1*	*0*	*0*
04	47	*43*	**70**	**67**	**80**	**98**	*6*	*0*	*0*	*0*
05	45	45	**57**	55	**68**	**82**	*14*	*0*	*0*	*0*
06	*37*	*22*	*41*	*40*	**65**	**75**	*7*	*0*	*0*	*0*
07	48	*28*	55	55	**70**	**86**	*5*	*0*	*0*	*0*
08	50	*30*	44	*40*	57	**58**	*10*	*1*	*1*	*1*
09	*38*	*17*	45	46	**77**	**82**	*14*	*2*	*4*	*3*
10	55	*30*	41	*40*	**65**	**67**	*10*	*0*	*0*	*0*
11	*34*	*24*	*28*	*30*	**65**	**59**	*11*	*7*	*2*	*2*

Displayed are percentages of classifications for which the sequencing results in an improvement (i.e. classifications of 4-day sequences have a lower rank than classifications without sequencing). Values over (below) 50 indicate a prevailing improvement (deterioration) in synoptic-climatological applicability caused by sequencing. The entries for which differences in ranks between the classifications of 4-day sequences and classifications without sequencing are significantly different from zero according to the Wilcoxon signed-rank test are marked in bold (italics) for the improvement (deterioration). TX, maximum temperature; TN, minimum temperature; TAVG, daily mean temperature; RR, precipitation amount.

in most domains for temperature and in several domains for precipitation (though mostly insignificant).

The change in performance caused by simultaneously adding all the three extra variables in both seasons reflects the changes because of single input variables. For both temperatures, improvements are observed in general, with the exception of D04 (and D01 in summer), while mostly deterioration or little effect is observed for precipitation, with the exception of southeastern and southern Europe (D08–D11).

In an attempt to generalize, we may claim that the addition of more input variables into classification is most beneficial in southeastern Europe, whereas it results in lesser improvements or even a prevalent deterioration in the most maritime domains (D01, D04, D05). Vorticity has much smaller potential to improve the stratification of climate variables than mid-tropospheric heights and lower tropospheric thickness.

7. Effect of seasonality

Here, the classifications defined separately for four seasons (seasonal definition) are compared with those defined for the whole year (annual definition). The effect of the seasonality is quantified analogously to the previous sections. It is worth mentioning that the number of available pairs of classifications is lower than in the previous two comparisons because the seasonally defined classifications were constructed for five classification methods only, for a single number of types (seven in each season), and for instantaneous fields, i.e. not for 4-day sequences (cf. Table 1, symbols 'A' and 'S' in the last column). Consequently, the confidence intervals around the mean differences in ranks are much wider, which results in a lower significance for differences of the same magnitude. Another issue to point out is the fact that all seasonal classifications are designed to have seven types. They are, however, compared with the annual classifications with the designed number of types of nine, which is the lowest number of types available. In view of a bias introduced by the unequal number of types in the compared classifications, the results should be treated with caution since a part of the improvement, if there is any, may be attributable to a lower number of types, not to the different definition of classifications.

The effect of the seasonality of the definition for the KS criterion is displayed in Table 8. The seasonal definition is better than the annual definition in almost all cases (all domains, all three climate variables, both seasons), although the effect is not significant many times. Only five entries (all in summer) indicate deterioration.

Table 7. Effect of the selection of classified variables: winter (top) and summer (bottom) for the KS criterion only.

DJF	SLP + HGT			SLP + TH			SLP + VOR			SLP + HGT + TH + VOR		
Domain	TX	TN	RR	TX	TN	RR	TX	TN	RR	TX	TN	RR
00	**74**	**74**	**76**	**71**	**76**	**76**	44	*33*	42	**67**	**65**	**65**
01	**62**	**65**	53	61	**79**	45	44	56	39	**64**	**76**	41
02	**62**	**68**	61	**61**	**70**	*39*	39	41	47	**65**	**74**	50
03	**82**	**73**	**77**	**65**	61	52	*32*	*29*	45	**70**	**73**	56
04	52	44	52	62	52	52	39	*45*	48	47	50	50
05	**73**	**74**	47	64	56	*38*	*35*	*27*	*29*	**64**	**71**	*35*
06	**70**	**70**	62	62	65	65	50	*39*	55	**79**	**77**	56
07	61	**65**	55	53	**68**	39	*33*	*33*	48	**73**	**76**	53
08	**77**	**88**	**62**	**76**	**80**	**65**	*35*	*36*	*35*	**74**	**77**	59
09	**80**	**58**	65	**62**	59	58	**61**	*30*	45	**88**	**76**	58
10	**77**	**83**	**68**	**61**	**76**	58	44	*33*	47	**85**	**80**	**79**
11	**73**	**76**	**82**	**76**	**73**	**76**	55	42	41	**85**	**70**	**67**
JJA	SLP + HGT			SLP + TH			SLP + VOR			SLP + HGT + TH + VOR		
Domain	TX	TN	RR	TX	TN	RR	TX	TN	RR	TX	TN	RR
00	**68**	**70**	55	62	**71**	50	45	53	47	**67**	**79**	58
01	*39*	**71**	*27*	*26*	58	*15*	39	55	*23*	*39*	**67**	*9*
02	**59**	**76**	*23*	58	**71**	*11*	**68**	**70**	61	**82**	**86**	*11*
03	**67**	**73**	*36*	56	**73**	*33*	**80**	**74**	41	**88**	**88**	*21*
04	56	**68**	*24*	45	**73**	*8*	35	42	56	56	**77**	*27*
05	**76**	**82**	*27*	53	**74**	*21*	**74**	**64**	44	**86**	**89**	*18*
06	**85**	**94**	62	**76**	**92**	*35*	**74**	**71**	55	**89**	**95**	56
07	**82**	**95**	56	**77**	**88**	50	**73**	**61**	64	**89**	**94**	53
08	**82**	**89**	55	**80**	**91**	42	**73**	53	56	**94**	**91**	**67**
09	**73**	**89**	**97**	**77**	**89**	**85**	55	55	**70**	**65**	**82**	**97**
10	**97**	**92**	**61**	**92**	**95**	53	**77**	65	55	**95**	**95**	**74**
11	**83**	**80**	**71**	**79**	**80**	**61**	58	52	53	**85**	**88**	**73**

Shown are the percentages of classifications for which the additional variable (HGT = 500 hPa height, TH = 1000/500 hPa thickness, VOR = 500 hPa vorticity) results in an improvement relative to the classification based on SLP only. Otherwise analogous to Table 6.

Table 8. Effects of the seasonality of the definition of classifications for the KS criterion.

Domain	Winter			Summer		
	TX	TN	RR	TX	TN	RR
00	**76**	**88**	**84**	**92**	**68**	**96**
01	**72**	**76**	52	**84**	72	**92**
02	**80**	**72**	**64**	**84**	64	**84**
03	52	60	**88**	44	52	**64**
04	**76**	**68**	72	**80**	**80**	64
05	64	60	72	68	60	**60**
06	68	**72**	**88**	48	32	60
07	**72**	**84**	72	64	**76**	**76**
08	**76**	**84**	68	**68**	**72**	**76**
09	72	**68**	**96**	**80**	**84**	56
10	52	68	**76**	**84**	**76**	**80**
11	**72**	**76**	48	52	*16*	48

Shown are the percentages of classifications for which the seasonal definition results in an improvement relative to the annual definition. Otherwise as in Table 6.

8. Conclusions

This study provides an evaluation of the database of classifications of circulation patterns, reported by Philipp *et al.* (2016), in terms of their ability to stratify surface temperature and precipitation on a daily basis – i.e. in terms of their synoptic-climatological applicability. It is worth saying that analogous conclusions for other climatic and environmental variables may be different. The main conclusions can be summarized in the following points.

The most general statement is that the synoptic-climatological applicability of classification methods considerably varies among climate variables (maximum temperature, minimum temperature, precipitation), across domains, and between seasons.

A strong sensitivity of the synoptic-climatological applicability, whatever criterion is used to its quantification, to the number of types, which was reported earlier on a much smaller set of classifications and domains, has been confirmed.

Some classifications contain types that are empty or infrequent in one of the seasons; such types are excluded from the evaluation. To be more specific, the types with the frequency of ten or fewer days in the analysed period are excluded. The unequal numbers of the types that are retained in the analysis (i.e. their frequency exceeds the threshold of 10 days) affect the results because the performance of a classification depends on the real number of types. This contamination of results is stronger in summer than in winter because empty and infrequent types are more common in summer.

Nevertheless, several well-performing methods can be identified. These include CKM (simple k-means clustering), CAP (k-means clustering preceded by

hierarchical cluster analysis), LIT (Litynski's method), and GWT (prototype classification).

Several methods must be used with great caution because their synoptic-climatological applicability appears to be inferior on many occasions (but they, of course, can have other positive properties for which they may be useful in other applications). They include PCT (obliquely rotated PCA), PTT (orthogonally rotated PCA), LND (Lund's correlation method), KIR (Kirchhofer's sums-of-squares method), and ERP (Erpicum's method).

The performance of methods does not manifest considerable differences among climate variables for which the methods are evaluated. In other words, there is no method that would be particularly suitable (or, conversely, unsuitable) just for one of the climate variables. Only LIT in summer seems to be more applicable to temperature than precipitation, but it is questionable whether this fact can be generalized.

There are hints of systematic differences in the synoptic-climatological applicability of the methods between the large (D00) and smaller (other) domains: GWT, LIT, KIR, and CKM perform relatively better on smaller domains, whereas JCT (Jenkinson–Collison), PXE (principal component's extreme scores), PCT, and to some extent also PXK (k-means clustering with extreme principal component scores as seeds) perform better on the large domain.

In summer, there is a tendency for several classification methods to display a geographical dependence of their synoptic-climatological applicability: JCT and ERP perform better in the southern domains, while GWT, KIR, and PXK perform worse there; GWT performs better in the northern domains, whereas JCT, PXE, and ERP tend to perform worse there. In winter, no such regional dependency of performance is noticed.

No similarities in behaviour can be found for classifications coming from the same family; in other words, the nature of the classification algorithm is only a weak driver for a classification's synoptic-climatological performance.

Classifications of 4-day sequences are usually better in stratifying surface temperature than ordinary instantaneous classifications; the opposite is true for precipitation. The improvement of temperature stratification because of sequencing is larger in winter than in summer. This has a clear interpretation: temperature in mid latitudes is governed by processes, whether advective or radiative, on relatively longer time-scales than precipitation; and the temporal scale of these processes is longer in winter when advective processes are stronger than radiative ones.

It is beneficial to use mid-tropospheric height or thickness as a classified variable in addition to SLP if one wishes to stratify temperature. The improvement because of the additional variable appears to increase with the continentality of the location. Adding mid-tropospheric vorticity to SLP as a classified variable provides some benefit in stratifying temperature in summer only. SLP seems to be sufficient for stratification of precipitation as the improvements caused by the additional classified variables are infrequent.

Acknowledgements

This study benefitted very much from the international cooperation within COST733 Action 'Harmonisation and Applications of Weather Types Classifications for European Regions'. Particular thanks go to several individuals who made the COST733 Action successful and smoothly running: Ole Einar Tveito, Meteorological Institute, Oslo, Norway, for chairing the action, Andreas Philipp, University of Augsburg, Germany, for his decisive contribution to the production of the classification database and software, and Massimiliano Pasqui, Institute of Biometeorology, IBIMET-CNR, Florence, Italy, and Pere Esteban, Snow and Mountain Research Centre of Andorra, Sant Julià de Lòria, Andorra, for leading the working group on the evaluation of classifications, within which this work was accomplished. The support from the Ministry of Education, Youth, and Sports of the Czech Republic, projects OC115 and LD12059, and from the Czech Science Foundation, project P209/12/P811, is acknowledged. We are grateful to two anonymous reviewers for their constructive comments, leading especially to improvements in the clarity of presentation. The authors declare no conflict of interest.

Supporting Information

The following supporting information is available as part of the online article:
Appendix S1. Performance of the classification methods in terms of the explained variance (EV) for ERA-40 data.

References

Anagnostopoulou C, Tolika K, Maheras P, Kutiel H, Flocas HA. 2008. Performance of the general circulation HadAM3P model in simulating circulation types over the Mediterranean region. *Int. J. Climatol.* **28**: 185–203.

Barry RG, Carleton AM. 2001. *Synoptic and Dynamic Climatology*. Routledge: London, New York, 620 pp.

Beck C, Philipp A. 2010. Evaluation and comparison of circulation type classifications for the European domain. *Phys. Chem. Earth* **35**: 374–387.

Beck C, Weitnauer C, Jacobeit J. 2014. Downscaling of monthly PM10 indices at different sites in Bavaria (Germany) based on circulation type classifications. *Atmos. Pollut. Res.* **5**: 741–752.

Beck C, Philipp A, Streicher F. 2016. The effect of domain size on the relationship between circulation type classifications and surface climate. *Int. J. Climatol.* **36**: 2692–2709, doi: 10.1002/joc.3688.

Bednorz E. 2008. Synoptic conditions of snow occurrence in Budapest. *Meteorol. Z.* **17**: 39–45.

Brisson E, Demuzere M, Kwakernaak B, van Lipzig NPM. 2011. Relations between atmospheric circulation and precipitation in Belgium. *Meteorol. Atmos. Phys.* **111**: 27–39.

Calinski T, Harabasz J. 1974. A dendrite method for cluster analysis. *Commun. Stat.* **3**: 1–27.

Casado MJ, Pastor MA, Doblas-Reyes FJ. 2010. Links between circulation types and precipitation in Spain. *Phys. Chem. Earth* **35**: 437–447.

Cassano EN, Lynch AH, Cassano JJ, Koslow MR. 2006. Classification of synoptic patterns in the western Arctic associated with extreme events at Barrow, Alaska, USA. *Clim. Res.* **30**: 83–97.

Dayan U, Tubi A, Levy I. 2012. On the importance of synoptic classification methods with respect to environmental phenomena. *Int. J. Climatol.* **32**: 681–694.

Demuzere M, Werner M, van Lipzig NPM, Roeckner E. 2009. An analysis of present and future ECHAM5 pressure fields using a classification of circulation patterns. *Int. J. Climatol.* **29**: 1796–1810.

Demuzere M, Kassomenos P, Philipp A. 2011. The COST733 circulation type classification software: an example for surface ozone concentrations in Central Europe. *Theor. Appl. Climatol.* **105**: 143–166.

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

Donat MG, Leckebusch GC, Pinto JG, Ulbrich U. 2010. Examination of wind storms over Central Europe with respect to circulation weather types and NAO phases. *Int. J. Climatol.* **30**: 1289–1300.

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

Fernau ME, Samson PJ. 1990. Use of cluster analysis to define periods of similar meteorology and precipitation chemistry in eastern North America. Part II: precipitation patterns and pollutant deposition. *J. Appl. Meteorol.* **29**: 751–761.

Fleig AK, Tallaksen LM, Hisdal H, Stahl K, Hannah DM. 2010. Inter-comparison of weather and circulation type classifications for hydrological drought development. *Phys. Chem. Earth* **35**: 507–515.

Gaetani M, Pasqui M, Crisci A, Guarnieri F. 2016. A synoptic characterization of the dust transport and associated thermal anomalies in the Mediterranean basin. *Int. J. Climatol.* **36**: 2779–2791, doi: 10.1002/joc.3615.

García C, Martí G, Oller P, Moner I, Gavaldà J, Martínez P, Peña C. 2009. Major avalanches occurrence at regional scale and related atmospheric circulation patterns in the Eastern Pyrenees. *Cold Reg. Sci. Technol.* **59**: 106–118.

Huth R. 2010. Synoptic-climatological applicability of circulation classifications from the COST733 collection: first results. *Phys. Chem. Earth* **35**: 388–394.

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

Huth R, Beck C, Tveito OE. 2010. Preface. *Phys. Chem. Earth* **35**: 307–308.

Jacobeit J, Rathmann J, Philipp A, Jones PD. 2009. Central European precipitation and temperature extremes in relation to large-scale atmospheric circulation types. *Meteorol. Z.* **18**: 397–410.

Jiang N, Hay JE, Fisher GW. 2005. Synoptic weather types and morning rush hour nitrogen oxides concentrations during Auckland winters. *Weather Clim.* **25**: 43–69.

Jones PD, Lister DH. 2009. The influence of the circulation on surface temperature and precipitation patterns over Europe. *Clim. Past* **5**: 259–267.

Kassomenos P. 2010. Synoptic circulation control on wild fire occurrence. *Phys. Chem. Earth* **35**: 544–552.

Kidson JW. 1997. The utility of surface and upper air data in synoptic climatological specification of surface climatic variables. *Int. J. Climatol.* **17**: 399–413.

Klok EJ, Klein Tank AMG. 2009. Updated and extended European dataset of daily climate observations. *Int. J. Climatol.* **29**: 1182–1191.

Kostopoulou E, Jones PD. 2007. Comprehensive analysis of the climate variability in the eastern Mediterranean. Part II: relationships between atmospheric circulation patterns and surface climatic elements. *Int. J. Climatol.* **27**: 1351–1371.

Küttel M, Luterbacher J, Wanner H. 2011. Multidecadal changes in winter circulation-climate relationship in Europe: frequency variations, within-type modifications, and long-term trends. *Clim. Dyn.* **36**: 957–972.

Leckebusch GC, Weimer A, Pinto JG, Reyers M, Speth P. 2008. Extreme wind storms over Europe in present and future climate: a cluster analysis approach. *Meteorol. Z.* **17**: 67–82.

Lorenzo MN, Taboada JJ, Gimeno L. 2008. Links between circulation weather types and teleconnection patterns and their influence on precipitation patterns in Galicia (NW Spain). *Int. J. Climatol.* **28**: 1493–1505.

Lykoudis SP, Kostopoulou E, Argiriou AA. 2010. Stable isotopic signature of precipitation under various synoptic classifications. *Phys. Chem. Earth* **35**: 530–535.

Makra L, Mika J, Bartzokas A, Béczi R, Borsos E, Sümeghy Z. 2006. An objective classification system of air mass types for Szeged, Hungary, with special interest in air pollution levels. *Meteorol. Atmos. Phys.* **92**: 115–137.

Maraun D, Osborn TJ, Rust HW. 2011. The influence of synoptic airflow on UK daily precipitation extremes. Part I: observed spatio-temporal relationships. *Clim. Dyn.* **36**: 261–275.

Nishiyama K, Endo S, Jinno K, Uvo CB, Olsson J, Berndtsson R. 2007. Identification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a self-organizing map. *Atmos. Res.* **83**: 185–200.

O'Hare G, Sweeney J. 1993. Lamb's circulation types and British weather: an evaluation. *Geography* **78**: 43–60.

Paegle JN. 1974. Prediction of precipitation probability based on 500-mb flow types. *J. Appl. Meteorol.* **13**: 213–220.

Philipp A, Bartholy J, Beck C, Erpicum M, Esteban P, Fettweis X, Huth R, James P, Jourdain S, Kreienkamp F, Krennert T, Lykoudis S, Michalides S, 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.

Philipp A, Beck C, Huth R, Jacobeit J. 2016. Development and comparison of circulation type classifications using the COST 733 dataset and software. *Int. J. Climatol.* **36**: 2673–2691, doi: 10.1002/joc.3920.

Pineda N, Esteban P, Trapero L, Soler X, Beck C. 2010. Circulation types related to lightning activity over Catalonia and the principality of Andorra. *Phys. Chem. Earth* **35**: 469–476.

Rasilla DF, García-Codrón JC. 2016. Regional and local scale atmospheric forcing upon sea level along the coast of SW Europe. *Int. J. Climatol.* **36**: 2792–2809, doi: 10.1002/joc.3646.

Rasilla DF, García-Codron JC, Carracedo V, Diego C. 2010. Circulation patterns, wildfire risk and wildfire occurrence at continental Spain. *Phys. Chem. Earth* **35**: 553–560.

Raziei T, Bordi I, Pereira LS, Corte-Real J, Santos JA. 2012. Relationship between daily atmospheric circulation types and winter dry/wet spells in western Iran. *Int. J. Climatol.* **32**: 1056–1068.

Schiemann R, Frei C. 2010. How to quantify the resolution of surface climate by circulation types: an example for Alpine precipitation. *Phys. Chem. Earth* **35**: 403–410.

Schuenemann KC, Cassano JJ, Finnis J. 2009. Synoptic forcing of precipitation over Greenland: climatology for 1961-99. *J. Hydrometeorol.* **10**: 60–78.

Sepp M, Saue T. 2012. Correlations between the modelled potato crop yield and the general atmospheric circulation. *Int. J. Biometeorol.* **56**: 591–603.

Stefan S, Necula C, Georgescu F. 2010. Analysis of long-range transport of particulate matters in connection with air circulation over Central and eastern part of Europe. *Phys. Chem. Earth* **35**: 523–529.

Trigo RM, Sousa PM, Pereira MG, Rasilla D, Gouveia CM. 2016. Modelling wildfire activity in Iberia with different atmospheric circulation weather types. *Int. J. Climatol.* **36**: 2761–2778, doi: 10.1002/joc.3749.

Tveito OE. 2010. An assessment of circulation type classifications for precipitation distribution in Norway. *Phys. Chem. Earth* **35**: 395–402.

Twardosz R. 2010. An analysis of diurnal variations of heavy hourly precipitation in Kraków using a classification of circulation types over Poland. *Phys. Chem. Earth* **35**: 456–461.

Ullmann A, Monbaliu J. 2010. Changes in atmospheric circulation over the North Atlantic and sea-surge variations along the Belgian coast during the twentieth century. *Int. J. Climatol.* **30**: 558–568.

Uppala SM, Kållberg PW, Simmons AJ, Andrae U, da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen A, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Bethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm 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.

Vicente-Serrano SM, López-Moreno JI. 2006. The influence of atmospheric circulation at different spatial scales on winter drought variability through a semi-arid climatic gradient in northeast Spain. *Int. J. Climatol.* **26**: 1427–1453.

Yarnal B. 1985. A 500 mb synoptic climatology of Pacific north-west coast winters in relation to climatic variability, 1948-1949 to 1977-1978. *J. Climatol.* **5**: 237–252.

Yarnal B. 1993. *Synoptic Climatology in Environmental Analysis. A Primer*. Belhaven Press: London, 195 pp.

Yarnal B, White DA, Leathers DJ. 1988. Subjectivity in a computer-assisted synoptic climatology II: relationships to surface climate. *J. Climatol.* **8**: 227–239.

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