An Assessment of Self-Organizing Maps and k-means Clustering Approaches for Atmospheric Circulation Classification

Despina Deligiorgi, Kostas Philippopoulos, and Georgios Kouroupetroglou

***Abstract*—** This study presents an analysis and comparison between the application of self-organizing maps (SOM) and the k-means clustering approaches in the field of atmospheric circulation classification, focusing in the area of southeastern Europe. Circulation type classification is a significant aspect of climate research in terms of examining the large-scale atmospheric variability and its relationship with local climate parameters. The study utilizes mean daily sea level pressure (MSLP) data for the spring months of a 62-year period (1948 to 2009) on a grid with 2.5°x2.5° resolution. Both schemes provide realistic classifications, differentiating in the number of the resulting circulation patterns. The two methods are compared by examining the distribution of each SOM circulation type members (days) to every k-means type and by investigating the pressure field correspondence along with their frequencies of occurrence. High similarity is observed, especially for the patterns where atmospheric circulation is controlled from high-pressure barometric systems. The SOM method is found to be superior, due to its ability to generate a non-linear classification and produce a map where closely related atmospheric modes are described by neighboring neurons and positioned in adjacent locations.

***Keywords*—** atmospheric circulation classification, data clustering, k-means clustering, Self-Organizing Maps.

I. Introduction

Synoptic climatology is defined as the linkage of atmospheric circulation and environmental response [1] and is often based on the successful classification of atmospheric conditions into a number of different representative states [2]. The procedure is called circulation type classification and deals with a small number of discrete circulation types for analyzing the variability of atmospheric circulation in terms of their frequency changes on different temporal and spatial scales [3]. The classification schemes can be subdivided into subjective and automated methods, depending on the procedure that is used to assign atmospheric fields into the resulting classes. The subjective or manual schemes employ the expert's knowledge for identifying the atmospheric circulation types and are typically based on the visual analysis of daily weather maps. On the contrary, automated classification schemes essentially employ statistical methods for analyzing atmospheric data, with the objective of generating groups of cases with increased internal similarity and at the same time increased external separability. An extensive database of weather and automated circulation type classification schemes in Europe is presented in [4]. According to Huth [5] the automated methods can be further classified into the following categories:

- Correlation method
- Sum-of-squares method
- Cluster analysis methods
- Principal components analysis.

The objective of this work is to examine and compare the resulting patterns from two different cluster analysis approaches by examining their correspondence using qualitative and quantitative criteria. The adopted methodology along with the essential theoretical background of the classification schemes is analyzed in the second section of this work, while the resulting circulation patterns and their comparison in the third part of this paper. In the concluding part of this work the results are discussed and a two-step classification scheme, based on the strengths and weaknesses of the two approaches, is proposed.

II. METHODS

A. Area of study and data

In this study and for classifying atmospheric circulation, mean daily averaged sea level pressure (MSLP) data are acquired from the NCEP/NCAR Reanalysis 1 project [6] that produces a global analyses record of atmospheric fields. The reanalysis dataset covers the period from 1948 to 2009 on a grid with a 2.5°x2.5° resolution. The selected spatial domain is from 30°N to 60°N and from 10°W to 37.5°E, which contains 260 grid points in total. The classification is performed for the transitional period of spring (March to May), leading to a subset of 5704 MSLP fields (days). In southeastern Europe, spring is one of the most significant seasons in terms of

This work was partially funded by the National and Kapodistrian University of Athens, Special Account for Research Grants.

D. Deligiorgi is with the Division of Environmental Physics and Meteorology, Department of Physics, National and Kapodistrian University of Athens, GR15784, Athens, Greece (corresponding author, phone: +30 2107276924; fax: +30 2106018677; e-mail: despo@phys.uoa.gr).

K. Philippopoulos with the Division of Environmental Physics and Meteorology, Department of Physics, National and Kapodistrian University of Athens, Greece (e-mail: kostasphilippopoulos@yahoo.com).

G. Kouroupetroglou is with the Division of Signal Processing and Communication, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, GR 15784 Athens, Greece (e-mail: koupe@di.uoa.gr).

atmospheric circulation as the weather alternates from cold to warm period types. The clustering algorithms treat each of the 5704 days as a different object while the 260 grid point MSLP values are the elements (variables) of each object.

B. Methodology

The atmospheric circulation for the period and area under study is examined using two different clustering approaches. The first classification scheme employs a traditional clustering algorithm (k-means). The k-means method is the most widely known data-clustering scheme and has been extensively used in environmental sciences for grouping objects into respective categories (e.g. [7] and [8]). It is a nonhierarchical clustering approach with the inherent advantage of allowing the reallocation of misplaced objects as the analysis proceeds [9]. The method defines k centroids, one for each cluster, and associates each object to the nearest centroid. It uses an iterative algorithm that finds the local minimum of the sum of object-to-centroid Euclidean distances, summed over all k clusters according to:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| \overrightarrow{x_\iota} - \overrightarrow{c_J} \right\| \qquad (1)$$

where $\left\| \overrightarrow{x_\iota} - \overrightarrow{c_J} \right\|$ is the Euclidean distance of the object $\overrightarrow{x_\iota}$ and the centroid $\overrightarrow{c_J}$.

The k-means method consists of two steps. Initially the Principal Components Analysis (PCA) is used to reduce the dimensionality of the dataset and subsequently the k-means clustering is performed. The PCA method transforms the high-dimensional space into fewer dimensions and in our case the initial 5704x260 dataset is reduced to a 5704x25 subset by using the first 25 principal components that describe the 99.01% of the total variation. This pre-processing step is essential for the efficient classification of MSLP data.

The second approach is the Self-Organizing Map (SOM) algorithm, introduced by Kohonen [10], which is an unsupervised neural network model used for classification and feature extraction of high-dimensional data. The SOM converts the complex, nonlinear statistical relations of the high-dimensional input data into simple geometric relations at a typically two-dimensional map [11]. Such a property is highly desirable in meteorology and synoptic climatology, where the nonlinearity is a primary characteristic of atmospheric field data [12]. A detailed survey of SOM applications in meteorology and oceanography is presented in [13], while a description of its applications in climate studies can be found in [14]. The SOM neural network model consists of an input layer and a two-dimensional lattice of neurons, the output or competitive layer, which is fully connected to the input space. Initially the number of neurons is selected and their weight vectors are randomly initialized. Subsequently a training vector is presented to the network and the Euclidean distances between the training vector and the neurons' weight vectors are calculated. The neuron that produces the smallest distance is called the Best Matching Unit (BMU) and its weight vectors along with its neighboring neurons weight vectors are updated towards the input vector. The input vectors are presented sequentially in the network and by using iterative training the neurons are adjusted in a way that different parts of the SOM respond similarly to certain input patterns. The final part of the SOM method is the visualization of the results, where each training vector is associated with one neuron, which represents the resulting patterns of the classification process. According to Haykin [15], the main properties of the SOM lattice are:

- The approximation of the input space, as it is estimated from the weight vectors
- Topological ordering, where a location within the lattice corresponds to a specific feature of the input patterns
- Density matching, as more neurons are allocated to represent dense areas of the input space
- Feature selection as the method selects the best features to approximate the underlying distribution.

The SOM methodology has been applied in southeastern Europe for associating wintertime precipitation and large-scale atmospheric variability [16] and for identifying synoptic patterns based on 500hpa level geopotential height [12].

The main drawback of both classification schemes is the requirement of a predefined number of clusters. In circulation type classification there is no a priori knowledge of the number of the resulting patterns and therefore both methods are repeated for a range of initial number of classes. In detail, for the k-means classification the procedure is repeated multiple times for centroids ranging from 6 to 13, while for the SOM classification for two-dimensional lattices that correspond to classes ranging from 12 to 36, with varying number of row and column neurons. The optimum number in both cases is selected from the qualitative examination of the resulting composite MSLP maps.

III. RESULS

A general remark from the multiple experiments of generating atmospheric circulation types from both classifications is that in many cases the resulting MSLP composites were suboptimal. The qualitative analysis of the resulting patterns identified an optimum number of ten clusters for the k-means classification (Figure 1) and twenty atmospheric states for the SOM classification (Figure 3), which are mapped along a 4-row and 5-column hexagonal topology. The relative frequencies of each type in both cases are presented in Figure 2 and Figure 4 respectively.

A. k-means circulation patterns

The k-means circulation classification (Figure 1) resulted into two types influenced by low-pressure systems (K1 and K2 types), in two patterns characterized by high-pressure systems (K3 and K4 types), in three smooth fields with minimal pressure gradient (K5, K6 and K7 types) and in three states where the atmospheric circulation is influenced by both high and low pressure systems (K8, K9 and K10 types). The description of the relevant circulation patterns is presented in terms of the most important atmospheric circulation characteristics in Table 1 and their relative frequencies of occurrence in Figure 2.

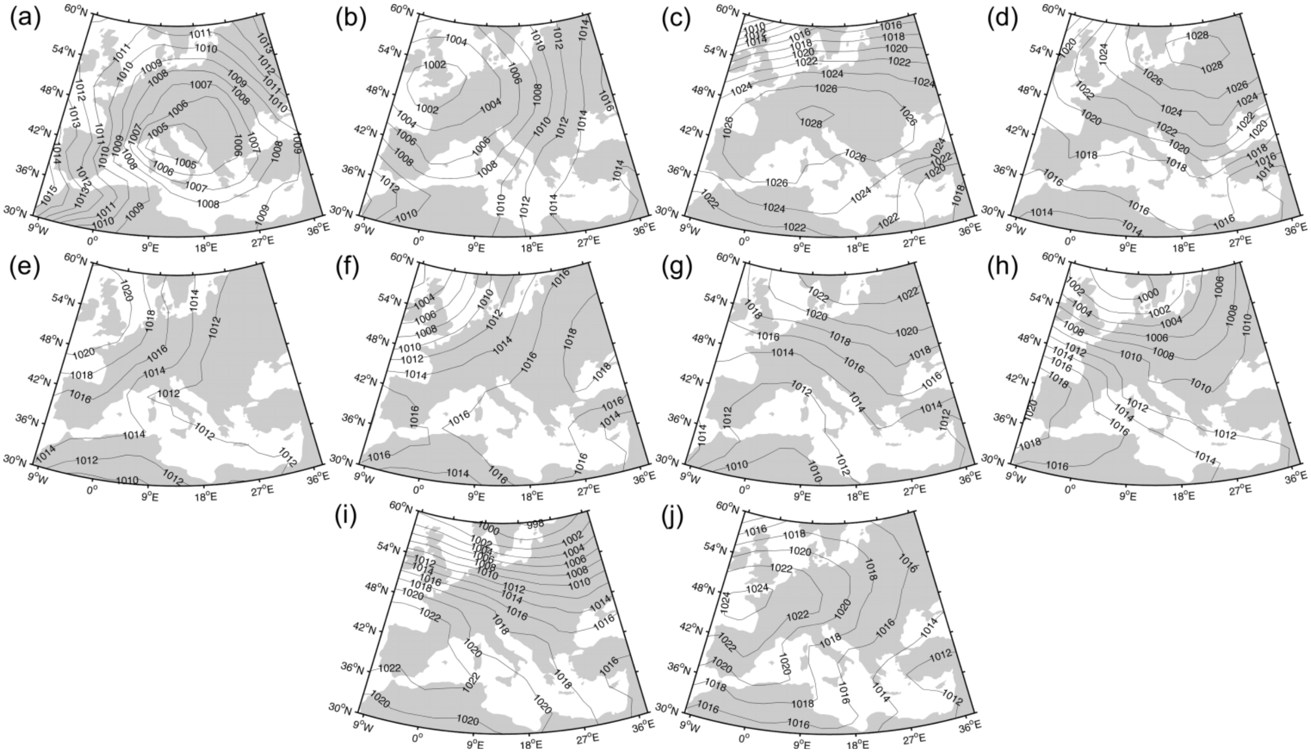

Fig. 1 The k-means classification circulation types, K1 (a), K2 (b), K3 (c)

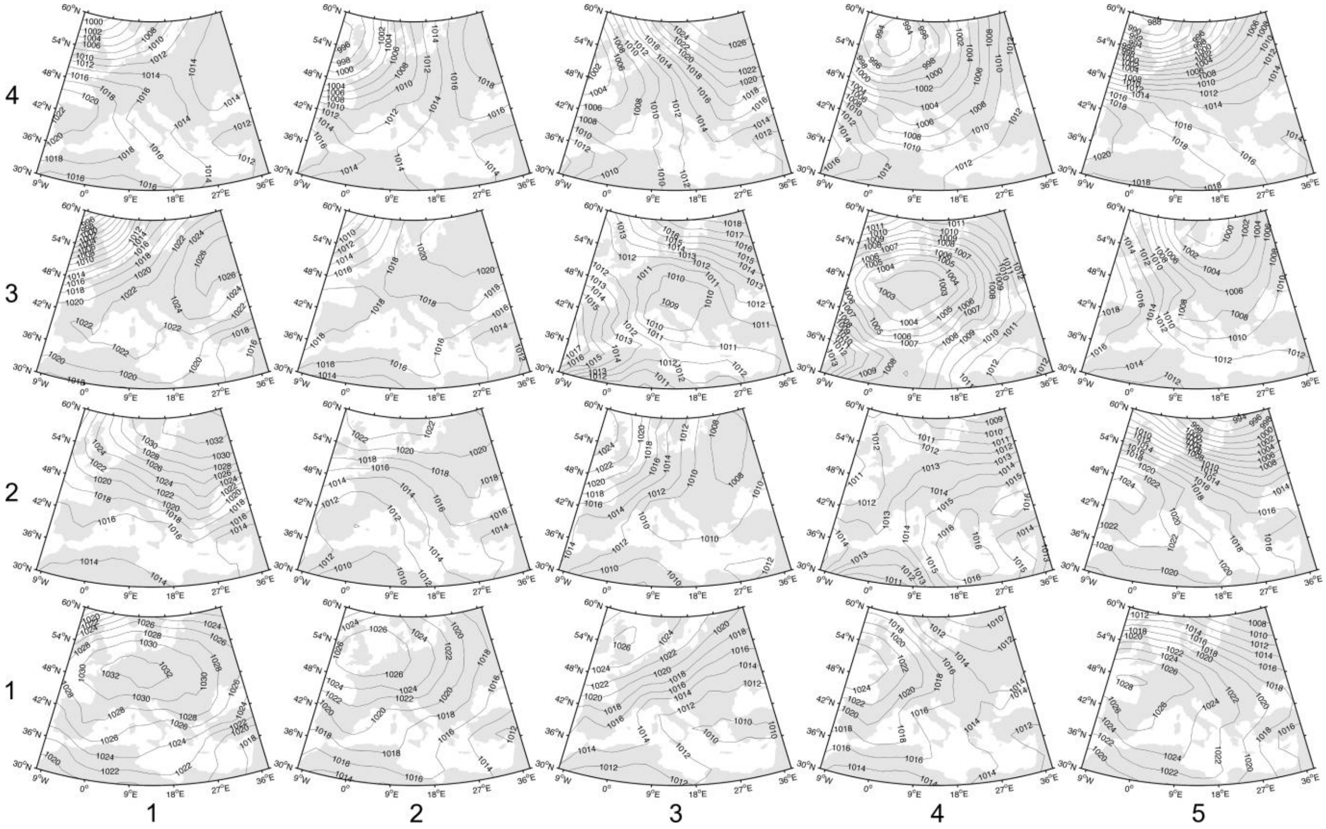

Fig. 3 The SOM classification circulation types

Table 1: Description of the k-means classification circulation types

Abbreviation	Circulation type	Description
K1	Cyclonic	Low-pressure system over central Italy
K2	Cyclonic	Low-pressure system over the British Isles
K3	Anticyclonic	Extended anticyclone over central Europe
K4	Anticyclonic	Extension of the Siberian anticyclone over western Russia and the Baltic countries
K5	Smooth	Smooth pressure field that favor the development of local flows
K6	Smooth	Smooth pressure field that favor the development of local flows
K7	Smooth	Smooth pressure field that favor the development of local flows
K8	High – Low combination	Low-pressure system in northern Europe over Nordic countries - Anticyclone in the Iberian Peninsula
K9	High – Low combination	Low-pressure system in northern Europe over Nordic countries - Anticyclone in the Iberian Peninsula
K10	High – Low combination	Easterly extension of the Azores anticyclone in western and central Europe in combination with the low-pressure field in the Middle East

Table 2: Description of the SOM classification circulation types

Abbreviation	Circulation type	Description
SOM1.1	High – Low Combination	Combination of the extended anticyclone in central Europe and the relative low-pressure field of the Middle East
SOM2.1	High – Low Combination	Combination of the extended anticyclone over the UK and the Netherlands and the relative low-pressure field of the Middle East
SOM3.1	Cyclonic	Relative low-pressure field over Greece and the Balkans
SOM4.1	High – Low Combination	High and low pressure fields at the west and east of Greece respectively
SOM5.1	Anticylconic	Anticyclone at the north of the Iberian peninsula which extends over the whole Mediterranean Sea
SOM1.2	Anticylconic	Anticyclone in northern Europe at the Baltics which extends over the Balkans and Greece
SOM2.2	Smooth	Smooth field that favor the development of local flows
SOM3.2	Smooth	Smooth field that favor the development of local flows
SOM4.2	Smooth	Smooth field that favor the development of local flows
SOM5.2	Anticylconic	High-pressure system in the Iberia peninsula which extends over the eastern Mediterranean
SOM1.3	Anticylconic	The Siberian anticyclone is extended over the Balkans
SOM2.3	Smooth	Smooth pressure field for the entire European continent
SOM3.3	Cyclonic	Low-pressure system of the Adriatic Sea and Italy
SOM4.3	Cyclonic	Extended low-pressure system over central Europe.
SOM5.3	Cyclonic	Low-pressure in northeastern Europe
SOM1.4	Anticylconic	Weak Azores high penetration in the eastern Mediterranean that favors the development of local flows
SOM2.4	Cyclonic	Deep low in the UK does not affect Southeastern Europe
SOM3.4	High – Low Combination	High-low combination over Western and Eastern Europe
SOM4.4	Cyclonic	Deep low-pressure system situated at North Sea
SOM5.4	Cyclonic	Low-pressure system, located at the north of Greece

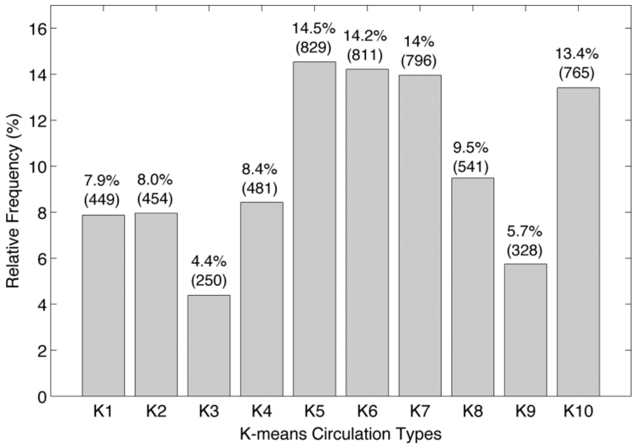

Fig. 2 Relative (and absolute) frequency of occurrence of the k-means classification circulation types

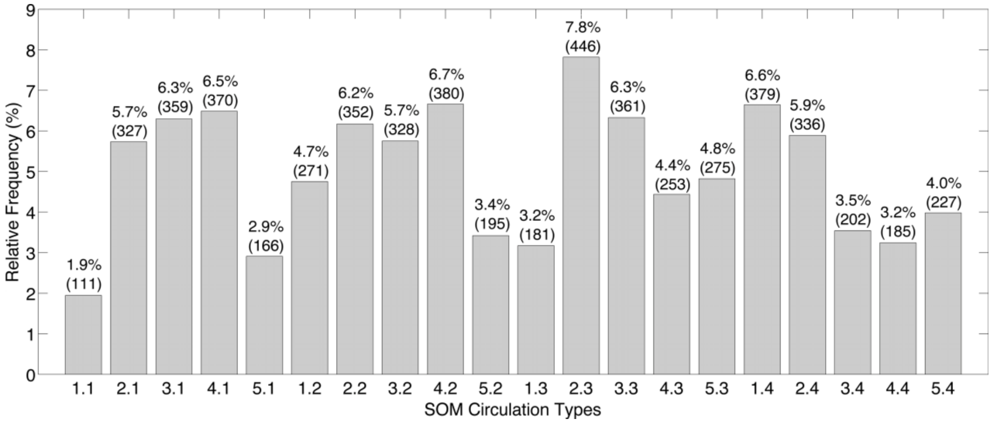

Fig. 4 Relative (and absolute) frequency of occurrence for the SOM classification circulation types

Table 3: Agreement in percent between the SOM and the k-means circulation patterns

	K1	**K2**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
SOM1.1	0.0	0.0	**76.6**	21.6	0.0	0.0	0.0	0.0	0.0	1.8
SOM2.1	0.0	0.0	1.5	33.0	0.0	0.0	4.0	0.0	0.0	61.5
SOM1.3	1.7	0.0	0.0	0.3	**50.1**	0.0	30.4	0.0	0.0	17.5
SOM1.4	0.0	0.0	0.0	0.3	39.2	1.6	3.2	0.5	2.4	52.7
SOM1.5	0.0	0.0	44.6	2.4	0.0	0.0	0.0	0.0	19.3	33.7
SOM2.1	0.0	0.0	0.4	**80.8**	0.0	0.0	17.7	0.0	0.0	1.1
SOM2.2	0.6	0.3	0.0	4.8	4.0	1.1	**88.9**	0.0	0.0	0.3
SOM2.3	22.0	1.5	0.0	0.0	**74.4**	0.0	1.5	0.0	0.6	0.0
SOM2.4	1.3	11.1	0.0	0.0	22.4	49.2	6.3	6.1	1.8	1.8
SOM2.5	0.0	0.0	2.1	0.0	1.5	0.0	0.0	10.3	**85.1**	1.0
SOM3.1	0.0	0.0	44.2	16.0	0.0	29.8	0.0	0.0	2.2	7.7
SOM3.2	0.0	0.0	0.2	15.9	1.3	30.7	17.3	0.0	0.0	34.5
SOM3.3	31.0	5.3	0.0	0.0	32.4	7.5	19.1	3.9	0.0	0.8
SOM3.4	44.3	52.6	0.0	0.0	0.0	0.0	0.0	3.2	0.0	0.0
SOM3.5	24.4	8.0	0.0	0.0	6.2	0.0	0.0	59.6	1.8	0.0
SOM4.1	1.8	0.0	0.0	0.0	4.7	41.2	0.0	25.3	10.0	16.9
SOM4.2	5.7	31.3	0.0	0.0	0.0	55.7	0.6	6.8	0.0	0.0
SOM4.3	9.4	14.9	0.0	3.5	0.0	10.9	61.4	0.0	0.0	0.0
SOM4.4	15.1	**51.4**	0.0	0.0	0.0	0.0	0.0	**33.5**	0.0	0.0
SOM4.5	0.0	0.9	0.0	0.0	0.0	13.7	0.0	56.8	28.6	0.0

B. *SOM circulation patterns*

The circulation patterns of the SOM classification are mapped according to the influence of the high and low-pressure systems (Fig. 3). The description of the relevant SOM circulation patterns is presented in terms of the most important atmospheric circulation characteristics in Table 2 and their relative frequencies of occurrence in Figure 4. In the lower left part of the map the patterns are mainly influenced by the existence of high-pressure systems in Europe, while the relative location of the low-pressure systems is the primary characteristic of the upper right part. This finding is in accordance with previous studies [14] and it is attributed to the inherent characteristic of the SOM method to self-organize. The nodes (neurons) exist in a continuum and enable the understanding of phases as well as the transitional nodes between phases [2].

C. *Comparison of the atmospheric circulation classifications*

The two classification schemes produce similar circulation types. The comparison of two classifications is presented in terms of examining the distribution of each SOM circulation type days to the k-means patterns (Table 3). Regarding the circulation types that are characterized from the existence of a low-pressure system in Europe, the SOM3.3 and the SOM3.4 types share common characteristics with K1 circulation type, differentiating in the relative position of the low-pressure system. Furthermore, 84.9% of the SOM4.4 days are classified as members of the K2 and K8 types, while the surface pressure distribution of the SOM3.5, SOM4.5 and the K8 types is depicted from the existence of a low-pressure system in northern Europe. The characteristic synoptic condition of the SOM2.5 is almost identical to the K9 type, resulting to a high agreement percentage (85.1%). For both classifications an increased number of days are classified into smooth pressure patterns with minimal pressure gradient in southeastern Europe. In detail, the days classified into the SOM1.3 and SOM2.3 patterns have high agreement percentages with the K5 circulation type (50.1% and 74.4% respectively), while the SOM2.4, SOM4.1 and SOM4.2 patterns are similar to the K6 circulation type. The SOM2.2 type is almost identical with the K7 pattern, with a total agreement of 88.9%. The two classification schemes provide more consistent results for the high-pressure system patterns. The SOM1.1 and SOM1.5 types, due to their cold period character, are mainly observed during March and are similar to the K3 type. Furthermore, high agreement percentage (80.8%) is observed between the SOM2.1 and the K4 types, which are also commonly observed during March. The synoptic situation for both SOM1.2 and SOM1.4 types share some common characteristics with the K10 pattern, where an anticyclone is located at the north of the Iberian Peninsula and in the British Isles. The similarity between the resulting patterns of the two classifications is further established from the high correspondence of their monthly frequency of occurrence.

IV. Conclusions

In this study two automated atmospheric circulation classification schemes are presented and examined for their ability to produce meaningful circulation types for the spring season in southeastern Europe. Both classifications, following the circulation-to-environment approach, can be used for relating the circulation types with regional or local scale meteorology and climatology. The k-means classification includes ten distinct types, while the SOM required more neurons to describe with discrete atmospheric states the daily MSLP distribution for the area and period under study. Both methods (k-means and SOM) are designed to achieve optimal distribution of objects (daily patterns) into the classes. The reason for reaching different result is that k-means can be trapped in local minima of the minimization function (reduction of within-type variance) while SOM is able to approach the global optimum. Meaningful relations are obtained in all cases. The correspondence of the two classifications is higher for the types where the high-pressure systems define the atmospheric circulation in the examined region. The SOM scheme has the ability to account for non-linear relationships and produce a map where synoptic states that are closely related are positioned in adjacent locations. In our case the high-pressure patterns are positioned in the lower left part of the map while the low-pressure patterns are located in the upper right part. Future work is proposed for developing a two-step classification scheme using both of the examined methods. The SOM can be used to decrease and reduce noise by producing a high number of atmospheric states which can be subsequently further grouped into a highly practical daily catalogue by applying k-means cluster analysis.

References

[1] B. Yarnal, *Synoptic Climatology in Environmental Analysis: A Primer*. London: Belhaven Press, 1993, 1st edition.

[2] C. S. Sheridan, and C. C. Lee, “The self-organizing map in synoptic climatological research”, *Prog. Phys. Geog.*, vol. 35, pp. 109-119, 2011.

[3] C. Beck, and A. Philipp, “Evaluation and comparison of circulation type classifications for the European domain”. *Phys Chem Earth, Parts A/B/C*, vol. 35 (9-12), pp. 374-387, 2010.

[4] A. Philipp, J. Bartholy, C. Beck, M. Erpicum, P. Esteban, X. Fettweis, R. Huth, P. James, S. Jourdain, F. Kreienkamp, T. Krennert, S. Lykoudis, S. C. Michaelides, K. Pianko-Kluczynska, P. Post, D. Rasilla Álvarez, R. Schiemann, A. Spekat, and F. S. Tymvios, “Cost733cat – A database of weather and circulation type classifications”. *Phys. Chem. Earth, Parts A/B/C*, vol. 35 (9-12), pp. 360-373, 2010.

[5] R. Huth, “An intercomparison of computer-assisted circulation classification methods”. *Int. J. Clim.*, vol. 16, pp. 893-922, 1996.

[6] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, and D. Joseph, “The NCEP/NCAR 40-year reanalysis project”. *Bull. Amer. Meteor. Soc.*, vol. 77, pp. 437-447, 1996.

[7] W. Enke, and A. Spekat, 1997. “Downscaling climate model outputs into local and regional weather elements by classification and regression”. *Clim. Res.*, vol. 8, pp. 195–207, 1997.

[8] G. Karvounis, D. Deligiorgi, and K. Philippopoulos, “On the sensitivity of AERMOD to surface parameters under various anemological conditions”. In Proceedings of the 11th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, pp. 43-47, 2007.

[9] D. S. Wilks, 2011. *Statistical Methods in the Atmospheric Sciences*. Amsterdam: Academic Press, 2011, 3rd edition.

[10] T. Kohonen, *Self-Organization and Associative Memory*. New York: Springer-Verlag, 1984, 3rd edition.

[11] T. Kohonen, T., *Self-Organizing Maps.* New York: Springer-Verlag, 2001, 3rd edition

[12] S. C. Michaelides, F. Liassidou, and C. N. Schizas, "Synoptic classification and establishment of analogues with artificial neural networks". *Pure Appl. Geophys.*, vol. 164, pp. 1347-1364, 2007.

[13] Y. Liu, and H. R. Weisberg, "A Review of Self-Organizing Map Applications in Meteorology and Oceanography", In *Self Organizing Maps - Applications and Novel Algorithm Design*, Rijeka: InTech Publishers, 2011.

[14] B. C. Hewitson, and R. G. Crane, "Self-organizing maps: applications to synoptic climatology". *Clim. Res.*, vol. 22, pp. 13-26, 2002.

[15] S. Haykin, *Neural Networks and Learning Machines*. Upper Saddle River: Pearson Education Inc., 2009, 3rd edition.

[16] T. Cavazos, "Using Self-Organizing Maps to Investigate Extreme Climate Events: An Application to Wintertime Precipitation in the Balkans". *J. Climate*, vol. 13, pp. 1718-1732, 2000.