On the Use of Self-Organizing Maps in Detecting Teleconnections. Part I: Spatial Aspects

Jan Stryhal (stryhal@ufa.cas.cz)
Institute of Atmospheric Physics, Czech Academy of Sciences

Romana Beranová
Institute of Atmospheric Physics, Czech Academy of Sciences

Radan Huth
Institute of Atmospheric Physics, Czech Academy of Sciences

Research Article

Keywords: SOMs, modes of variability, teleconnections, atmospheric circulation classifications

DOI: https://doi.org/10.21203/rs.3.rs-719193/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Self-organizing maps (SOMs) represent a popular classification tool, output of which has been used to
link typical synoptic-scale circulation patterns to large-scale teleconnections, or modes of low-frequency
circulation variability. However, recently there have been attempts to interpret SOM output directly as a
continuum of teleconnections. In the present paper, we provide a theoretical study dealing with how
predefined idealized modes of variability project onto SOM arrays. Three orthogonal modes are defined
and their various linear combinations used to generate datasets of known structure. Multiple variants of
SOMs that differ in their initialization are generated for SOMs of various shapes and sizes. The results
show that how a mode projects on a SOM array is sensitive not only to data structure, but also to various
SOM parameters. The leading mode of variability projects rather strongly on SOMs if its explained
variance is markedly higher than that of the second-order mode; the remaining modes project
considerably more weakly, and all modes tend to blend when their explained variance is close to each
other, which leads to underrepresentation of some phases of modes and/or combinations of modes
among the SOM patterns. Furthermore, non-linear features appear in SOM arrays even if the underlying
modes of variability are strictly linear. The findings point to limitations in the applicability of SOMs in the
research of teleconnections.

1 Introduction

Self-organizing maps (SOMs; Kohonen 2001; Hewitson and Crane 2002; Sheridan and Lee 2011) have
recently become one of the most popular methods in synoptic climatology. Their utilization have been
manifold but the application towards classifying circulation patterns—typically mean sea level pressure
(SLP) or geopotential height fields—has been most common. Similarly to other classification methods,
such as the Jenkinson-Collison and Lund methods, T-mode principal component analysis (TPCA), and
cluster analysis—see for example Huth et al. (2008) and Philipp et al. (2016) for detailed reviews of the
topic of circulation typing—SOMs enable approximating the broad variability in circulation fields by only a
small number of representative nodes, or circulation types. Additionally, SOMs offer the possibility to
organize these types into an array in which the distance between individual types approximates the
(typically) Euclidean distance between them in the original multidimensional space; that is, SOMs are not
only able to classify data but also represent their topological structure. This latter characteristic of SOMs
is arguably the main reason of their popularity, since it eases visualization and, consequently, also
interpretation of results.

An alternative utilization of SOMs, proposed originally by Johnson et al. (2008) and used since then in
several spin-off studies (Johnson 2013; Chang and Johnson 2015; Rousi et al. 2015, 2020; Yuan et al.
2015) consists in a different interpretation of the SOM output. Instead of interpreting individual types of a
SOM array as “stand-alone” typical geophysical fields, one interprets the whole SOM array as a
continuum of teleconnections. Compared to typical fields, teleconnections have predominantly been
interpreted as patterns expressing statistical relations between climatic variables at distant locations
(Ångström 1935; Wallace and Gutzler 1981; Hannachi et al. 2017) and as such investigated via local-
based indices, one-point correlation or regression maps (e.g., Wallace and Gutzler 1981; Corti et al. 1997; Franzke and Feldstein 2005) and S-mode PCA (SPCA; e.g., Horel 1981; Barnston and Livezey 1987; Corti et al. 1997; Feldstein 2000; Jung et al. 2003; Beranová and Huth 2008; Davini and Cagnazzo 2014; Comas-Bru and Hernández 2018), the latter method having arguably become the most popular one.

The SPCA decomposes data transformed into a similarity (typically correlation or covariance) matrix whose values represent relations between time series of a typically gridded input meteorological variable. (Note that TPCA, mentioned earlier among classification methods, also decomposes a similarity matrix; however, this matrix represents similarity between pairs of geophysical fields, such as pattern correlation, instead of relations between pairs of variables, which yields qualitatively different results. A reader is referred to Preisendorfer (1988) for detailed information on various modes of decomposition and Compagnucci and Richman (2008) for a comparison of SPCA and TPCA in climatological context). The SPCA results in a set of new variables produced as linear combinations of input variables, which are mutually orthogonal, maximize the explained variance, and have traditionally been referred to as modes of variability (Barnston and Livezey 1987). Since it was shown by Horel (1981) that the leading modes of variability obtained by rotated SPCA well reproduce the teleconnection patterns found earlier by Wallace and Gutzler (1981), the terms teleconnections and modes of variability have been used more or less interchangeably. Although some instances can be found in literature in which the term mode of variability refers to variability patterns in general, including circulation types and/or transitions between them (e.g., Reusch et al. 2005; Lennard and Hegerl 2015; Risbey et al. 2015; Cannon 2020), here we use the term to address spatial variables or functions of variability, such as PC loading vectors or point-correlation maps. On the other hand, the term teleconnection is here understood more loosely such that it accommodates other definitions than that of a spatial function.

One particular motivation behind the application of SOMs (and less often also cluster analysis) to analyses of teleconnections instead of SPCA and other linear methods is to study their non-linear features, such as differences in the spatial structure teleconnections may have in their opposite phases, which cannot be captured by the abovementioned traditional methods due to their imposed linearity and/or orthogonality (Risbey et al. 2015). A typical application of SOMs is to study the extent to which occurrences of typical atmospheric patterns are conditioned by intensity and phase of leading modes of variability (e.g., Reusch et al. 2007; Verdon-Kidd and Kiem 2009; Johnson and Feldstein 2010; Reusch 2010; Gao et al. 2019). Because SOMs preserve topology of data, the opposite phases of leading modes of variability tend to strongly project on specific regions of a SOM array, for example on its two opposite corners. In the continuum perspective of teleconnections, such regions are then interpreted to form positive and negative continua of the mode’s “members” (e.g. Johnson et al. 2008; Chang and Johnson 2015), or “flavors” (Johnson 2013; Verdon-Kidd 2018; Rousi et al. 2020), which are in turn applied to describe the mode’s temporal changes (in phase and intensity) and/or changes in its spatial patterns by means of analyzing shifts in the frequency of occurrence among the pertinent SOM types. Note that compared to the traditional “discrete modal perspective” (Chang and Johnson 2015), in the continuum perspective a teleconnection does not have one fixed spatial pattern, although its generalized positive and negative patterns can be retrieved by weighted averaging of pertinent SOM types (e.g., Chang and
The definition of a teleconnection is in the continuum perspective shifted from a statistical link expressing co-variability among distant locations to a less constrained “low-frequency anomaly pattern” (Yuan et al. 2015), which shows a particular temporal realization of such a statistical link rather than the link itself, although in reality typical daily anomalies have been interpreted as teleconnections in this context more often than low-pass filtered data (e.g., Johnson et al. 2008; Chang and Johnson 2015; Rousi et al. 2015; Yuan et al. 2015).

A typical way to delimit individual “canonical” teleconnections (leading modes of variability defined by linear methods) within a SOM teleconnection continuum consists in projecting modes of variability (principal components, PCs) on SOM arrays and comparing similarities between the two representations of teleconnections in the spatial (e.g., Yuan et al. 2015, Rousi et al. 2020) or temporal (e.g., Johnson et al. 2008) domain, which links the two perspectives of teleconnections together. Theoretical studies on how modes of variability project on SOM arrays seem, therefore, particularly relevant to the continuum perspective. It is, however, an area of research that has so far received only very little attention.

Jiang et al. (2013) have shown that influences of various teleconnection indices on the frequency, intensity and airmass characteristics of typical geopotential height patterns are complex and can counteract one another; consequently, attributing changes in these patterns to one particular teleconnection seems far from trivial. Most of previous studies indicate that the leading variability mode tends to project strongly on types around two opposite corners of a SOM array, and the very corners are often representing positive and negative polarities of the mode. According to Lennard and Hegerl (2015), the two diagonals of a SOM array are analogous to first two principal components and thus to the leading two teleconnections. This somewhat contradicts the suggestion of Reusch (2010) that if the leading mode of variability projects on one diagonal of a SOM array, the second-order mode may or may not project on the other diagonal. Furthermore, Reusch (2010) suggests that differences in the length of two diagonals on which the leading modes of variability do project—measured in the phase space that approximates Euclidean distances among data (refer to Sect. 2, Sammon mapping for more detail)—stem from data nonlinearities. Only very limited attention has been given to the third-order mode, which is likely due to its very limited and scattered signal in two-dimensional SOMs. This means that all lower-order modes of variability still project on SOM arrays and influence any results, but are ignored in the interpretation.

In order to contribute to the understanding of the continuum perspective of teleconnections as well as the links and differences between different perspectives of teleconnections, we present a two-part study that deals with spatial (Part I) and temporal (Part II) aspects of using SOMs as a tool for studying teleconnections. In the present Part I of the study, we analyze how three predefined modes of variability used to construct various synthetic datasets project on SOMs trained with a wide range of parameters. The paper is organized as follows: in Sect. 2, we describe the datasets and methodologies of SOMs and Sammon mapping (Sammon 1969), which is used to visualize the results; in Sect. 3, we analyze the links of SOM arrays to modes of variability and to data structure, and in Sect. 4 we discuss the results and summarize main implications for the continuum perspective and nonlinearity of teleconnections.
2 Data And Methods

2.1 Synthetic data

In the study, multiple synthetic datasets with predefined—and thus known—structure are used to train SOMs. The datasets are built using three modes of circulation variability that describe pattern-to-pattern changes in zonal (Z), meridional (M), and circular (C) flow. The three modes are defined as mutually orthogonal and of unit length, in order that contribution of individual modes to the total variability can be fully controlled by generated indices (amplitudes). In order to avoid calculation with very small numbers and to make the resulting generated fields easier to read, all three patterns are scaled up by the same factor, the size of which is not relevant, but explains the larger isoline intervals in the spatial patterns of modes (Fig. 1). The patterns can be interpreted as anomalies of an unspecified circulation variable that represent the modes in their positive phase and intensity of +1.0.

Each dataset comprises 1,000 circulation fields; each field consisting of 20×20 grid points is a certain linear combination of the three patterns, whereby each mode is multiplied by a scalar index. The index therefore represents the intensity of a particular mode at a particular time instant. For each mode, the indices are drawn from normal distributions, the mean of which is zero and variance is a certain fraction (explained variance hereafter) of the total variance $\sigma^2 = 1$. A total of twelve combinations of explained variances (EV), and thus datasets, are predefined (Table 1) such that $\sigma^2_Z \geq \sigma^2_M \geq \sigma^2_C$. For example, in the dataset “Z70-M20-C10”, 70% of the total variability is in the zonal flow, while the contribution of the meridional and circular flow is 20% and 10%, respectively. We also define several simple “bivariate” datasets in which the third mode is “turned off” (e.g., Z70-M30). Note that datasets in which another mode than the zonal mode was the strongest (e.g., Z30-M70) were not included in the study as they did not lead to any new findings.

| Mode | Explained variance (%) |
|------|------------------------|
| Z    | 90 70 70 60 60 50 50 50 40 40 33 |
| M    | 10 30 20 30 20 50 40 30 25 40 30 33 |
| C    | 0  0 10 10 20 0 10 20 25 20 30 33 |

2.2 Training of self-organizing maps and quality metrics

Here, we describe the methodology of training, evaluating and selecting SOMs. The methodology is indicated in Fig. 2; choices of parameters along with their effects of results are further discussed in Sect. 4.2. A reader is referred to Kohonen (2001) and Sheridan and Lee (2011) for a more detailed description.
of SOMs. Note that in the paper we use the term *type* instead of *node*, in order to prevent confusion between the terms *node* and *mode*.

The training of a SOM constitutes an iterative and unsupervised process. At the start, \( n \) starting types, or vectors (in our case, generated circulation fields) are randomly selected from a dataset to initialize the SOM array; \( n \) also equals the resultant number of circulation types. Subsequently, each of the daily patterns is step by step and in many iterations projected on the array and the position of types that are close (in terms of Euclidean distance) to the pattern is adjusted. The adjustment depends on three parameters that have to be specified in advance. First, the neighborhood radius (further referred to as radius for short) defines the size of the affected area around the projected data point that is adjusted. Second, the neighborhood function defines the mode of the adjustment, and third, the learning rate specifies how much the types within the affected area are attracted to the projected data point. The Gaussian neighborhood function was chosen here as it led to SOMs with somewhat better topology than the commonly used bubble function. The value of the remaining two parameters is typically chosen to decrease in each iteration step, which allows for capturing the overall topological structure of the dataset first and then fine tuning the resulting classification during the later iteration steps. All choices of parameters are discussed in Sect. 4.2.

The “Kohonen” R package (Wehrens and Buydens 2007; Wehrens and Kruisselbrink 2018) was used to train SOMs. For each dataset, SOMs of 13 different shapes (number of rows × number of columns) of arrays were calculated: 2×3, 2×4, 3×3, 3×4, 3×5, 4×4, 4×5, 4×6, 5×5, 5×6, 5×7, 6×6, and 6×7, resulting in classifications into between six (2×3 SOMs) and forty-two (6×7) types. The number of iterations was set to 1,000; a higher value did not lead to higher quality SOMs, likely due to the relatively simple structure of data. The initial value of the learning rate was set to between 0.04 and 0.12 for smallest and largest SOMs, respectively, and to linearly decrease to 0.01 in all cases. The radius parameter was initialized as the 66th percentile of type-to-type Euclidean distances, calculated from coordinates of types within the SOM array (i.e., row and column), and was set to linearly decrease over the 1,000 iterations to a specific terminal radius value. Since Jiang et al. (2015) and Gibson et al. (2017) pointed to the sensitivity of SOMs to this particular parameter, and this sensitivity was also very apparent in our preliminary analyses, multiple values of the terminal radius were tested for each SOM shape – dataset combination. This complex issue will be described in detail in Sect. 3.1.

Two kinds of quality metrics are typically used to assess SOMs—quantization error and topological error (Kohonen 2001). Although their exact specification varies among studies, their purpose does not: While the quantization error shows the degree of data separation into clusters—and can be therefore understood as a measure of quality of a classification, the topological error shows the degree to which the organization of types within a SOM array captures the structure of classified data. Since both metrics depend on predefined parameters, they are typically used to select a single optimum SOM of multiple ones trained. Here, quantization error of a SOM is calculated as the Euclidean distance between each circulation field and its closest type, averaged over all 1,000 classified fields. The topological error
represents the percentage of circulation fields whose two closest types do not lie next to each other in the SOM array.

2.3 Sammon mapping

Sammon mapping is a non-linear dimension reduction method that maps multivariate data points (here, 400-dimensional circulation fields) to lower- (here, two-) dimensional space such that the data structure is approximately preserved; that is, inter-point distances (typically, Euclidean) in the original and the projection Sammon space correspond (Sammon, 1969). This is achieved by iteratively looking for a solution that minimizes the error function

\[
E = \frac{1}{\Sigma_{i<j} d_{ij}^*} \sum_{i<j} \left( \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \right),
\]

where \(d_{ij}^*\) denotes the Euclidean distance between \(i\)th and \(j\)th circulation fields in the original space and \(d_{ij}\) denotes their Euclidean distance in the projection (Sammon) space.

The purpose of Sammon mapping in the study is twofold. First, it is used as an additional quality measure to assess the topological structure of trained SOMs. Second, it is utilized as a means of analyzing the links between the generated datasets, the position of SOM types within the data, and the underlying modes of variability (structure of data). The following approach is used: First, a Sammon map is calculated for each dataset; the function “sammon” of the R package MASS (Venables and Ripley 2002) is used to this end. To plot a SOM onto a Sammon map, each type is projected onto the map by finding the point in the Sammon phase space (at a step of 1×1 unit) for which the projection error \(F\) (Eq. 2) is the smallest:

\[
F = \frac{1}{\Sigma_t d_{it}^*} \sum_t (d_{it}^* - d_{it})^2 \frac{d_{it}^*}{d_{it}^*},
\]

where \(d_{it}^*\) (\(d_{it}\)) denotes the Euclidean distance between the \(i\)th circulation field and type \(t\) in the original (projection) space. The same method can be used to project any circulation field, which we utilized for projection of multiples of patterns of variability modes to analyze how the modes are oriented within the maps.

In Fig. 3, Sammon map of one dataset is shown together with one SOM and its projection onto the Sammon map. Note that in the Sammon map in Fig. 3a, relative to the orthogonal array of types in Fig. 3b, the array is slightly deformed such that the distance between individual SOM types approximates the Euclidean distance between the spatial patterns of the types in the original space.

3 Results
3.1 Training and selection of SOMs

Although calculation of a SOM constitutes an objective process that does not require any interventions on the part of the researcher, multiple parameters affect the eventual properties of a SOM, many of which require subjective decisions to be made a priori. Above all, the number of types (size of the SOM array) determines the degree of generalization, and the asymmetry of the SOM array (difference between the numbers of columns and rows) potentially affects the ability of a SOM to represent datasets with various structures (differences in the number and explained variance of individual modes of variability). Furthermore, previous research has shown (see, e.g., Rousi et al. 2015) that when selecting a SOM, one typically has to balance between the quality of classification (as expressed, for example, by the quantization error) and the quality of SOM topology—the goal of the study being the decisive factor. The aim of the present study required selecting SOMs with good topological structure, however, without compromising quality of classifications.

First, for each of the twelve datasets and thirteen SOM shapes, a total of 320 (200) variants of SOMs were trained for SOMs up to (larger than) 4×4; the difference stems from markedly longer computation time of larger SOMs. The variants differ in their initialization (20 variants were calculated for each dataset – SOM shape combination) and also in the terminal radius: For smaller (larger) SOMs, values between 0.0 and 1.5 (0.9 and 1.8) were tested, in both cases at a step of 0.1. For small SOMs (especially if combined with datasets with simple and clear structure, that is, two modes of variability of dissimilar strength), best results (in terms of both quantization and topology) are typically achieved for low terminal radius. For larger SOMs, one has to choose larger values of the terminal radius to avoid deterioration of the topological structure (except for the simplest data).

Second, one SOM was eventually chosen for each dataset and SOM shape (therefore, a total of 156 SOMs were analyzed), by selecting the SOM with the smallest quantization error among the SOMs in whose projections onto the Sammon space folding of arrays (crossing of types-to-type links) did not occur. In case multiple SOMs met these conditions, the one with the smallest topological error was selected. In Fig. 4, quantization and topological errors and their dependence on terminal radius, SOM shape and dataset structure are shown for selected cases.

Not surprisingly, SOMs with lower quantization error are those with larger arrays, which better explain data variations. Furthermore, the overall larger error in the case of datasets with variance spread more evenly among the modes is due to overall larger inter-pattern Euclidean distance in these datasets (not shown). Of particular note is a marked drop in the error when the terminal radius drops below 1.0—but no further decrease is observed beyond. For future studies, experimenting with the terminal radius close to the threshold of 1.0 may, therefore, suffice. However, this may not be the case if large SOMs were combined with data of little structure (similar variance of modes). Here, for example, terminal radius of 1.4–1.6 was the minimum that led to SOMs with reasonable topological structure for arrays of 30–42 types combined with datasets of little structure (not shown here).
The link between the topological error and SOM parameters is somewhat less clear, especially for terminal radius above 1.0: the link can be indifferent, directly as well as indirectly proportional, depending on the particular dataset – SOM shape combination. A sharp increase of the error can be, nevertheless, seen in nearly all of the 156 combinations tested when the terminal radius drops below 1.0, except for a few cases concerning the simplest datasets. The relatively high spread of the error, especially when terminal radius is set under 1.0, highlights the sensitivity of SOMs to its initialization and, thus, the importance of calculating multiple variants of SOMs to avoid the risk of achieving only suboptimal local optima of classification. Despite comparable quantization error for all variants differing only in their initialization, classifications with identical parameters can lead to two or even more local optima of the topological structure—note for example the three local optima for Z50-M40-C10 dataset, 3×4 array, and radius of 0.7 in Fig. 3b. The Sammon maps of three SOMs that represent the three optima are shown in Fig. 5 (notice the SOM in Fig. 5a has markedly better topology compared to the other SOMs). Multiple local optima in SOM topology are often found for oblong SOMs of data that have equal or close to equal strength of two modes (zonal and meridional, or meridional and circular). Therefore, an assessment of data structure by PCA may represent a useful tool prior to SOM classification in terms of selection of some SOM parameters.

3.2 SOM topology and the modes of variability

Here, we present how SOM types are organized within an array according to their spatial similarity to modes of variability. Sammon mapping is utilized to visualize the topological structure of both data and SOM arrays, and Pearson's pattern correlation is used to measure the spatial similarity between patterns of types and modes. Additionally, several spatial patterns representative of various intensities and both phases of the modes are projected onto Sammon maps to obtain a generalized non-linear projection of the underlying linear modes.

First, the simplest cases are presented, in which data variations are determined only by two modes of variability (zonal and meridional). Second, SOMs of datasets generated from all three modes of dissimilar strength are presented. Last, we focus on the specific behavior of SOMs that occurs if all three modes have nonzero variance and equal strength of two or all three modes. In each case, we first focus on how the modes project on the Sammon maps of data, and then analyze the relation between the projections of modes and SOMs.

3.2.1 Two modes of variability

In the three simplest datasets (Z90-M10, Z70-M30, Z50-M50) we test the organization of SOM types relative to the two underlying variability modes, and mainly whether projections of the two modes are orthogonal to each other in the nonlinear Sammon projection, and whether the modes can be linked to the diagonals of SOM arrays, as suggested before by Lennard and Hegerl (2015).
The generated circulation fields are clearly organized in the Sammon map according to their spatial similarity to patterns of variability modes. In all three datasets, projections of the two modes are orthogonal to each other. In Fig. 6, this is apparent both indirectly from correlations between data and modes and directly from projections of idealized patterns representing the modes onto the Sammon maps. Furthermore, projections of these idealized patterns are linear. In the two datasets of unequal strength of modes, the modes project along the two axes of the Sammon space (Fig. 6a,b). In the special case of their equal strength (Fig. 6c), the modes remain orthogonal after projection, but do not project on the horizontal and vertical axes of the Sammon map. This is likely due to the fact that the horizontal axis is by default oriented in the direction of maximum explained variance (in our case, this is true for all datasets with one leading variability mode), which is unstable for data of spherical (rather than elliptic) shape. However, this does not affect our analysis, as the relative links between the projected objects do not change if the Sammon map is linearly transformed (e.g., rotated).

In Fig. 6, three SOMs (4×4, 4×5, and 4×6) are projected onto each of the three Sammon maps (data sets) to illustrate the link between their topology and projection of modes. The diagonals of the square 4×4 SOMs project on the axes of the Sammon space in the case of Z90-M10 and Z70-M30 datasets, that is, opposite corners of SOM arrays depict opposite phases of modes in their rather extreme state. In the case of Z50-M50 (Fig. 6c), however, the modes strongly project on the (theoretical) center row and column of the array, with the corner types representing fields with about equal contribution of both modes.

In general, with an increasing elongation of SOM arrays, its center row tends to be aligned with the leading mode rather than a diagonal, but a wide range of results was obtained in different datasets. Consequently, the leading mode may strongly project on two opposite corners (as in 4×5 and 4×6 SOMs of Z90-M10 data), on only one corner (4×5 SOM of Z70-M30 data) or none at all (4×6 SOM of Z70-M30 data). The weaker mode does not project most strongly on the remaining two corners but rather on the types next to and second next to the corners in 4×5 and 4×6 SOMs, respectively. The corner types themselves represent more complex fields with non-negligible contribution of both modes. The results obtained for SOMs of other shapes are similar to those presented above: Modes project well on the diagonals (center row and column in the case of Z50-M50) of square SOM arrays regardless of their size, while only the leading mode occasionally projects on a diagonal in oblong SOMs.

### 3.2.2 Three modes of variability of different strength

Four datasets composed of three modes of variability of different strength were analyzed in the study (Z70-M20-C10, Z60-M30-C10, Z50-M40-C10, and Z50-M30-C20). Relative to the bivariate datasets, shifting 10–20% of total variability to the third-order circular mode does not markedly affect the projection of the first two modes onto the Sammon map: the zonal (meridional) modes project on the horizontal (vertical) axes of the projection plane (Fig. 7). When projecting the circular mode on the Sammon map, its stronger realizations (i.e., ±1×C and beyond) appear along the vertical axis (i.e., orthogonally to the leading mode, that is, more or less parallel to the second-order mode). However, fields with such strong contribution of circular flow are generated only very rarely in these four datasets, the maximum generated indices of the mode being—in absolute values—1.23 (1.49) in C10 (C20) datasets,
and the 95th percentile being only about 0.6. On the contrary, the ± 0.5×C patterns project in all four cases very close to the center of the Sammon map. In general, fields highly correlated with the circular mode are scattered across the whole Sammon map, but their highest density is close to the center. In the center, fields with only weak circulation (“turned-off” zonal and meridional modes) occur, and therefore, only a small contribution of the circular mode leads to relatively high pattern correlation between its pattern and data fields.

The orientation of SOM arrays relative to modes of variability and the shape of the arrays also resemble those obtained for bivariate datasets: the two leading modes project onto diagonals in all square SOMs except those of the Z50-M40-C10 dataset. On the other hand, the two leading modes project along the central row and column in most SOMs that are strongly elongated (i.e., those with two more columns than rows), the corner types then comprising linear combinations of the two leading modes. The orientation of SOMs with one more column than rows varies between the typical orientation of square and strongly elongated SOMs, depending on dataset and SOM parameters (number of types, terminal radius). The corner types of these slightly elongated SOMs can be either linear combinations of two or three modes or strong projections of one particular mode. In the latter case, patterns in opposite corners typically are not mirror images of each other (see Fig. 8a,b); such a situation is further referred to as “non-mirrored patterns”.

Consistent with the projection of the ± 0.5×C patterns close to the center of Sammon maps, also types resembling the pattern of this mode occur close to the center of SOM arrays. Note, however, that the proximity to the center is in terms of the Sammon map rather than SOM array coordinates. This is apparent in types [1,3] in Fig. 8b, and [4,3] and [4,4] in Fig. 8c, which—despite their location at the edge of SOM arrays—lie relatively close to the center of their respective Sammon maps, due to a rather deformed shape of the SOM array in the former case (Fig. 7b) and rather small array (consequence of a higher terminal radius necessary to produce a SOM with good topological structure) in the latter (Fig. 7c). However, the occurrence of types resembling the circular mode in the center does not mean that other types do not show signs of circular flow; rather the opposite holds, which is apparent from the spatial patterns of types (Fig. 8).

3.2.3 Three modes of variability of equal strength

One example of equal strength of modes was presented above, regarding the bivariate Z50-M50 dataset. Additional five datasets were defined in which the first- and second-order modes (Z40-M40-C20), the second- and third-order modes (Z60-M20-C20, Z50-M25-C25, and Z40-M30-C30), and all three modes (Z33-M33-C33) have equal strength.

Both the Sammon map and projection of modes in the case of Z40-M40-C20 data (not shown) resemble those of Z50-M40-C10 (Fig. 7b), except the leading two modes are not aligned with the Sammon map axes, similarly to Z50-M50 (Fig. 6c).
The three cases with equal strength of the meridional and circular modes are presented in Fig. 9, including Sammon maps of 4×4, 4×5 and 4×6 SOMs—the spatial patterns of the latter are shown in Fig. 8d–f. In all three cases, the leading mode is approximately aligned with the horizontal axis of the Sammon map. However, despite the equal strength of the two remaining modes, the three cases differ in how these two modes project onto the Sammon map of data as well as onto the SOMs.

In the case of Z60-M20-C20 (Fig. 9a), both modes project onto the vertical axis of the Sammon map, such that their positive (negative) values occur mostly above (below) the horizontal axis. A similar distribution is also apparent in patterns of SOM types (Fig. 8d): excluding the clearly zonal types, the upper two rows tend to capture southerly (M+) and cyclonic (C+) flow, and their combinations, while northerly (M-) and anticyclonic (C-) flow prevails in the bottom two rows. Consequently, the SOM selectively and clearly underrepresents concurrence of phases M + C- and M-C+ (in general, types with circular features in the right-hand side of their patterns). Additionally, the two types that project onto the vertical axis of the Sammon map ([1,3] and [4,4]) have non-mirrored patterns, the former resembling M+, while the latter being a linear combination of M- and C-. This illustrates the fact that the position of types within an array and their distance from the center of its Sammon map are functions of Euclidean distance and do not necessarily have a link to the position of anomalies, at least not to the same degree as in the case of pattern correlation.

In the case of the Z50-M25-C25 dataset (Fig. 9b), the three modes project onto the Sammon map similarly to Z50-M30-C20 (Fig. 7c), with the zonal and meridional modes aligned with the two axes and the circular mode scattered across the map with increased density of high correlations in the center of the map. The SOM is in this case deformed such that its upper left corner (see type [1,1] in Fig. 8e) and the center of the bottom row (type [4,4]) lie close to this area of increased data density and represent both phases of the circular mode. On the other hand, strong projections of the meridional mode occur in types that lie farthest from the horizontal axis in the Sammon map (types [1,3] and [4,6]). Thus, while two corners of the SOM that lie close to the horizontal axis show opposite phases of the leading mode, the remaining two corners show one phase of one of the remaining two modes each.

The Sammon map of the Z40-M30-C30 dataset (Fig. 9c) reveals that in this case it is the circular mode that projects orthogonally to the leading zonal mode. The projection of the meridional mode on the vertical axis is somewhat less clear than in Z60-M20-C20, but still apparent; note, however, that it does not have a clear representation among the types (Fig. 8f). Nonetheless, the northerly (southerly) component of flow does project on most types in the upper (bottom) half of the SOM. Given the alignment of the zonal and circular modes with the central row and column of SOMs (in the case of this particular dataset), it matters whether the number of rows and columns of a SOM is odd or even; the odd number causing the center row/column to clearly represent the spectrum of realizations of one particular mode, the remaining types being linear combinations of (typically) all three modes (not shown). However, this particular alignment occurs rarely; here, its most cases can be linked to three datasets (Z50-M40-C10, Z50-M30-C20, and Z40-M30-C30) and also to most of small strongly elongated (2×4, 3×5) SOMs.
4 Discussion And Concluding Remarks

4.1 On data

In research on methodology to study spatial-temporal variability of atmospheric circulation, synthetic datasets play an essential role. The structure of such datasets is a very important factor since it constrains the results to a large extent. This seems to be of particular importance if one compares the two alternative perspectives of teleconnections, that is, the spatial patterns of temporal co-variability on one hand and the regime-based approach that sees teleconnections as typical anomaly patterns on the other.

In the two previous studies that directly compared the skill of SPCA and SOMs in extracting known patterns (modes) from synthetic datasets (Reusch et al. 2005; Liu et al. 2006), most of datasets were generated by repeating predefined patterns, which led to datasets consisting of a few distinct clusters. These clusters were represented well by SOMs in the latter study if there were at least as many SOM types as clusters; in the former study, considerably more types were necessary to avoid blending of clusters. SPCA, on the other hand, was in most cases unable to capture the spatial structure of the predefined modes, even though the data were strictly linear. Datasets generated in such a way seem beneficial to study methods and their ability to recognize such clusters; however, the application of SPCA on such datasets is inherently meaningless, since it is not designed to recognize clusters in realizations, linear or nonlinear, unlike TPCA (Compagnucci and Richman 2008). This, however, does not mean that SPCA is useless in the continuum perspective, since the principal components are typically projected on SOM arrays to discern the relevant types from those related to other teleconnections (e.g., Johnson et al. 2008; Rousi et al. 2020; Yuan et al. 2015).

Consequently, to obtain data that could be used to compare different perspectives on teleconnections and avoid biased results, we construct synthetic datasets in a way that can be understood as reversed SPCA—that is, constructing data from predefined orthonormal vectors (loadings) that carry the information on how grid points co-vary and times series of amplitudes randomly drawn from normal distributions that carry the variability itself. A resulting dataset forms a continuum of circulation fields rather than a few distinct clusters, which better approximates the structure of real circulation—see, e.g., Fig. 7 in Philipp et al. (2016). More importantly, datasets defined in such a way can still be successfully classified and do not limit the applicability of SOMs in any way, but they can also be linked to underlying modes of variability. Therefore, they make it possible to study both perspectives of teleconnections in parallel, in our case by means of visualizing both SOMs and the underlying modes of variability in one projection plane, for which the nonlinear Sammon mapping is utilized.

We did not add noise to the generated datasets as according to the results of Reusch et al. (2005) and Liu et al. (2006) we did not believe that doing so would be beneficial. Furthermore, our data is generated as white noise. Based on results in Part II of the study, where generated and real-world data are compared, we do not think that generating auto-correlated data would change results of the present study. Last,
results by Liu et al. (2006; see Fig. 9 there) clearly show that linear methods blend those modes whose spatial patterns are correlated. Therefore, the three modes were defined such that they are mutually uncorrelated. We also defined the modes to be mutually equidistant, since one can hypothesize that SOMs may also tend to blend modes that are relatively close to one another, in terms of the distance metric.

All in all, even though our data are more complex than data used in previous similar studies, they are still relatively simplistic compared to real-world data. In Part II, additional datasets will be generated from modes extracted from reanalysis data to assess whether findings based on simplistic datasets hold in the real world.

### 4.2 On methodology

Beside the datasets, two other specific issues seem particularly important for our study due to its aim at methodological aspects. First, any results are constrained by the quality of SOMs, while the assessment of quality is to a large extent goal-specific and a result of choice of multiple parameters (e.g., Gibson et al. 2017). Previous studies provide only limited guidelines regarding these choices, and in many cases the parameters were selected randomly or not mentioned at all (Sheridan and Lee 2011). The second issue regards the application of Sammon mapping beyond a mere tool for validation, toward providing a visual link between data, modes and SOMs.

The choice of parameters was indicated in Fig. 2 and Sect. 2.2. The parameters were chosen experimentally based on an initial sensitivity study, with the exception of the algorithm type and grid topology, which were set to sequential and rectangular, respectively. Kohonen (2013) suggests that the batch algorithm be used instead of the sequential one, as it is faster and safer. However, it is not clear to what extent this recommendation relates to the typical utilization of SOMs in synoptic climatology consisting in defining relatively very small SOMs used for classification rather than for exploratory data projection. According to Jiang et al. (2015), the sequential algorithm was used by the majority of research in synoptic climatology.

The training was initialized using randomly selected data points. We did not choose the alternative initialization from leading two principal components to assure that the relative orientation of SOMs and predefined modes in the Sammon space was not biased by this step. Since random initialization was quite sensitive to initial conditions, 20 variants of SOM were calculated for each combination of the remaining parameters. Furthermore, the commonly used Euclidean distance was chosen as the metric of similarity, as other metrics provided by the package (e.g., Manhattan, sum-of-squares) did not lead to higher quality SOMs. Last, we kept the number of training iterations fairly low (1,000), and the training was done in a single phase. Increasing the number of iterations and/or splitting the training into two phases of “rough” and “fine” ordering, as suggested by Kohonen (2013), did not lead to qualitatively different SOMs, probably due to the combination of small arrays and small datasets.
The package provides two neighborhood functions: bubble and Gaussian. Given the limited options provided by the algorithm, we did not test other functions such as the Epanechnikov function that was found superior by Liu et al. (2006), although Jiang et al. (2015) found no notable difference between Epanechnikov and Gaussian functions for SOMs optimized for classification. Both available functions led to SOMs with good topology; however, each of them required a specific range of neighborhood radii. In general, the Gaussian function led to SOMs less sensitive to the radius, notably to its lower terminal value. According to Wehrens and Kruisselbrink (2018) the Gaussian neighborhood always adjusts all types to some extent, even if the terminal radius is close or equal to the threshold of 1.0. For the bubble neighborhood, iterations calculated at the threshold account only for the winning type, which turns the process into k-means (Kohonen 2001). Regardless the function, in our case such iterations typically led to a marked change in quantization (decrease) and topological (decrease) errors and to dissipation of SOM topology. Gibson et al (2017) suggest training two sets of SOMs that differ in the inclusion of “k-means” iterations, arguing that the two sets may complement each other, one providing good topology, the other good clustering. Since SOM topology itself was the subject of our study, SOMs were only allowed to include a “k-means” phase if no folding of SOMs was detected in the Sammon projection.

Note that radius less than 1.0 mentioned in the study has no physical meaning and is specific to the “Kohonen” R package; there may be differences in other available software (Gibson et al. 2017). It allows for specifying the fraction of iterations with no neighbor modification: for example, radius set to linearly decrease from 3 to 0 means that the last one third of training does not modify neighbors.

We experimentally found that the effect of the learning rate and the initial radius was relatively small, once obvious errors were removed. In a few cases, which were discarded, an insufficient initial radius led to poor initial ordering of types in larger arrays. Additionally, too large learning rate in later iterations broke the topology of SOMs. These findings are in line with other studies (e.g., Sheridan and Lee 2011; Gibson et al. 2017).

The last point regards the use of Sammon mapping. Strictly speaking, Sammon mapping adds a step into the analysis that could be considered redundant or even introducing new uncertainty due to the black-box-like character of the projection that utilizes the nonlinear relations among all data points to define the projection plane. However, the idealized rectangular array of a SOM limits the study of its topological structure to such an extent that the benefits of Sammon mapping outweigh the negatives. Furthermore, utilization of a data projection technique allows one to see individual data points instead of SOM type frequencies only, which makes it possible to show how the linear modes of variability project on data in a nonlinear continuum framework.

4.3 On using SOMs to study teleconnections

Given the high number of combinations retained for the analysis (13 SOMs for each of the 12 datasets), only a relatively small subset of results could be chosen for presentation that, nevertheless, well represents the range of results. While the number of SOM types (size of array) does not seem to play that
important a role in SOM topology (contrary to its strong link to explained variation), the shape of the array (ratio of the number of columns and rows) as well as the structure of data are important factors.

The projection of two leading modes of variability on SOM array diagonals, a topology pattern suggested in earlier studies (e.g., Reusch et al. 2007), seems quite rare. An alignment of SOM diagonals with two leading modes—which results in mirrored patterns of positive and negative phases of modes in opposite corners—was detected only in square SOMs (in about 50% of cases). On the other hand, in about 60% of strongly elongated SOMs (2×4, 3×5, 4×6, and 5×7) another topology pattern was observed, in which the two leading modes are approximately aligned with the center row and column of the array, the diagonals then representing their combinations. Consequently, SOMs with an odd number of rows/columns show the spatial structure of modes somewhat better, especially those with small arrays. However, most of the remaining cases, and also most large SOMs and SOMs of datasets in which two (or all three) modes have approximately the same strength, cannot be described with either of the two idealized topology patterns, the projection and identification of modes becoming less straightforward.

The projection of the third-order mode on SOM arrays has so far received only little attention. Here, when the circular mode is weaker than the leading two modes, it projects strongly only on a few types close to the center of the array; but some curvature of isolines is apparent in most types. This is in line with how the mode projects on the Sammon map of data—fields relatively strongly correlated with the pattern of the mode may appear anywhere in the phase space, but in the center of the Sammon map such cases are more common, and refer to fields in which both leading modes are close to their neutral phase. As a result, such fields do not comprise marked overall anomalies and occur, therefore, close to the center of the Sammon space. A similar SOM structure can be seen in the results of Reusch et al. (2007) who analyzed North-Atlantic monthly mean SLP patterns (see Figs. 1 and 6 of their study). While the leading two PCs, which explained 39% and 13% of variability, projected relatively strongly on all the four corners of the SOM, the third-order PC (11%) did not have a clear representation within the 30-member SOM array. Despite similarities between the PC pattern and several types in the center of the array (e.g., type 16, Fig. 1), the two phases of this PC were not recognized as noteworthy regimes of North-Atlantic variability, owing to their overall weak projection on a SOM, at variance with the similar strength of the second- and third-order modes.

The identification of the second and third modes in SOM arrays becomes particularly challenging if their strength is identical. Several different topology patterns were found here, each pointing to a particular limitation of SOMs. First, one of the modes may project relatively weakly onto one or a few types (typically in the center of the SOM) as if its variance was smaller than that of the other mode—the mode thus being underrepresented. Second, parts of the SOM array not occupied by patterns similar to the leading mode (e.g., two opposite corners) may represent one (any) phase of one mode each, the remaining phases being underrepresented. Third, one part of a SOM array (e.g., upper half) may predominantly show one particular combination of phases (positive or negative) of the two modes, the other (e.g., lower) half showing a different combination of phases, but not necessarily opposite to the upper half for both modes, the remaining combinations of phases being underrepresented. Each of these
three topology patterns—if forced—may have specific utility, except that the pattern itself is largely outside of researchers’ control due to the unsupervised nature of SOMs. In real data, the underrepresentation of specific combinations of phases of modes may be, but equally likely may not be, a consequence of these combinations occurring less frequently in the training dataset—here it was a result of only minor differences due to sampling. This illustrates the fact that one has to pay a considerable price for the topology preservation constraint of SOMs, relative to simple cluster analysis.

Last, it has to be stressed that in both the Euro-Atlantic and the Pacific-American regions at least four modes of variability are necessary to describe variations in winter circulation, summer circulation being even more complex (Barnston and Livezey 1987; Casado and Pastor 2012; Hynčica and Huth 2020). Therefore, selective and uncontrollable underrepresentation of important modes of variability and their interactions will be an even more important issue in SOMs of real data.

### 4.4 On using SOMs to study the nonlinearity of teleconnections

We clearly show that the topology of SOMs strongly depends on the structure of data. Even if the leading two modes of variability project along elements of a SOM array (such as its diagonals), marked irregularities can be identified in the spatial patterns of types if the data comprise more than the two modes of variability. One way these irregularities demonstrate themselves in the spatial patterns of SOM types is a presence of non-mirrored patterns in opposite types within a SOM array, for example in opposite corners. Reusch et al. (2007) pointed out that such non-mirrored patterns are a feature that distinguishes SOMs from linear analyses (such as PCA). However, one needs to remember that such apparent nonlinearities do not necessarily point to nonlinearity of the underlying modes of variability—non-mirrored patterns in (any) two opposite positions within a SOM array clearly prevail even if data are linear. In the case of our synthetic datasets, these differences may be due to sampling, or simply be artifacts of projecting multidimensional data onto a two-dimensional phase space. We showed that especially lower order modes can project on SOM arrays in various ways, which, nevertheless, all lead to some kind of selective underrepresentation of co-occurrences of some phases of modes. Our data can be understood bivariate or trivariate—as regardless of the number of grid points, or input variables, all variability can be linked to only two or three modes. The fact that nonlinearity appears so clearly in SOMs of such simplistic datasets suggests that it will occur in any complex (multimodal, noisy) real-world data regardless of how well such data may be approximated by linear models. The tendency of nonlinear methods to detect nonlinearity where there is none has already been documented (Christiansen 2005), and one should, therefore, not forget the strong predisposition of results by the character of the used research method.

### 4.5 Outlook

The topology of SOMs relative to underlying modes of variability seems to be rather sensitive to data sampling. Here, we focused only on the topology; however, the response of the frequency of occurrence of individual SOM types to changes in the underlying modes may also be of interest, especially regarding
the ability of SOMs to capture inter-decadal changes in teleconnections and distinguishing them from sampling variability. This will be further analyzed in Part II of the study. Furthermore, the modes of variability predefined in the present study were highly idealized; generating synthetic datasets from real modes of variability—and potentially from a larger number of modes, adding noise to the generated data, and/or defining more complex nonlinear functions of data variability may represent other useful steps in studying the links between different perspectives of teleconnections. Last, other methods than SOMs might be utilized to study the continuum of teleconnections to assess whether they may improve on some of the limitations of SOMs discovered here.

Declarations

Funding

This research was supported by the Czech Science Foundation, project 17-07043S

Conflicts of interest/Competing interests

The authors declare no conflict of interest

Availability of data and material (data transparency)

The predefined modes of variability and all generated datasets can be accessed under the CC 4.0 license at dx.doi.org/10.6084/m9.figshare.13288889

Code availability (software application or custom code)

Not applicable

Authors' contributions

Not applicable

Additional declarations for articles in life science journals that report the results of studies involving humans and/or animals

Not applicable

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication
Acknowledgements

This research was supported by the Czech Science Foundation, project 17-07043S. We acknowledge the work of all authors of the Kohonen and MASS R packages, without which this study would not have been possible.

References

1. Ångström A (1935) Teleconnections of climatic changes in present time. Geografiska Annaler 17:242–258. doi:10.1080/20014422.1935.11880600
2. Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon Wea Rev 115:1083–1126. doi:10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2
3. Beranová R, Huth R (2008) Time variations of the effects of circulation variability modes on European temperature and precipitation in winter. Int J Climatol 28:139–158. doi:10.1002/joc.1516
4. Cannon AJ (2020) Reductions in daily continental-scale atmospheric circulation biases between generations of global climate models: CMIP5 to CMIP6. Environ Res Lett 15:064006. doi:10.1088/1748-9326/ab7e4f
5. Casado MJ, Pastor MA (2012) Use of variability modes to evaluate AR4 climate models over the Euro-Atlantic region. Climate Dyn 38:225–237. doi:10.1007/s00382-011-1077-2
6. Chang C, Johnson NC (2015) The continuum of wintertime Southern Hemisphere atmospheric teleconnection patterns. J Climate, 28, 9507–9529. doi:10.1175/JCLI-D-14-00739.1
7. Christiansen B (2005) The shortcomings of nonlinear principal component analysis in identifying circulation regimes. J Climate 18:4814–4823. doi:10.1175/JCLI3569.1
8. Comas-Bru L, Hernández A (2018) Reconciling North Atlantic climate modes: revised monthly indices for the East Atlantic and the Scandinavian patterns beyond the 20th century. Earth Syst Sci Data 10:2329–2344. doi:10.5194/essd-10-2329-2018
9. Compagnucci RH, Richman MB (2008) Can principal component analysis provide atmospheric circulation or teleconnection patterns? Int J Climatol 28:703–726. doi:10.1002/joc.1574
10. Corti S, Giannini A, Tibaldi S, Molteni F (1997) Patterns of low-frequency variability in a three-level quasi-geostrophic model. Climate Dyn 13:883–904. doi:10.1007/s003820050203
11. Davini P, Cagnazzo C (2014) On the misinterpretation of the North Atlantic Oscillation in CMIP5 models. Climate Dyn 43:1497–1511. doi:10.1007/s00382-013-1970-y
12. Feldstein SB (2000) The timescale, power spectra, and climate noise properties of teleconnection patterns. J Climate 13:4430–4440. doi:10.1175/1520-0442(2000)013<4430:TTPSAC>2.0.CO;2
14. Franzke C, Feldstein SB (2005) The continuum and dynamics of Northern Hemisphere teleconnection patterns. J Atmos Sci 62:3250–3267. doi:10.1175/JAS3536.1

15. Gao M, Yang Y, Shi H, Gao Z (2019) SOM-based synoptic analysis of atmospheric circulation patterns and temperature anomalies in China. Atmos Res 220:46–56. doi:10.1016/j.atmosres.2019.01.005

16. Gibson PB, Perkins-Kirkpatrick SE, Uotila P, Pepler AS, Alexander LV (2017) On the use of self-organizing maps for studying climate extremes. J Geophys Res 122:2891–3903. doi:10.1002/2016JD026256

17. Hannachi A, Straus DM, Franzke CLE, Corti S, Woollings T (2017) Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. Rev Geophys 55:199–234. doi:10.1002/2015RG000509

18. Hewitson BC, Crane RG (2002) Self-organizing maps: applications to synoptic climatology. Climate Res 22:13–26. doi:10.3354/cr022013

19. Horel JD (1981) A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. Mon Wea Rev 109:2080–2092. doi:10.1175/1520-0493(1981)109<2080:ARPCAO>2.0.CO;2

20. Huth R, Beck C, Philipp A, Demuzere M, Ustrnul Z, Cahynová M, Kyselý J, Tveito OE (2008) Classifications of atmospheric circulation patterns. Ann N Y Acad Sci 1146:105–152. doi:10.1196/annals.1446.019

21. Hynčica M, Huth R (2020) Modes of atmospheric circulation variability in the Northern Extratropics: a comparison of five reanalyses. J Climate 33:10707–10726. doi:10.1175/JCLI-D-19-0904.1

22. Jiang N, Griffiths G, Lorrey A (2013) Influence of large-scale climate modes on daily synoptic weather types over New Zealand. Int J Climatol 33:499–519. doi:10.1002/joc.3443

23. Jiang N, Luo K, Beggs PJ, Cheung K, Scorgie Y (2015) Insights into the implementation of synoptic weather-type classification using self-organizing maps: an Australian case study. Int J Climatol 35:3471–3485. doi:10.1002/joc.4221

24. Johnson NC (2013) How many ENSO flavors can we distinguish? J Climate 26:4816–4827. doi:10.1175/JCLI-D-12-00649.1

25. Johnson NC, Feldstein SB (2010) The continuum of North Pacific sea level pressure patterns: Intraseasonal, interannual, and interdecadal variability. J Climate 23:851–867. doi:10.1175/2009JCLI3099.1

26. Johnson NC, Feldstein SB, Tremblay B (2008) The continuum of Northern Hemisphere teleconnection patterns and a description of the NAO shift with the use of self-organizing maps. J Climate 21:6354–6371. doi:10.1175/2008JCLI2380.1

27. Jung T, Hilmer M, Ruprecht E, Kлеппе K, Gulev SK, Zolina O (2003) Characteristics of the recent eastward shift of interannual NAO variability. J Climate 16:3371–3382. doi:10.1175/1520-0442(2003)016<3371:COTRES>2.0.CO;2

28. Kohonen T (2001) Self-Organizing Maps. Springer, New York
29. Kohonen, T (2013) Essentials of the self-organizing map. Neural Netw 37:52–65. doi:10.1016/j.neunet.2012.09.018
30. Lennard C, Hegerl G (2015) Relating changes in synoptic circulation to the surface rainfall response using self-organising maps. Climate Dyn 44:861–879. doi:10.1007/s00382-014-2169-6
31. Liu Y, Weisberg RH, Mooers CNK (2006) Performance evaluation of the self-organizing map for feature extraction. J Geophys Res Oceans 111:C05018. doi:10.1029/2005JC003117
32. Philipp A, Beck C, Huth R, Jacobite J (2016) Development and comparison of circulation type classifications using the COST733 dataset and software. Int J Climatol 36:2673–2691. doi:10.1002/joc.3920
33. Reusch DB, Alley RB, Hewitson BC (2005) Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. Polar Geogr 29:188–212. doi:10.1080/789610199
34. Reusch DB, Alley RB, Hewitson BC (2007) North Atlantic climate variability from a self-organizing map perspective, J Geophys Res 112:D02104. doi:10.1029/2006JD007460
35. Reusch DB (2010) Nonlinear climatology and paleoclimatology: capturing patterns of variability and change with self-organizing maps. Phys Chem Earth 35:329–340. doi:10.1016/j.pce.2009.09.001
36. Risbey JS, O’Kane TJ, Monselesan DP, Franzke C, Horenko I (2015) Metastability of Northern Hemisphere teleconnection modes. J Atmos Sci 72:35–54. doi:10.1175/JAS-D-14-0020.1
37. Rousi E, Anagnostopoulou C, Tolika K, Maheras P (2015) Representing teleconnection patterns over Europe: A comparison of SOM and PCA methods. Atmos Res 152:123–137. doi:10.1016/j.atmosres.2013.11.010
38. Rousi E, Rust HW, Ulbrich U, Anagnostopoulou C (2020) Implications of winter NAO flavors on present and future European climate. Climate 8:13. doi:10.3390/cli8010013
39. Sammon JW (1969) A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C-18:401–409. doi:10.1109/T-C.1969.222678
40. Sheridan SC, Lee CC (2011) The self-organizing map in synoptic climatological research. Prog Phys Geogr 35:109–119. doi:10.1177/0309133310397582
41. Venables WN, Ripley BD (2002) Modern applied statistics with S, Fourth edition. Springer, New York
42. Verdon-Kidd DC (2018) On the classification of different flavours of Indian Ocean dipole events. Int J Climatol 38:4924–4937. doi:10.1002/joc.5707
43. Verdon-Kidd DC, Kiem AS (2009): On the relationship between large-scale climate modes and regional synoptic patterns that drive Victorian rainfall. Hydrol. Earth Syst Sci 13:467–479. doi:10.5194/hess-13-467-2009
44. Wallace JM, Gutzler DS (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. Mon Wea Rev 109:784–812. doi:10.1175/1520-
48. Wehrens R, Buydens L (2007) Self- and super-organizing maps in R: the Kohonen package. J Stat Softw 21:1–19. doi:10.18637/jss.v021.i05

49. Wehrens R, Kruisselbrink J (2018): Flexible Self-Organizing Maps in kohonen 3.0. J Stat Softw 87:1–18. doi:10.18637/jss.v087.i07

50. Yuan J, Tan B, Feldstein SB, Lee S (2015) Wintertime North Pacific teleconnection patterns: seasonal and interannual variability. J Climate 28:8247–8263. doi:10.1175/JCLI-D-14-00749.1

**Figures**

![Figure 1](image-url)

**Figure 1**

Spatial patterns of predefined circulation modes in their positive phase: a) zonal, b) meridional, c) circular. The interval of isolines is 2 units; positive (negative) values are in red (blue), the zero isoline is in green.
Figure 2

Schematic description of the methodology. The process was repeated for each of 12 datasets and each of 13 SOM shapes
Figure 3

Sammon map and a 4×5 SOM of Z60-M30-C10 dataset: a) Sammon map of 1,000 circulation fields (grey dots) and a SOM projected onto the Sammon map (red circles and lines); b) Spatial patterns of SOM types in panel a (units as in Fig. 1), with coordinates denoting row and column of types
Figure 4

The dependence of quantization and topological errors on terminal radius (horizontal axes), SOM shape (columns) and dataset structure (rows). 20 variants of SOMs were defined for each radius; variants selected for the study are highlighted by red crosses. The red box highlights three local optima of the topological structure (see text and Fig. 5 for more information). The units of vertical axes are (a) Euclidean distance of an unspecified variable and (b) fraction of 1.00 (see Section 2 for definition of the errors)
Figure 5

Sammon maps of three variants of 3x4 SOMs representing three local optima of topological structure. TE and QE indicate the topological and quantization errors associated with the respective SOMs. A few data points of high y-value were omitted from the figure. See Fig. 3 for legend.
Figure 6

Sammon maps of 4×4, 4×5 and 4×6 SOMs of bivariate datasets. Unlike in Fig. 3, here the Sammon maps of datasets are colored according to Pearson's correlation between patterns of circulation fields and positive phase of respective modes (Fig. 1). Each row shows results for one dataset: (a) Z90-M10, (b) Z70-M30, and (c) Z50-M50. In each row, the two leftmost panels show the identical 4×4 SOM array (in black dots and lines), while each panel shows correlation of data with one mode (Z for zonal, M for meridional). In the right-hand half of the figure, the two columns show projections of 4×5 and 4×6 SOMs; correlations with the zonal mode only are shown. Larger blue/red dots denote projections of negative/positive phases of a respective mode, each dot representing a certain multiple of the pattern in Fig. 1, the dots closest to the Sammon map center [0,0] being ±0.5 times the pattern, the second closest being ±1.0 times the pattern, etc.
(a) Z60-M30-C10

(b) Z50-M40-C10

(c) Z50-M30-C20

Figure 7

As in Fig. 6, except for (a) Z60-M30-C10, (b) Z50-M40-C10, and (c) Z50-M30-C20, and inclusion of the third-order (C) mode. In each row, the three panels on the left show the Sammon map of one dataset and projections of the 4×6 SOM array (in black dots and lines) and circulation modes (larger blue and red dots) onto it, while the two figures on the right show shape of 4×4 and 4×5 SOMs.

Figure 8
Spatial patterns of selected SOMs: (a)–(c) 4×6 SOMs shown in Fig. 7, (d)–(f) 4×6 SOMs shown in Fig. 9. For legend, see Fig. 3

(a) Z60-M20-C20

(b) Z50-M25-C25

(c) Z40-M30-C30

Figure 9

As in Fig. 7, except for (a) Z60-M20-C20, (b) Z50-M25-C25, and (c) Z40-M30-C30