Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

On the fuzziness of circulation types derived from the application of obliquely rotated principal component analysis to a T-mode climatic field

Chibuike Chiedozie Ibebuchi (chibuike.ibebuchi@uni-wuerzburg.de)
University of Würzburg

Research Article

Keywords: Circulation types, Principal component analysis, T-mode, Africa south of the equator, separability, threshold values

Posted Date: May 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-530514/v1

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

Loading [MathJax]/jax/output/CommonHTML/jax.js

Abstract

This study examined the separability of circulation types (CTs) classified from the application of principal component analysis (PCA) to the T-mode matrix (variable is time series and observation is grid points) of a climatic field that explains atmospheric circulation; in addition to the uncertainty introduced on (i) the probability of occurrence, (ii) the mean shape of the CTs, (iii) the trend in the annual frequency of occurrence, (iv) the frequency distribution of the CTs, by using varying threshold values within the range of 0.2–0.35 to assign days to a given CT. The study region is Africa, south of the equator. Some large clusters were classified with most days in the analysis period assigned to them; these classes are interpreted as the dominant states of the atmosphere and generally, their existence results in the poor separability of the CTs since their features overlap with other CTs. Qualitatively, the choice of the threshold values within the defined range has little or no influence on the overall structure of the probability of occurrence of the CTs, the mean shape of the CTs, and the year-to-year variations in the annual occurrence of the CTs. However, it significantly impacts the frequency distribution of the CTs and the statistical significance of the trend in the annual occurrence of the CTs. Stringent threshold values within the defined range might benefit studies that aim to isolate days when specific CTs are most expressed and analyze their mechanism using composite maps, without focus on the frequency distribution and annual occurrence of the CTs. Overall, for the study region, lower threshold values within the defined range might be recommended since relatively, they do not tend to further constrain the probability of group membership, and equally seem to reveal the mechanisms that might be consistent when a given CT occurred regardless of the strength of its signal at a given time.

1 Introduction

A major goal of synoptic climatology is weather forecasting. It is focused on exploring the relationship between a local climate and classified large-scale atmospheric circulation patterns (Yarnal 1993). Synoptic climatological classifications can have a wide range of applications such as finding teleconnection patterns (e.g. Kim et al. 2021), regime classification (e.g. Plaut and Simonnet 2001), identification of circulation types (CTs) linked to specific weather conditions (e.g. Beck and Philipp 2010), and more. Using principal component analysis a spatial decomposition (S-mode) of an appropriate climate field can be used to obtain teleconnection patterns (Compagnucci and Richman 2008). Also, an S-mode PCA can be used for regionalization purposes which involves the classification of regions (grid points or stations) that covary with respect to a given time development (Richman and Lamb 1985; Ibebuchi 2021a). According to Huth et al. (2008) regime classification differs from circulation typing. The spatial scale, time scale, persistence of the patterns, and the number of classes mark the difference between the two concepts. Regime classification is typically done at a continental or hemispheric spatial scale; the time scale of the classified variable is needed to be longer than daily; the classified patterns persist for a relatively long time, and the numbers of classified patterns are relatively smaller. Circulation typing which is the focus of this study can be done at a relatively smaller spatial and time scale; the number of classified patterns can be larger and the patterns do not necessarily have to appear repeatedly.

Loading [MathJax]/jax/output/CommonHTML/jax.js

The meteorological field used in classifying the atmospheric circulation patterns is continuous (Compagnucci et al. 2001). For this reason, the separation of the inherent structures of variability making up an atmospheric field used in classifying atmospheric circulation into discrete (well-separated) patterns might affect the meteorological robustness of the classification output, and to what extent the stratification of surface variables is physically robust (Richman and Gong 1999). The quest for well-separated classes defies the continuum nature of atmospheric circulation and also makes it look like patterns associated with extremes occur (nearly) equally in time with patterns associated with normal weather conditions. This is why Mo and Ghil (1988) argued that every good synoptic classification should not have equal and well-separated classes. Thus there is the need to investigate the classification of atmospheric circulation from the perspective of the continuum and fuzzy nature of the atmospheric field used in the classification.

The most recent developments in circulation type classifications, addressed in the regional context of Europe, were discussed by Philip et al. (2010) and interested readers are referred to their work. A major drawback in clustering techniques when it comes to synoptic climatological classifications is the application of "hard clustering" to continuous data. Hard clustering implies that a variable can be assigned to only one class in the classification process. Among the hard cluster analysis algorithms, the K-means is the most used in synoptic classifications. It is based on partition (see Xu and Tian 2015 for more details). According to Huth et al. (2008) *"K-means procedure tends to produce equally-sized groups, which may be unrealistic, the groups being very well separated".* This attribute can nevertheless be useful when the research goal aims at a good stratification of surface variables without probing the physical validity of the CTs used in the stratification. Also, the K-means can be sensitive to the pre-set parameters leading to instability of the classification, though this drawback has been addressed by the simulated annealing algorithm which has proved to be superior to the K-means. Gong and Richman (1995), however, warned that in applied geophysical research, the application of hard clustering techniques to a continuous and fuzzy dataset is questionable. The fuzzy K-means can be used to overcome this hurdle as applied by Harr and Elsberry (1995). The Self Organising map is another clustering algorithm based on model. Its clustering approach might outperform the hard cluster analysis algorithms based on partition since it tends to treat the data set as continuous (Hewiston and Crane 2002), and its ability to generate a non-linear classification (Philippopoulos et al. 2014). Another major issue with clustering a climatic field is that the climatic data set is subjected to sampling error (North et al. 1982) and it might be necessary to extract the feature in the data set before the cluster analysis (Bartholv et al. 1987). For this purpose, PCA might be used to denoise the climatic data and extract the important features.

A detailed review of clustering techniques by Xu and Tian (2015), Gong and Richman (1995), showed that the probability of group membership, which is an interesting property of geophysical classifications, can only be achieved with fuzzy classifications – i.e. classifications that allow overlapping of the classified variables. While every cluster analysis techniques have their advantages and disadvantages, a detailed review by Xu and Tian (2015) noted that fuzzy classifications have relatively high accuracy of clustering. Based on fuzzy K-means, Harr and Elsberry (1995) noted: "*the fuzzy analysis allows identification of*

Loading [MathJax]/jax/output/CommonHTML/jax.js

veral clusters. These may be associated with slowly varying

circulation features that may identify the mechanisms that cause a circulation pattern to change from one cluster to another". By this statement, they have implied that fuzzy classifications can aid the identification of dominant patterns (and rare patterns), in addition to a deeper understanding of how CTs might evolve at a given time, due to dynamical linkages.

Richman (1981) advanced the use of PCA as a classification tool for atmospheric circulation patterns but with the condition that the retained components must be rotated, preferably obliquely, to suppress orthogonality constraint. Gong and Richman (1995) explained that the rotated PCA when used as a classification tool is inherently fuzzy. The introduction of threshold values within the range of $\pm 0.2 - \pm 0.35$ as recommended by Richman and Gong (1999) to separate the PCs results in some degree of uncertainty in the cluster membership but allows overlapping of the classified variables despite the binarization (Gong and Richman 1995). The choice of the threshold values to use within the defined range according to Richman and Gong (1999) should be based on the sample size. However, in the regional context of Africa, south of the equator and beyond, when rotated PCA is used as an eigenvector-based synoptic classification tool, no study has addressed in detail, the uncertainty introduced to the classification output by using varying thresholds within this range. A study by Barreira and Compagnucci (2011) used ± 0.3 to separate the PCs for the analysis of the spatial fields of Antarctic sea-ice concentration anomalies for summer–autumn and their relationship to Southern Hemisphere atmospheric circulation, but no justification was provided for the choice of this threshold value.

When rotated PCA is applied to a raw field that explains atmospheric circulation and represented in the T-mode structure (i.e. variable is time series and observation is grid points), several studies have indicated that the time decomposition achieved with this approach can reproduce CTs known *a priori* in a given region (Richman 1986, 1983; Vargas and Compagnucci 1983, Compagnucci and Ruiz 1992; Huth 1996, 2008; Compagnucci and Richman 2008). According to Companucci et al. (2001) "*The T-mode analysis can lead to the determination of frequent synoptic situations, improving the basic knowledge essential to weather forecasting, among other things. The application of such a tool to a wide range of processes, ranging from the daily synoptic developments to the monthly or annual mean developments is valuable for an ample set of atmospheric processes, including both daily variability and climate fluctuations and change"*. Comparisons between climatological classification achieved with rotated PCA and other hard clustering algorithms have been made both for circulation typing (e.g. Huth 1996) and regionalization (e.g. Gong and Richman 1995), the conclusions are that classifications with rotated PCA preserves the underlying physics in the data set and can reproduce classes known *a priori* compared to hard clustering algorithms. The rotated T-mode PCA has been widely applied in classifying CTs in different parts of the globe, such as in the northern hemisphere (e.g. Bartzokas 1996), in Europe (e,g. Huth 1993); in Asia (e.g. Zhao et al. 2019), but remains unpopular for circulation typing in Africa, south of the equator where the hydroclimate is vulnerable to climate change (IPCC 2013). The usefulness of this clustering tool to investigate climate change as highlighted by Compagnucci et al. (2001) makes it deserve further attention in studying the hydroclimate of Africa, south of the equator. Some studies have indicated that the rotated T-mode PCA produces CTs that are poorly separated (e.g. Huth 1996). The poor separability of

Loading [MathJax]/jax/output/CommonHTML/jax.js

xamined in this study to evaluate if a physical justification

might be attributed to it. Also, for assigning realizations (days in this case) to a given CT, the uncertainty introduced by using the recommended threshold values within the range of 0.2–0.35 on (i) the probability of occurrence, (ii) the mean shape of the CTs, (iii) the trend in the annual frequency of occurrence, (iv) the frequency distribution of the CTs are examined also in this work. Thus given the subjectivity of some decisions made in using this clustering approach, for the regional context of Africa, south of the equator, the strength of this work is to provide methodological guidelines that might guide researchers, who wish to use this classification tool in the region, in optimizing its applicability for circulation typing and the linkage of the CTs to rainfall variability at specific local climates in southern Africa and beyond.

2 Data And Methodology

2.1 Choice of data set

For the circulation typing, Sea level pressure (SLP) is obtained from ERA5 (Hersbach et al. 2020). The horizontal resolution of the SLP field is 0.25° longitude and latitude and the temporal resolution is daily from 1979-2018. According to Kidson (1997), SLP provides a good representation of synoptic-scale systems; explains the relationship between topography and low-level flow; and between 1000 hPa to 500 hPa, the choice of the level to use in the classification of CTs has little influence on the explanation of surface variables. The choice of ERA5 in this study is due to the relatively high horizontal resolution of its SLP field.

2.2 Choice of domain size

The choice of domain size has been reported to influence the classification output and the relationship between CTs and surface climate (Beck et al. 2013). The spatial coordinate for the circulation typing in this study is 5.25°E-55.25°E and 0°-50.25°S. Since the CTs are aimed to be used in studying the local hydroclimate of southern Africa and beyond, the choice of this domain is based on prior knowledge of the spatial extent of the major rain-bearing synoptic systems that influence the local hydroclimate of the target regions within the domain. Though some of these systems are subjected to move around, Fig. 1 presents their typical locations during their active periods. The adjacent oceans are included since they act as moisture sources to the landmasses (Reason and Mulenga 1999). Parts of the

Loading [MathJax]/jax/output/CommonHTML/jax.js

tropics are included to capture the cross-equatorial northeast trade winds that transport moisture into the Angola low. To the east, the western branch of the Mascarene high drives southeast moisture fluxes into southern Africa (Cook 2000), though the fluxes are deflected southward by the high Madagascar topography, and northwest by the Mozambique Channel trough (Barimalala et al. 2019). To the south, the ridging of the South Atlantic Ocean high-pressure south of South Africa can be associated with enhanced southeast moisture advection to the eastern landmasses (Ndarana et al. 2018), and also the northward track of the mid-latitude cyclones can be associated with cold fronts sweeping across the Western Cape and bringing about rainfall over there (Reason and Ronault 2005). To the west, the South Atlantic Ocean high-pressure drive moisture offshore. The activities of these synoptic systems, with the inclusion of evaporation rate at the Agulhas warm current which flows down from 27°S to 40°S (Gordon 1985), directly or indirectly, influence convergence into the South Indian Ocean Convergence Zone (Cook 2000) which is the major large-scale system that controls the (austral summer) hydroclimate of southern Africa. The South Indian Ocean Convergence Zone is related to the Inter-tropical Convergence Zone through the Angola low (Reason and Smart 2015).

2.3 Methods

For the CT classification, obliquely rotated PCA is applied to the z-score standardized SLP field represented in the T-mode structure. The annual cycle of the SLP field is not removed since to obtain CTs that are physically interpretable, Huth (1996, 2008) noted that the annual cycle has to be retained. The correlation matrix is used to relate the time series. Singular value decomposition is used to obtain the eigenvalues and eigenvectors. The eigenvectors were loaded with the square root of their corresponding eigenvalues making them longer than and unit length referred to as loadings. Richman and Lamb (1995) noted that this step makes the eigenvectors responsive to rotation. The loadings localize in time the spatial patterns captured by the PC scores (Compagnucci and Richman 2008).

Loading [MathJax]/jax/output/CommonHTML/jax.js

For the decision of the number of components to retain, an iterative approach was considered in line with expert knowledge of synoptic situations in the study region. First, North et al. (1982) explained that a climatic field is subjected to sampling errors and there is the need to analyze part of the data that explains most of the variability in the field. In line with the philosophy of North et al. (1982), the scree-test is the usual approach to decide the number of components to retain based on the separability of the eigenvalues, but since subsequent steps require rotating the vectors, this approach might be trivial. However, the scree-test was used as a first guess to know the number of components where the bulk of the variance is concentrated. This was followed by adding more components and rotating them iteratively to investigate if the added component has already been delineated by previous vectors as suggested by Richman (1986). This was statistically analyzed by computing the congruence coefficient between the scores and a visual inspection of the maps. While no classification can be considered as a truth (Huth et al. 2008), this step ensures that the addition of a new component does not represent in the absolute sense a subset of an already classified pattern. Also, this approach follows the philosophy of Harr and Elsberry (1995) in deciding the number of clusters in a fuzzy K-mean analysis, they noted: "*the number of clusters is increased until the additional cluster either represents a subset of extreme cases from previously defined clusters or the number of members within the new cluster is less than 10% of the next smallest cluster.*"

The retained components are rotated obliquely using Promax at a power of 2 (Richman 1981, 1986). The oblique rotation eliminates orthogonality constraint and maximizes the number of near-zero loadings so that each retained component clusters a unique number of days. Each of the retained components is associated with a unique spatial pattern (mode of variability), and time series of the analysis period - represented by positive and negative loadings. The different phases of the loadings can form dipoles that can be physically interpretable for a well-designed analysis (Compagnicci and Richman 2008). The magnitude of the loadings is an important signal and for a given retained component, loadings below specific threshold values can be considered as noise (Richman and Gong 1999) - implying that those days do not contribute to the PC scores (Compagnicci and

Loading [MathJax]/jax/output/CommonHTML/jax.js

Richman 2008). In other words, for a given retained component, days with loadings above a defined threshold can be interpreted as days when the mode occurred and for a given day, the magnitude of the loadings is the amplitude of the mode for the day in question. To improve the weather coherency of the days in a given mode, they are clustered into negative loadings and positive loadings above a threshold. Thus each retained components form two classes and the mean SLP of the days assigned to each class is the CT. Gong and Richman (1995) explained the introduction of a threshold value to separate the PCs as a binarization that tends to harden the classification, but regardless, it does not constrain overlapping of the classified variables insofar the threshold values are within a reasonable range that allows the probability of group membership. So a day can have its loadings greater than the defined threshold under more than one retained component so that more than one CTs can occur in a given day. Richman and Gong (1999) have recommended that both for S-mode and T-mode analysis, threshold values within the range of 0.2-0.35 can be used to separate the PCs. Thus for each retained component, threshold values of and are applied, and the result of the uncertainty introduced by using the threshold values are examined on (i) the probability of occurrence, (ii) the mean shape of the CTs, (iii) the trend in the annual frequency of occurrence, (iv) the frequency distribution of the CTs. Spatial correlation analysis is used to assess changes in the spatial variation of the SLP field under each scenario, the statistical significance of the correlation is done using the Kendall-Tau test at a 95% confidence level. Comparison of the frequency distribution of the CTs under each case is made using the Wilcoxon test; and the test of statistical significance of the trends in the annual frequency of occurrence of the CTs is made using the Mann-Kendall test (Mann 1945; Kendall 1975) at a 95% confidence level.

Finally, since a day can be assigned to more one than class, Eq. 1 is developed to measure the pair-wise separability of the classes. For two given CTs, it assesses the number of days both CTs have in common, these days are subtracted from the total number of days classified under each of the two CTs in question to obtain for each of the CTs, respectively, and the result is expressed in the form of a percentage. Values closer to 100% show satisfactory separability of a given CT directly compared to another CT and vice versa. The

Loading [MathJax]/jax/output/CommonHTML/jax.js

poor separability of the CTs also implies that the relative frequencies of the individual CTs will not add up to 100%.

$$P_{unique_{CT_i,CT_j}} = \frac{N_{unique}}{N_{total}} \times 100$$

Eq. (1)

$P_{unique_{CT_i,CT_j}}$ = Percentage of unique observations (days) in CT_i in comparison to CT_j

N_{unique} = Number of observations (days) in CT_i that were not clustered in CT_j

N_{total} = Total number of observations (days) in CT_i

3 Results

By retaining 9 components, 18 CTs were classified. Fig. 2 shows the CTs and Table 1 shows the range and the median value of the positive and negative loadings from each of the retained components that yield the CTs. Each of the classes in Table 1 is associated with a unique range of loadings. The median loadings of CT1+, CT2-, CT3- and CT4+ (i.e. before the thresholds are introduced) are generally greater than and are within the range of 0 to approximately. These CTs are highlighted in Fig. 2 by the thick black frames. They can have relatively more days assigned to them and are representative of the dominant states of the atmosphere. Regardless of the choice of the threshold to use in separating noise from the actual signal, Table 1 indicates that CT1+ will have most of the days in the analysis period clustered under it; Moleteni et al. (1990) identified such large cluster class, which they explained to be close to the climatological mean state of the atmospheric circulation in the region. Using a T-mode analysis, a similar result was obtained by Huth and Canziani (2003), whereby the first retained component yielded the class close to the mean pattern of atmospheric circulation in the region. Analysis of the annual cycle of the CTs (Ibebuchi 2021c) showed that CT1+ tends to dominate during austral winter, while CT3- and CT4+ are close to the austral summer mean patterns. CT2- does not tend to exhibit seasonality. From Fig. 3, the aforementioned CTs have relatively the highest probability of occurrence regardless of the choice of the threshold values used. However, with a less stringent threshold, more days will be assigned to a given CT and vice versa. Under T0.2, the probability of occurrence of the CTs is relatively higher and gradually decreases when the threshold value is increased towards T0.35. Gong and Richman (1995) reported that hardening the classification obtained from rotated PCA decreases the quality of the classification output, and this might be related to the fact that increasing the threshold value, decreases the probability of

Loading [MathJax]/jax/output/CommonHTML/jax.js

group membership which is among the most important added value of fuzzy classifications (Xu and Tian 2015).

The range and median value of some classes are relatively very low, for example, CT6-; and from CT8+ to CT9-. These classes from Fig. 3 have a relatively low probability of occurrence and they can be designated as rare patterns. When the CTs are related to rainfall in southern Africa, CT6- was found to be associated with widespread rainfall in many parts of southern Africa (e.g. Ibebuchi 2021a). Studies have reported that the synoptic state of CT6- is associated with widespread rainfall in southern Africa (e.g. Cook 2000), since a stronger circulation at the western branch of the Mascarene high, ridging into the eastern parts of southern Africa, during austral summer, leads to enhanced onshore moisture fluxes by southeasterlies. CT6- is also dominant during austral summer, thus its timing and synoptic feature correspond with the synoptic situation known *a priori* to be linked to above-average rainfall in southern Africa. Table 1 indicates that relatively lesser days can be assigned to CT6- and its probability of occurrence is equally low (Fig. 3). Physically, this is reasonable given that extreme weather conditions can be relatively rare. Depending on the target region of interest, other CTs found to be associated with extremes are CT5+, CT7+, CT9- (e.g. Ibebuchi 2021d, 2021e), thus given the high regional heterogeneity of the climate of southern Africa, each of the CT can be relevant depending on the target local climate where the rainfall response is focused. Generally, the CTs associated with extremes are relatively rare which fits the idea that extremes in rainfall are not dominant atmospheric states, as most rigid classification algorithms tend to portray.

Table 1 and Fig. 3 suggest that the use of stringent threshold value (i.e. T0.35) might relatively have less impact on the CTs designated to be close to the mean patterns, but can significantly influence the probability of occurrence and the days assigned to CTs that are rare. More often, the rare patterns associated with extremes are of interest in geophysical research, thus care has to be taken in the classification process to ensure that they are well represented. Physically, given the continuum nature of atmospheric circulation, hard clustering or use of stringent thresholds such as T0.35 might under-represent days when the rare patterns even though occurred but had relatively weaker signal compared to other patterns that occurred the same day; since hard clustering or using stringent thresholds tend to simplify the classification and this (over) simplification might degrade the extent to which the occurrence of rare patterns are detected.

To this end, from Table 1, it is vital to note that at higher components, the range and median value of the loadings decreases, and lesser days are assigned to the CTs from higher components, thus it was inspected to what point that adding a new class yields a new CT that has not been previously delineated. At the 9th (last) component, very CTs which are known by expert knowledge were detected. These CTs captured the seasonal variability in the Mozambique Channel Trough which modulates the hydroclimate of southern Africa (Barimalala et al. 2019). Thus CT9+ and CT9- are useful, and from the 10th component onwards, the CTs either represented a subset of already classified CTs and/or have very few days assigned to them even when the least threshold value (T0.2) is used.

Loading [MathJax]/jax/output/CommonHTML/jax.js

Fig. 4 and Table 2 show the implication of using the different thresholds on the mean shape of the CTs. From Table 2, the spatial variation in the SLP field measured by the correlation coefficient is not significantly affected as the threshold value increases to T0.35. The correlation coefficient is greater than 0.9 in all cases and is statistically significant. However, from T0.3, the spatial variations of SLP in the rare patterns were relatively altered and this was more pronounced at T0.35; Fig. 4 exemplifies this, using the CT1+ (which is close to the overall mean pattern); CT6- and CT9- (which are rare patterns associated with extremes). It can be seen that CT1+ is least affected in all cases, but in CT6- and CT9-, the centers of action seem to be more expressed at higher threshold values. For example under CT6-, at T0.35, the Kalahari low, the Angola low, the Angola tropical low and the weaker circulation at the South Atlantic Ocean high-pressure are relatively more expressed compared to T0.2; and under CT9-, at T0.35, the low-pressure system in the Mozambique Channel and the east coast of Madagascar are well expressed compared to T0.2. The explanation is that using T0.35 will only cluster days when the signal of the CT was relatively most expressed, compared to T0.2.

Fig. 5 presents the annual occurrence of the CTs exemplified from the first eight CTs. In all cases based on the Wilcoxon test, the frequency distribution of the CTs is non-identical at a 95% confidence level. Thus the frequency distribution of the CTs is significantly impacted by using different threshold values. However, the year-to-year variability in the annual occurrence of the CTs is captured under each case. When the annual occurrence of the CTs is analyzed, the trend significance is mostly of interest. For the analysis period, a significant trend was found in the frequency of occurrence of CT1+ and CT4+ in all cases except for T0.35. On the other hand, a significant trend was found in all cases for CT4-. Thus the result suggests that it is likely that between T0.2 to T0.3, the statistical significance of the trends might be coherent.

Finally, the result of the pair-wise separability of the CTs is presented in Fig. 6. For clarity, the two CTs from a given retained component are 100% separated and a CT is equally 100% separated with itself. For the interpretation of Fig. 6, is directly compared to and interpreted as being separated from using the color scale in the legend. If on a 100% scale, 0-20% can be designated as very poor separability, and 85-100% designated as perfect separability, then it can be seen that generally most of the classes are perfectly separated, except for CT2+ that is very poorly separated from CT3-, CT3+ that is very poorly separated from CT1+ and CT2-. The very poor separability of CT3+ from CT1+ and CT2- can be due to dynamical linkage since during austral winter when CT3+ prevails, the semi-permanent high-pressure systems are situated on the (southern) landmasses, moving from west to east as presented by CT1+ and CT2-, and the northward track of the mid-latitude cyclone during CT3+ causes the disintegration of the high-pressure system into an eastern and western branch. A similar argument might be inferred from the very poor separability of CT2+ from CT3- whereby CT3+ occurs and triggers when CT2+ when conditions are favorable. Thus when two classes exhibit (very) poor separability, a physical explanation might be sought when the researcher is familiar with the large-scale physical processes in the study region. Generally, in most cases, the CTs are poorly separated (0-45%) with the classes designated as being close to the mean patterns (i.e. CT1+, CT2-and CT3-) which justifies that they are truly dominant states of the

Loading [MathJax]/jax/output/CommonHTML/jax.js

atmosphere, with slowly varying features, and with their features overlapping with other CTs (Harr and Elsberry 1995).

4 Discussions

In this study the fuzziness of CTs in Africa, south of the equator, classified from the application of rotated PCA to a climatic field that explains atmospheric circulation and represented in the T-mode structure was examined in the light of the pair-wise separability of the CTs and the uncertainty introduced by using threshold values within the range of 0.2–0.35 as recommended by Richman and Gong (1999), to assign days to a given class. It was found that some CTs have a relatively higher probability to occur, and some other CTs are rare patterns. Harr and Elsberry (1995) explained that an advantage of fuzzy analysis is that it allows the identification of such CTs that may (i) contain the attributes of several CTs, (ii) have slowly varying features that may identify mechanisms through which a circulation pattern change from one CT to the other. Huth (1996) explained such classes as the dominant or actual occurring states of the atmosphere; Mo and Ghil (1988) and Moleteni et al. (1990) insisted that every good classification of circulation pattern should have such large classes. This is unlike hard clustering algorithms (e.g. K-means clustering) whereby all the classes tend to be well separated, which might be physically artificial. Generally, the existence of such classes was found to be a reason why the T-mode classification is not well separated; otherwise, the classes show satisfactory separability. Poorly separated classes were found to be also possibly due to dynamical linkages whereby a CT has to precede the occurrence of another CT. The linkage of the CTs to precipitation at different domains in southern Africa has been proved to be physically interpretable and validated (Ibebuchi 2021b, 2021c, 2021d, 2021e) whereby a physical circulation mechanism from the CT in question (and known *a priori* from existing works of literature based on dynamical simulations) physically correlates with the spatial distribution and intensity of rainfall in the local climate of interest. This is unlike some classifications whereby the focus is on how well a surface variable is stratified without investigating if the CTs used for the stratification are actual synoptic situations and if the linkage between the CT and the surface variable has a physical justification. Moreover, a *physical correlation* was found between analyzed variables and rainfall. For example, CT7- associated with large-scale subsidence was found to be one of the driest CTs in southern Africa (Ibebuchi 2021a). Its dominant period is in austral winter and according to Dedekind et al. (2016), during austral winter, large-scale subsidence is the major mechanism leading to dryness. Ibebuchi (2021d) investigated the climatic modes associated with an ample set of CTs associated with extremes in wet and dry conditions in Mozambique, and the result equally showed that there is a physically interpretable correlation between the CTs and major climatic modes in the south Indian Ocean.

Richman and Gong (1999) noted: *"Researchers should consider if the geophysical research question and the real world data are hard in nature before choosing a clustering analysis program. In our experiment, the causal mechanisms for precipitation are fuzzy and overlapping. For an applied research problem we suggest an examination of physical mechanisms prior to embarking on a specific methodology and then choosing techniques that best preserves the underlying physics".* Huth 1996 equally stated, *"The ability to* Loading [MathJax]/jax/output/CommonHTML/jax.js *et (i.e. the ability to reflect the underlying physics) is the most*

important property of classification methods". Thus this study aims to further suggest that researchers should try to reconcile statistical methods with physical interpretation; there is the need to have a physical understanding of the classified variable, region of interest, and synoptic systems, including their variability and linkage with rainfall in different seasons in the region before embarking on circulation typing, otherwise one might not be able to separate statistical artifacts from actual synoptic situations. In principle, statistical tests do not validate the actual existence of CTs, when they are not known *a priori*.

A plausible way to examine the synoptic situations *a priori*, for example, synoptic situations associated with extremes in rainfall, is to investigate the pressure map or any other field that represents atmospheric circulations at specific period(s) (e.g. day) when an extreme event was recorded in the target region where the response of the surface variable to atmospheric circulation is sought. A good classification output will reproduce the synoptic feature represented on the analyzed day(s) and will reveal that a CT designated to be linked to extremes occurred on that particular day so that its occurrence can be linked to the flood/dry episode (e.g. Ibebuchi 2021d). A similar approach was used by Richman (1983), whereby the 700 mb height data set was examined before the classification to get an insight into the nature of the height anomalies over the domain. Moreover, there have been successful efforts to reconcile CTs from T-mode analysis with physically justifiable patterns of atmospheric circulation (e.g. Huth and Canziani 2003).

The domain size (Beck and Philipp 2010) and choice of the reanalysis data set (Stryhal and Huth 2017) are factors that also influence the CT classification output. Ibebuchi (2021b) classified the CTs using the same classification scheme but removing the deep tropics (i.e. 0–5°S) and this did not influence the classification output, but when the domain size was far extended further beyond the chosen domain in this work, the CTs obtained were different and by expert knowledge they did not correspond to synoptic situations known in the region *a priori*. The selected domain which resulted in CTs that is physically meaningful is based on a physical understanding of the spatial extent of the synoptic systems that modulate the hydroclimate of southern Africa and beyond. However, it should be noted that due to the weak height gradients in the Tropics, classifications that are focused on the tropical regions might prefer wind data to identify the variability of circulation features in the tropics (e.g. Harr and Elsberry 1995). The application of the CTs in this work should be in addressing the hydroclimate of southern Africa and beyond and not necessarily the deep tropics (i.e 0–5°S).

Comparison of the CTs and their statistical properties as obtained from the different reanalysis products, e.g. ERA-Interim and NCEP-NCAR, in addition to CTs derived from appropriate GCMs, generally reproduced the CTs and their statistical properties (e.g. Ibebuchi 2021c) though the frequency distributions are not exact. Generally, in all cases, the CTs were reproduced with a one-to-one correspondence regardless of the choice of the SLP data set used in the classification. Classifications from ERA5 and NCEP-NCAR are however in better agreement compared to ERA-Interim.

The impact of using different threshold values to assign days to a given CT was found to be more pronounced on the frequency distribution of the CTs and the statistical significance of the trend in the

Loading [MathJax]/jax/output/CommonHTML/jax.js

annual occurrence of the CTs. Under each selected threshold value, the spatial variation of the composite SLP field of the CTs, the mean shape of the CTs, the overall structure in the probability of occurrence of the CTs, and the year-to-year variations in the annual occurrence of the CTs were not significantly affected. For this reason, studies that aim to analyze trends in the annual occurrence of the CTs have to ensure that the analysis is made using the set of defined thresholds and even different reanalysis products, to ensure robustness. However, between T0.2 to T0.3 might be recommended for such studies since the statistical significance of the trends within this range seem to be coherent. Also, studies that aim to isolate specific CTs (e.g. CTs associated with extremes) and study the centers of action of such CTs without focus on the frequency distribution of the CTs might use threshold values within the range of T0.3 to T0.35. However, as Gong and Richman (1995) noted, the more climatological classifications with rotated PCA are hardened the more the quality of the classification degrades in the physical sense. Thus to have a physically meaningful classification output, and to ensure that the pre-set threshold does not further constrain the probability of group membership – which is the major added value for using a fuzzy classification scheme - stringent thresholds should be avoided, and in the author's opinion, T0.2 might be preferred, at least for circulation typing in southern Africa.

In summary, while there will be some level of uncertainty associated with fuzzy classifications such as rotated PCA, this uncertainty is within the reach of the physical reasoning that one might not accurately and objectively specify when the amplitude of a given CT can be regarded as noise or signal. However, this study has provided guidelines so that the researcher in line with the research design might be guided on the choice of the best thresholds within the defined range, at least for the study region in this work. Uncertainty is an indispensable attribute of atmospheric processes and it is good when this attribute is incorporated in synoptic climatological classifications. Hard clustering algorithms, on the other hand, reach exact solutions without providing a physical justification for the exact decision. This study does not discourage researchers from using hard clustering algorithms but suggests that the classification output from hard clustering might be over-simplified at the expense of the rare patterns; thus it is hypothesized that hard clustering might work best for regime classifications - where only the dominant types are of interest - compared to circulation typing.

5 Conclusions

This study examined the separability of CTs classified in Africa, south of the equator, using obliquely rotated T-mode PCA, and the uncertainty introduced on (i) the probability of occurrence of the CTs, (ii) the mean shape of the CTs, (iii) trend in the annual frequency of occurrence of the CTs, (iv) frequency distribution of the CTs, by using threshold values within the range of 0.2–0.35 to assign days to a given CT. The result showed that (i) generally, the CTs are satisfactorily pair-wise separated except with the CTs designated as mean patterns, (ii) the overall structure in the probability of occurrence of the CTs, the mean shape of the CTs, and the year-to-year variations in the annual occurrence of the CTs are not significantly affected by the choice of the threshold values within the defined range, (iii) the frequency distribution and the statistical significance of the trend in the annual occurrence of the CTs might be

Loading [MathJax]/jax/output/CommonHTML/jax.js

reshold. A more stringent threshold value (e.g. T0.35) further

limits the probability of group membership and the number of days assigned to a given class but might have the advantage of revealing the centers of action during the active state of a CT, when composite maps are analyzed. Relatively, using low threshold values (e.g. T0.2) allows the probability of group membership, and allows a relatively more accurate analysis of the time series of the CTs but might limit the extent to which all the centers of action under the active state of a given CT are detected when composite maps are analyzed.

Declaration

Conflict of interest: There are no conflicts of interest in this paper

Funding statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Barimalala, R., R. C. Blamex, F. Desbiolles, C. J. C. Reason, 2019: Variability in the Mozambique Channel Trough and impacts on Southeast African rainfall. J. Clim., 2, 749-765

Barreira, S., R. H. Compagnucci, 2011: Spatial fields of Antarctic sea-ice concentration anomalies for summer–autumn and their relationship to Southern Hemisphere atmospheric circulation during the period 1979–2009. Ann. Glaciol. 57, 140-150

Bartholv, J., A. G. Barnston, R. E. Livezey, 1987: The use of cluster analysis and rotated empirical orthogonal function analysis in describing the macrocirculation pattern of the northern hemisphere and the Atlantic-European region on different heights. In Preprints 3rd International Meeting on Statistical Climatology, Vienna, 23–27

Bartzokas, A., D. A. Metaxas, 1996: Northern Hemisphere gross circulation types. Climatic change and temperature distribution. Meteorologische Zeitschrift. 5, 99-109

Beck, C., Philipp A., 2010: Evaluation and comparison of circulation type classifications of the European domain. Phys. Chem. Earth, Parts A/B/C, 374-387

Compagnucci, R. H., N. E. Ruiz, 1992: On the Interpretation of Principal Component Analysis applied to meteorological data. In Proceedings of the Fifth International Meeting on Statistical Climatology. Atmospheric Environmental Service of Canada: Toronto; 241–244.

Compagnucci, R. H., D. Araneo, P. O. Canziani, 2001: Principal sequence pattern analysis: a new approach to classifying the evolution of atmospheric systems. Int. J. Climatol. 21, 197–217

Compagnucci R. H., M. B. Richman, 2008: Can principal component analysis provide atmospheric circulation or teleconnection patterns? Int. J. Climatol. 6, 703–726

Loading [MathJax]/jax/output/CommonHTML/jax.js

Cook, K. H., 2000: The South Indian Convergence Zone and Interannual Rainfall Variability over Southern Africa. J. Clim. 13 21, 3789–3804

Dedekind, Z., F. A. Engelbrecht, J. Merwe, 2016: Model simulations of rainfall over southern Africa and its eastern escarpment. Water SA 1, 129

Gong, X., M. B., Richman, 1995: On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies. J. Clim. 4, 897-931

Gordon, A.L., 1985: Indian-Atlantic transfer of thermocline water at the Agulhas Retroflection. Science, 227, 1030-1033.

Harr, A. P., R. L. Elsberry, 1995: Large-scale circulation variability over the Tropical Western North Pacific. Part 1: spatial patterns and tropical cyclone characteristics. Mon. Wea. Rev., 5, 1225–1246.

Hersbach, H., et al, 2020: The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 730, 1999-2049

Hewitson, B. C., R. G. Cran, 2002: Self organizing maps: applications to synoptic climatology. Clim. Res., 22, 13–26.

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

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

Huth, R., P. Canziani, 2003: Classification of hemispheric monthly mean stratospheric potential vorticity fields. Ann. Geophys. 21, 808-817.

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

Ibebuchi, C. C. 2021a: On the combination of rotated principal component analysis regionalization technique and linear regression in seasonal rainfall prediction. Preprint, available at Research Square. Viewed 1 February 2021, DOI: 10.21203/rs.3.rs-164806/v1

Ibebuchi, C. C., 2021b: Circulation pattern control of wet days and dry days in Free State, South Africa. Preprint, available at Research Square. Viewed 1 February 2021, DOI: 10.21203/rs.3.rs-160597/v1

Ibebuchi, C. C., 2021c: Can synoptic patterns influence the track and formation of tropical cyclones in the Mozambique Channel? Preprint, available at Research Square. Viewed 19 February 2021, DOI: 10.21203/rs.3.rs-200536/v1

Loading [MathJax]/jax/output/CommonHTML/jax.js

Ibebuchi, C. C., 2021d: Synoptic situations in Africa south of the equator linked to wet events in Namibia; a case study with the February 2008 flood episode. Preprint, available at Research Square. Viewed 1 May 2021, DOI: 10.21203/rs.3.rs-368816/v1

Ibebuchi, C. C., 2021e: Circulation patterns linked to extreme wet and dry conditions in Mozambique, relationship with climatic modes, and changes since 1961. Preprint, available at Research Square. Viewed 1 May 2021. DOI: 10.21203/rs.3.rs-462878/v1.

IPCC, Climate Change, 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker T. F., D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels,Y. Xia, V. Bex, and P. M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp, https://doi:10.1017/CBO9781107415324

Kendall, M. G., 1975: Rank Correlation Methods. Griffin, London, UK

Kidson, J. W., 1997: The utility of surface and upper air data in synoptic climatological specification of surface climatic variables. Int. J. Climatol. 4, 399–414

Mann, H. B., 1945: Non-parametric tests against trend. Econometrica 3, 245-259

Mo, K., M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. J. Geophys. Res. Atmos. 93

Molteni, F., S. Tibaldi, T. Palmer, 1990: Regimes in the wintertime circulation over northern extratropics. I: observational evidence. Q. J. R.Meteorol. 116, 21–67

North, G., T. Bell, F. R. Cahalan, F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. Mon. Wea. Rev. 7, 699-706.

Philipp, A., J. Bartholy, C. Beck, M. Erpicum, P. Esteban, X. Fettweis et al., 2010: Cost733cat – A database of weather and circulation type classifications. Phys. Chem. Earth, 35, 360–373

Philippopoulos, K., D. Deligiorgi, G. Kouroupetroglou, 2014: Performance Comparison of Self-Organizing Maps and k-means Clustering Techniques for Atmospheric Circulation Classification. Int. J. Energy Environ., 8, 171-180

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

Kim, M., C. Yoo, M. Sung, S. Lee, 2021: Classification of Wintertime Atmospheric Teleconnection Patterns in the Northern Hemisphere. J. Clim. 34, 1847–1861

Ndarana, T., M. Bopape, D. Waugh, L. Dyson, 2018: The influence of lower stratosphere on ridging Atlantic Ocean Anticyclone over South Africa. J. Clim. 15, 6175-6187

Loading [MathJax]/jax/output/CommonHTML/jax.js

Reason, C. J. C., H. Mulenga, 1999: Relationships between South African rainfall and SST anomalies in the southwest Indian Ocean. Int. J. Climatol. 19, 1651–1673.

Reason, C. J. C., M. Rouault, 2005: Links between the Antarctic Oscillation and winter rainfall over western South Africa. Geo. Phys. Res. Lett. 32, 7.

Reason, C. J. C., S. Smart, 2015: Tropical Southeast Atlantic warm events and associated rainfall anomalies over Southern Africa. Front. Environ. Sci. 3, 24.

Richman, M. B., P. J. Lamb, 1985: Climatic pattern analysis of three and seven-day summer rainfall in the Central United States: some methodological considerations and regionalization. J. Climate Appl. Meteor., 12, 1325–1343

Richman, M. B., 1986: Rotation of principal components. J. Climatol. 3, 293-335.

Richman, M. B., X. Gong, 1999: Relationships between the definition of the hyperplane width to the fidelity of principal component loadings patterns. J. Clim. 6, 1557–1576.

Stryhal, J., R. Huth, 2017: Classifications of Winter Euro-Atlantic Circulation Patterns: An Intercomparison of Five Atmospheric Reanalyses. J. Clim., 30, 7847-7861

Vargas, W. M., R. H. Compagnucci, 1983: Methodological aspects of principal component analysis in meteorological fields [preprints, Second International Conference on Statistical Climatology]. National Institute of Meteorology and Geophysics, Lisbon. pp. 531-539

Xu, D., Y. A. Tian, 2015: Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci., 2, 165–193

Yarnal, B. 1993: Synoptic Climatology in En6ironmental Analysis. Belhaven Press: London

Zhao, Z., H. Xi, A. Russo, H. Du, Y. Gong, J. Xiang, Z. Zhou, J. Zhang, C. Li, C. Zhou, 2019: The Influence of Multi-Scale Atmospheric Circulation on Severe Haze Events in Autumn and Winter in Shanghai, China. Sustainability. 11, 5979

Tables

Table 1: Range and median value of the loadings for each phase of the retained components

Loading [MathJax]/jax/output/CommonHTML/jax.js

CT	Maximum	Median	Minimum
1+	1	0.53	0
1-	-0.9	-0.2	0
2+	0.8	0.18	0
2-	-1	-0.46	0
3+	0.7	0.16	0
3-	-1	-0.39	0
4+	0.91	0.22	0
4-	-0.81	-0.15	0
5+	0.75	0.13	0
5-	-0.9	-0.13	0
6+	0.84	0.11	0
6-	-0.51	-0.08	0
7+	0.66	0.1	0
7-	-0.7	-0.1	0
8+	0.7	0.08	0
8-	-0.46	-0.06	0
9+	0.55	0.08	0
9-	-0.55	-0.07	0

Table 2: Spatial correlation coefficients between the CTs as obtained from T0.2 and the counterparts as obtained from the other threshold values

Loading [MathJax]/jax/output/CommonHTML/jax.js

CT0.2	T0.25	T0.3	T0.35
1	0.99	0.99	0.99
2	0.99	0.99	0.99
3	0.99	0.99	0.99
4	0.99	0.99	0.99
5	0.99	0.99	0.99
6	0.99	0.99	0.99
7	0.99	0.99	0.98
8	0.99	0.99	0.98
9	0.99	0.99	0.97
10	0.99	0.99	0.97
11	0.99	0.98	0.95
12	0.99	0.97	0.93
13	0.99	0.99	0.97
14	0.99	0.98	0.95
15	0.99	0.96	0.93
16	0.99	0.95	0.88
17	0.99	0.97	0.95
18	0.99	0.97	0.94

Figures

Loading [MathJax]/jax/output/CommonHTML/jax.js

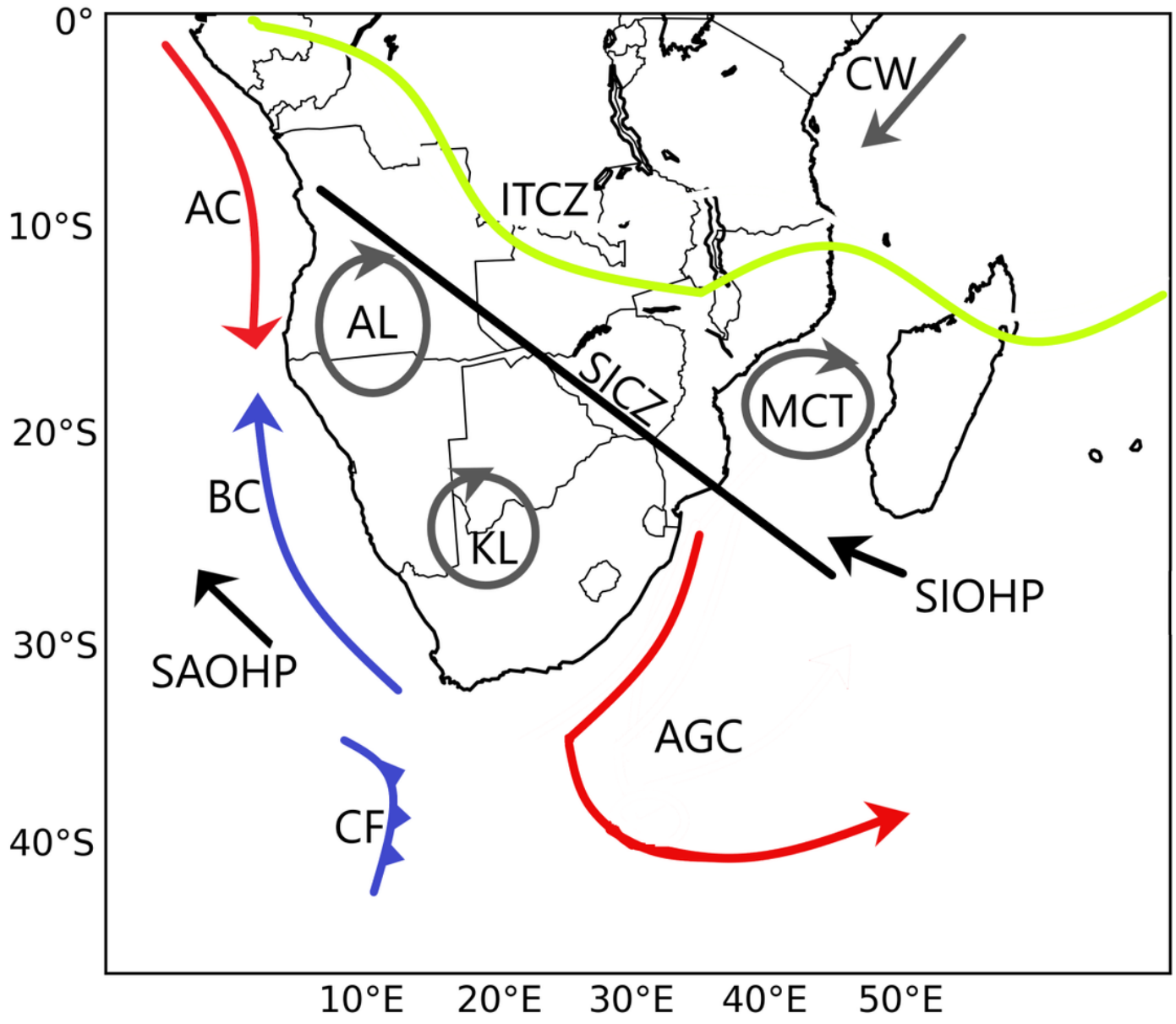

Figure 1

Idealized locations of synoptic rain-bearing systems in the study region during their active periods. CW: cross-equatorial northeast trade wind; MCT: Mozambique Channel Trough (it is associated with cyclonic circulation); SIOHP: western branch of the South Indian Ocean high-pressure (southeast moist winds deflected southward by the Madagascar topography penetrates southern Africa through its anticyclonic circulation); AGC: Agulhas warm current (it runs from 27°S to 40°S, retroflecting at about 21°E); CF: cold fronts (it moves from west to east when the mid-latitude cyclones track northward); SAOHP: South Atlantic Ocean high-pressure (it is associated with anticyclonic circulation driving moisture offshore); AC: Angola warm current; BC: Benguela cold current; ITCZ: Inter-tropical convergence Zone during its southward track in austral summer; KL: Kalahari low (it is associated with cyclonic circulation); AL: Angola low (it is associated with cyclonic circulation); SICZ: South Indian Ocean Convergence Zone

Loading [MathJax]/jax/output/CommonHTML/jax.js

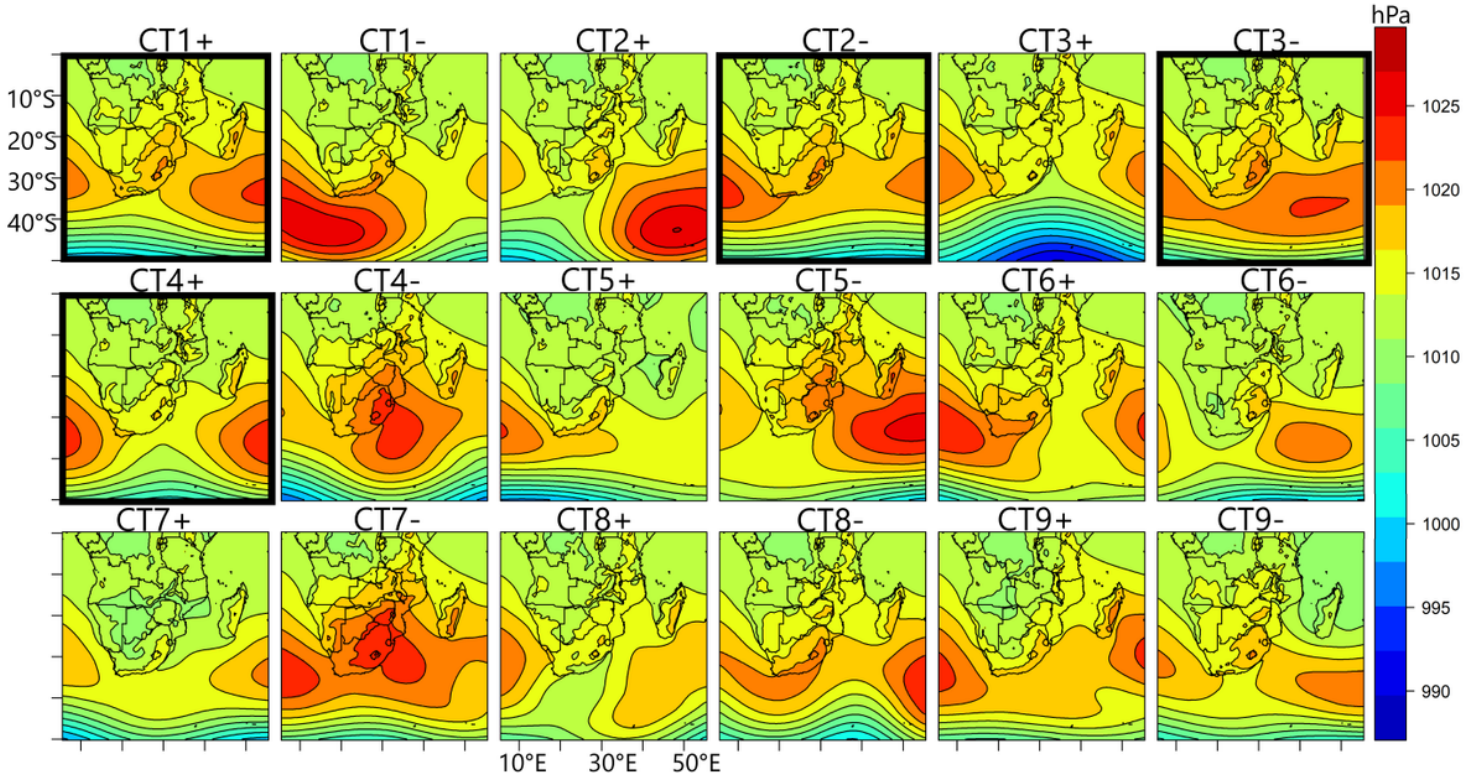

Figure 2

Circulation types in Africa, south of the equator

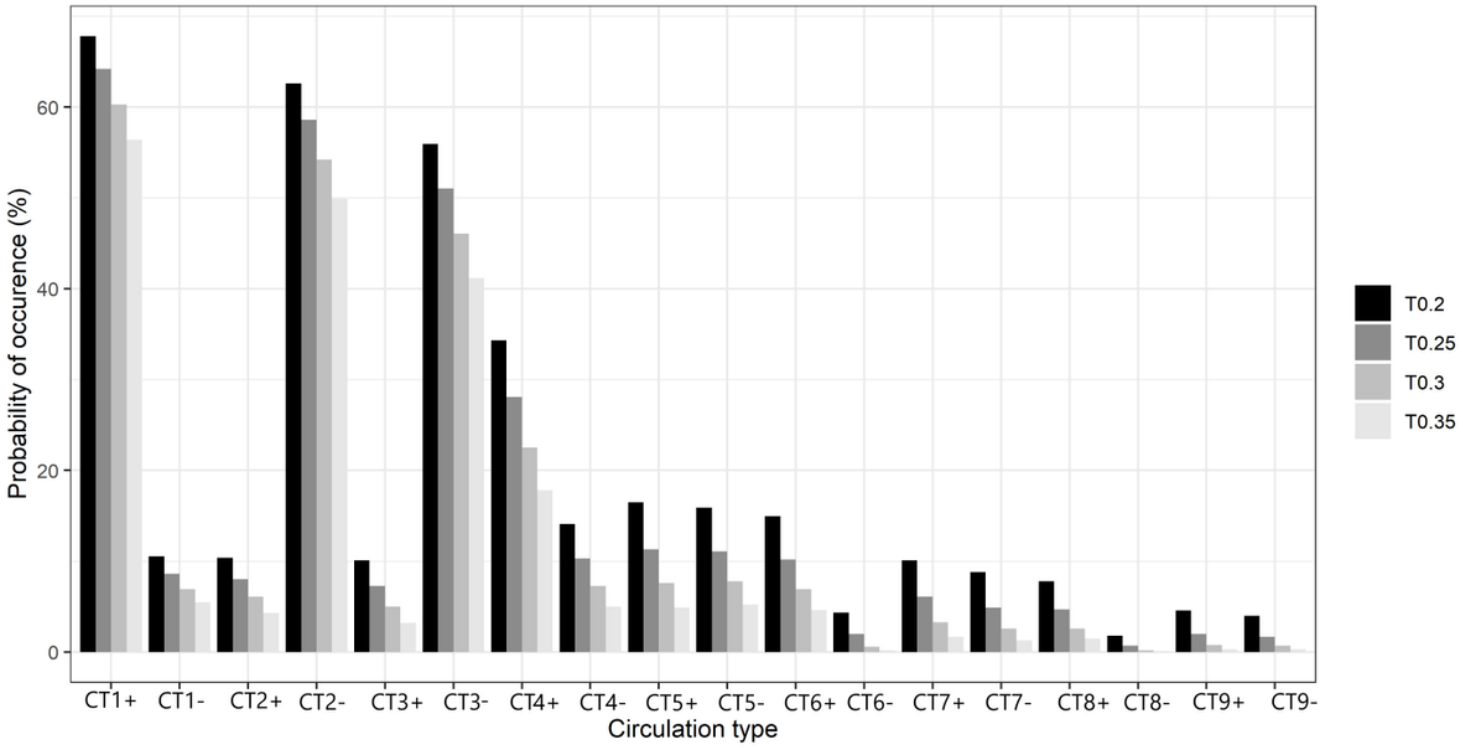

Figure 3

Loading [MathJax]/jax/output/CommonHTML/jax.js

Probability of occurrence of the CTs under the different threshold values

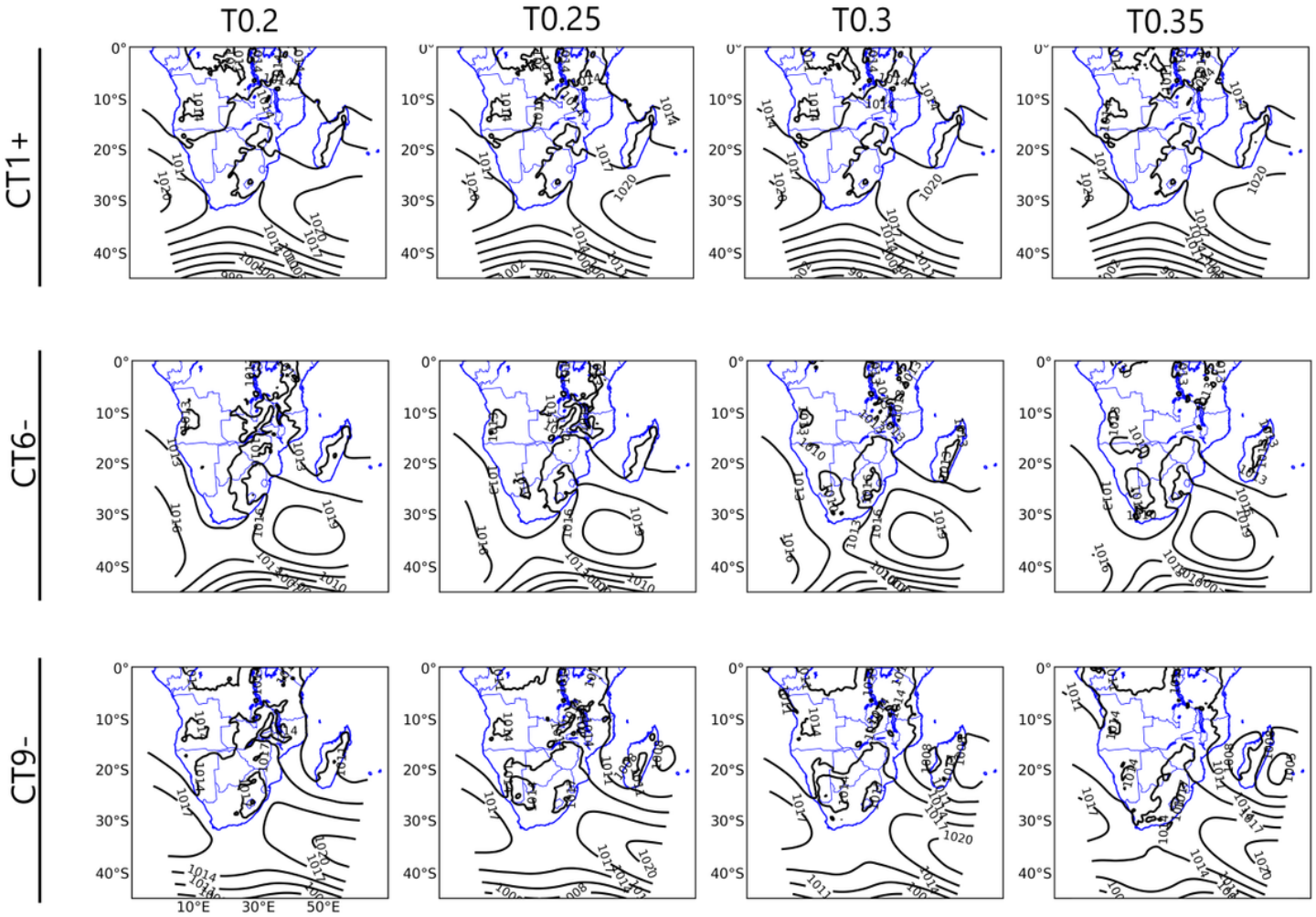

Figure 4

Mean shape of the CTs under the different threshold values exemplified from CT1+, CT6- and CT9-

Loading [MathJax]/jax/output/CommonHTML/jax.js

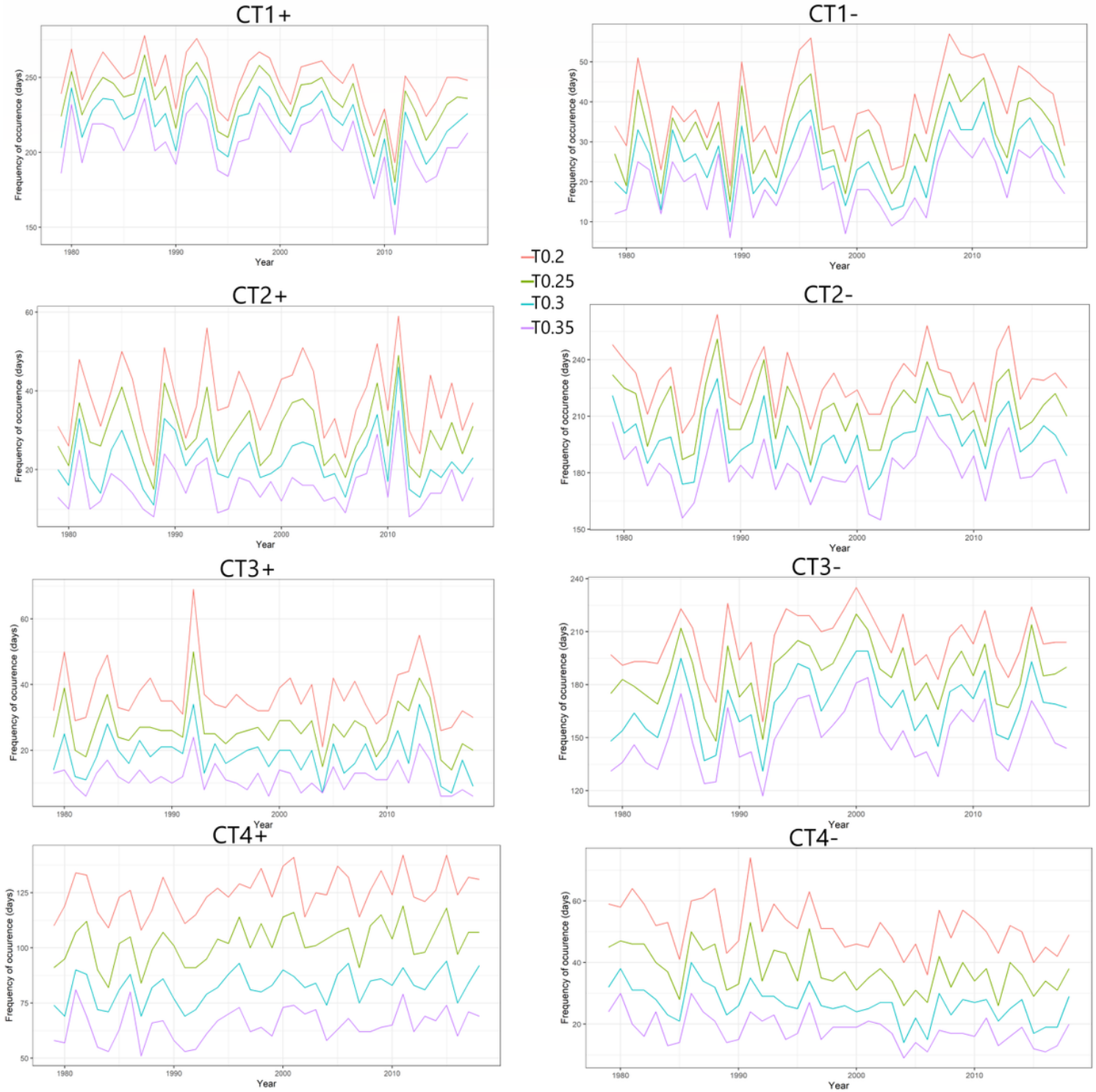

Figure 5

Annual frequency of occurrence of the CTs for the 1979-2018 period

Loading [MathJax]/jax/output/CommonHTML/jax.js

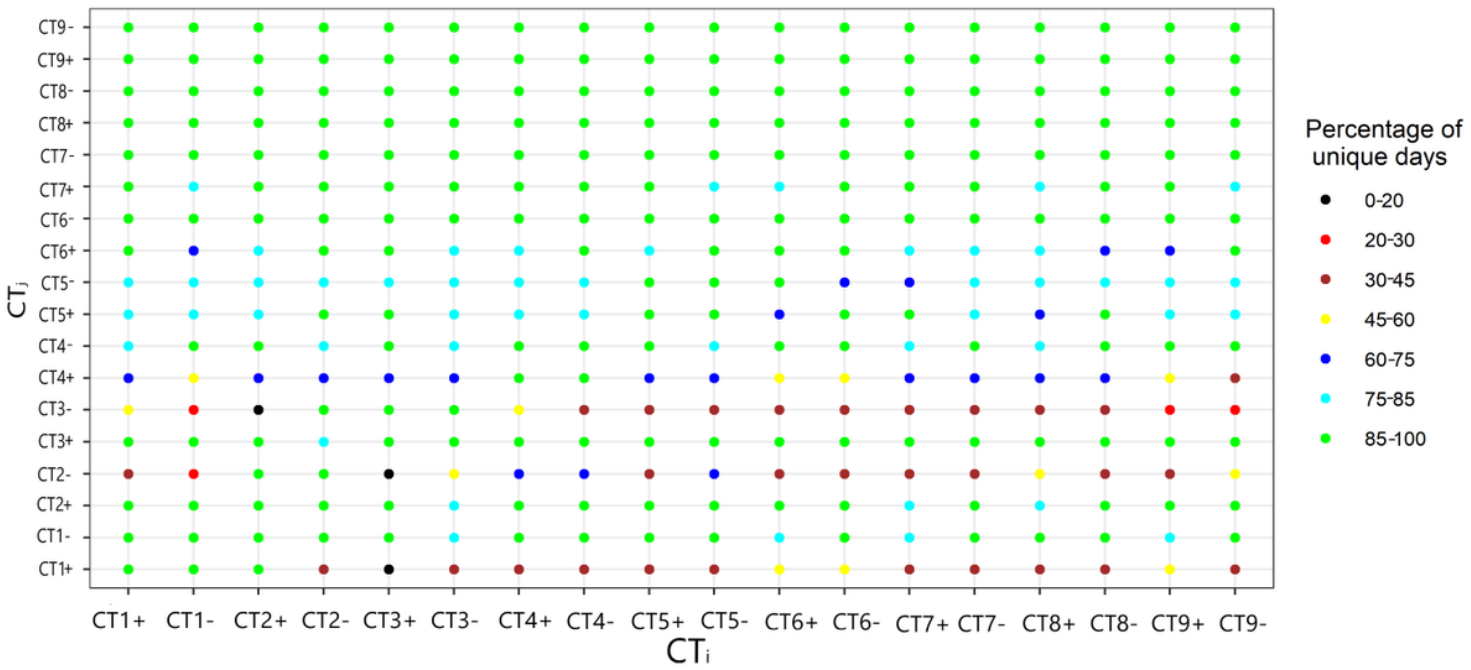

Figure 6

Measure of pair-wise separability between the 18 circulation types

Loading [MathJax]/jax/output/CommonHTML/jax.js