Università degli Studi di Padova

Dipartimento di Fisica e Astronomia Galileo Galilei

Project Report

Unsupervised Clustering Methods for Meteorological European Configurations

Author: Eugenia Boccanera
Supervisor: Francesco Domenichini, Federico Chiariotti

26/09/2025

1 Abstract

The European Atmospheric Configurations arise from the movement of air masses across different latitudes that interact with each other and have typical recurrent patterns. The observed weather phenomena are closely related to these characteristic configurations. Therefore, identifying those atmospheric patterns is crucial to performing weather forecasts and climate analysis.
Unsupervised clustering approaches offer a strong tool for this identification. This project aims to identify, develop, and validate an unsupervised clustering method in order to recognize these meteorological recurring configurations. Furthermore, the project aims to discover the most suitable variables that will yield the best results. Moreover, another purpose is to verify whether any configuration is increasing or decreasing in frequency over the years, highlighting an evolution in the characteristic structures of the climate due to climate change.
The dataset was sourced from the ERA5 reanalysis system from 1961 to 2020. The variable analyzed for this study was the geopotential measured at a pressure altitude of 500 hPa and the temperature at a pressure altitude of 850 hPa. Data were taken on a 0.25 ° x 0.25 ° grid across Europe.

Before starting with the analysis, the data have been preprocessed through standardization and Principal Component Analysis in order to reduce the dimensionality and to mitigate excessive noise in the data. Once the data were prepared, the K-means clustering algorithm was applied, and to evaluate the quality of the analysis, different coefficients and methods were utilized, including the elbow method, silhouette and percentage inertia reduction.
The best result of the study is a classification of data into 7 clusters representing European weather regimes. we then analyzed the individual decades and the year-by-year evolution of the atmospheric configurations identified by k-means, identifying strong increases or decreases in some of them, very probably caused by climate change.

2 Type of Data

2.1 Atmospheric Introduction

The Earth's atmosphere is a very complex fluid system characterized by continuous pressure and temperature variations and by the circulation of air masses. The atmosphere is subdivided into five main temperature-based layers: the troposphere, stratosphere, mesosphere, thermosphere, and exosphere. All meteorological phenomena occur in the lowest layer: the troposphere, which extends from 0 to almost 12 km.

There are two significant pressure arrangements possible in the air: air pressure depends on the density of air molecules and air temperature. In low-pressure systems, air is rising and the water vapor it contains is condensed to form clouds, typically precipitation, and cyclonic whirls that move in a counterclockwise direction (in the Northern Hemisphere). Conversely, high-pressure systems consist of winds being pushed away from the center of high pressure with clockwise rotation in our hemisphere, showing an anticyclonic flow, with largely clear skies. The alternation and interaction of anticyclones and cyclones regulate the large-scale atmospheric circulation.

The classification of atmospheric circulation was applied for the first time in meteorology for weather forecasting. Thanks to computer development, it is now possible to deal with large databases, and so classification methods have become a significant part of statistical and synoptic climatology [14]. Cluster analysis has the main goal of distinguishing distinct groups of atmospheric circulation states from observations, reanalyses, and predictions [15].

2.2 ERA5 Reanalysis Dataset

Based on the results on clustering quality and computational efficiency, and due to the fact that most of the studies rely on z_{500} [14], we select z_{500} and t_{850} as end variables for our final results. Using both of them we have accurate informations regarding the atmospheric structure vertical and horizontal, which allows more thorough analysis of the large-scale circulation and clustering pattern.

Geopotential height is a measure that represents the height of a specific pressure surface in the atmosphere

above mean sea level and is a measure of a block of air's potential energy per unit of mass. It is applied very diffusely across atmospheric studies because it provides a true picture of big-scale circulation patterns and there is a direct relationship with the mass and pressure distribution of the troposphere [3]. Air temperature, however, is a root atmospheric variable that expresses the condition of warmth in the air. Density, pressure, gradients and stability within the air are dependent upon it, and it plays an important role in forcing circulation regimes.

The observations used in our analysis are taken from ERA5. ERA5 is the 5th generation atmospheric reanalysis of the Copernicus Climate Change Service (C3S) of the European Centre for Medium-Range Weather Forecasts (ECMWF). Covers the period January 1940 to date. In this work, we take into account the period 1961-2020. This time period includes the two standard climatological reference 30-year periods, is sufficiently long to record interannual variability, and at the same time is computationally feasible for analysis. ERA5 makes available hourly estimates of a wide range of atmospheric, land, and oceanic variables.

Our study is focused on Europe, i.e., the latitude range is 70°N and 20°S, and the longitude range is 40°E and 40°W. The study region was divided into a rectangular grid at a spatial resolution of 0.25 ° x 0.25 °, which corresponds to 201 latitude points and 321 longitude points, or a total of 64,521 points in space. For each day, only a single map at 00:00 UTC was considered for analysis. Spatial and temporal resolution is needed for the definition of circulation types, together with the variables on which they depend. We also attempt to clarify what the best representative and appropriate combination of variables is. Hence, we considered different sets of variables for analysis.

2.3 Feature selection

According to [14] the predominant number of atmospheric experiments (approximately 84%) are based on values of sea level pressure, geopotential heights or wind fields. In particular, we selected the geopotential height at pressure levels of 850 hPa, 500 hPa, 250 hPa ($z_{850}, z_{500}, z_{250}$), as well as the temperature at 850 hPa (t_{850}). We perform the following groupings for analysis and clustering: $z_{850,500,250}$; $z_{850,500}$; z_{500} with t_{850}; and $z_{250,500,850}$ with t_{850}.

Based on the results on clustering quality and computational efficiency, and because most studies rely on z_{500} [14], we select z_{500} and t_{850} as end variables for our final results. Using both of them, we have accurate information regarding the atmospheric structure, vertical and horizontal, which allows more thorough analysis of the large-scale circulation and clustering pattern.

The geopotential height is a value that represents the height of a given pressure surface in the atmosphere with respect to the mean sea level and reflects the potential energy per unit mass of an air block. Its use is very diffuse in atmospheric studies because it provides a clear representation of large-scale circulation patterns and is directly related to the distribution of mass and pressure in the troposphere [3].

However, air temperature is an atmospheric variable that explains the condition of heat in the air. Its density, pressure, gradients, and stability depend on it and play an important role in forcing circulation regimes.

3 Analysis

3.1 Preprocessing

The rough data are organized this way: each variable (geopotential and temperature) has a 4D structure (time, pressure, latitude, longitude). Before the analysis, it is essential to transform the 4D structure into a 2D structure (sample x feature)[7]. This remapping of the data is fundamental because the Principal Component Analysis is defined over a bidimensional matrix n x p, where n is the number of samples and p the number of variables or features. In our case, the samples correspond to the days, while the variables correspond to all possible combinations of pressure, latitude, and longitude. In this way, it is possible to compute the variance and the correlation between the variables. Moreover, for clustering, a bidimensional structure is needed.

Considering the 60-year period and focusing on the two variables, geopotential z and temperature t, we first filtered the entire dataset to select only the geopotential at the 500 hPa isobaric level and the temperature at the 850 hPa isobaric level. Then, for each variable, all non-temporal dimensions are combined and flattened into a single feature axis, so the final shape of the matrix is: (21915, 129042).

To conclude the preprocessing, it's also necessary to normalize the data. As explained in [7], the goal of PCA is to maximize variance, and variance is sensitive to the unit of measurement; thus, if the variables have different scales (such as °C for temperature and hPa for geopotential), those with larger numerical ranges tend to dominate the first principal components. This is why normalization of each variable before applying PCA is necessary.

Many atmospheric variables have very long autocorrelation times; therefore, in our 60-year dataset, there are a limited number of statistically independent samples [9]. Because of that, we performed a temporal subsampling, selecting one day every five days, to reduce the temporal redundancy and the computational costs, but preserving the climate variability. We divided the sampled data set into two halves. The first is the part containing the temperatures, while the second contains the geopotential values. We normalized each half; in particular, each data value x_{ij} is centered on the global mean μ and divided by the global standard deviation σ:

$$X_{ij}^{(\mathrm{std})} = \frac{X_{ij} - \mu}{\sigma}$$

Both quantities μ and σ are computed for both parts of the dataset. This separate standardization approach ensures balanced contribution of both variables to the principal component analysis while maintaining physical interpretability.
Then we recreate a unique dataset by reuniting the two halves, and we save this normalized subsampled dataset. After this step, our goal is to reconstruct a normalized and dimensionally reduced dataset that includes all daily records, without any temporal sub-sampling. Therefore, we divided the dataset into six decades: 1961-1970, 1971-1980, 1981-1990, 1991-2000, 2001-2010, and 2011-2020. Then we loaded

one decade at a time and normalized it with the same procedure as adopted for the sub-sampled dataset.

3.2 Pca

The dimension of the dataset is very large, even if the dataset has been sampled:(4320, 129042). It is computationally expensive to use the clustering algorithm directly. Therefore, the data set's dimensionality must be reduced. Principal Component Analysis (PCA), is the most popular method for this purpose in atmospheric science [9]. Additionally, it minimizes information loss while improving interpretability [7]. PCA creates new uncorrelated variables (Principal Components, PC) that capture as much of the variability (i.e. statistical information) in the data as possible. They are linear functions of the centered original variables and are obtained by solving the eigen-decomposition of the covariance matrix, which is a matrix that sums up the covariances between all pairs of variables. The explained variance is a measure to be used to understand the amount of total variability captured by each Principal Component.

For our analysis, we performed Incremental PCA (IPCA) over the sampled dataset to avoid memory issues. Instead of computing the entire covariance matrix, it processes the data in batches, making the whole process faster and more scalable. Once the loading matrix (or Empirical Orthogonal Functions, EOFs in climatology [9]) is computed, it is saved because it contains the projection vectors that, if multiplied by the normalized data, allow us to derive the Principal Components (PCs).
Analyzing the explained variance in the left panel in Figure 1, it is evident that already from the 15th principal component, the contributions of the following PC are no longer significant, and are extremely close to 0.0. However, we choose 20 principal components that allow us to reach more than 90% of the cumulative explained variance (in particular more than 92%), which is a good limit that allows us to reduce the dimensionality of the data set while still preserving most of the information contained in it. This is evident in the right panel in Figure 1.

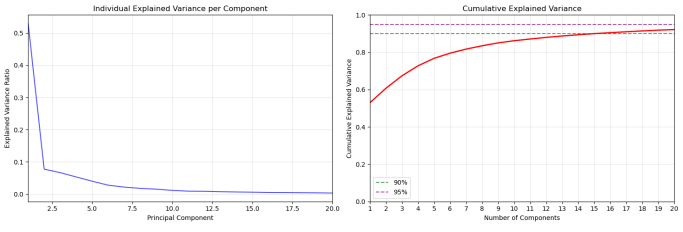

FIGURE 1: The left panel shows the explained variance for each principal component. The right panel shows the cumulative explained variance of principal components.

Now, as we make for the normalization, we multiply the loading matrix by each decade that has been normalized before. Once the six-10-year period has been normalized and reduced to 20 Principal Components, we unify them to obtain the entire 60-dataset normalized and reduced, ready for further analysis.

3.3 K-means clustering on principal components

Now, the shape of our dataset is: (21915, 20). To classify all of that data, we employed the k-means algorithm directly from the library: scikit-learn. This method is very diffuse because of its simple algorithm and its fast convergence [17]. The number of clusters k should be specified a priori, but in order to explore different clustering granularities, we pick k in a range of 2 to 50. The chosen initialization algorithm is k-means++. It selects initial cluster centroids that spread them out across the dataset. The first centroid is chosen randomly, and each subsequent centroid is determined with a probability inversely proportional to the distance squared from the next nearest centroid.
Additionally, the algorithm has been initialized with 20 initializations to ensure robustness and also to minimize the effect of the initial conditions on the outcomes. For each initialization, the algorithm is executed using different centroid seeds, and the resulting output is the best output among the different runs according to inertia.
The K-Means algorithm is an iterative distance-based clustering algorithm that assigns each observation to the nearest cluster centroid [16]. The algorithm procedure follows these steps:

1. Determine the number of clusters k and the maximum number of iterations.
2. Choose randomly the initial centroids.
3. For each point of the dataset, compute the Euclidean distance to each centroid and assign the point to the nearest cluster.
4. Recalculate the location of the centroids as the average of all points assigned to such a cluster.
5. If the location of any centroid is updated or the maximum number of iterations is not reached, then repeat step 3. Otherwise, terminate the algorithm and report final cluster assignments.

The objective of K-Means is to minimize the sum of squared distances $J = \sum_{i=1}^{n} \sum_{j=1}^{K} a_{ij}, |x_i - c_j|^2$, [16], between data points and their assigned cluster centroids, where n is the number of data points, k the number of clusters, x_i a data point, c_j the centroid of the cluster j, and a_{ij} is 1 if x_i belongs to the cluster j, 0 otherwise.

4 Evaluation of the clustering

The identification of the best value of the number of cluster k has been made through a systematic method that combines different statistical validation techniques, which have been applied to the k interval that goes from 2 to 50. The first approach is the Elbow Method, and then we also compute the silhouette coefficient and Calinski–Harabasz Index.

4.1 Elbow and Inertia Analysis

The Elbow Method is a classical method widely used in statistics to define the best number of clusters [17]. It uses the square of the distance between the sample points in each cluster and the centroid of the cluster. The sum of squared errors (SSE) or inertia (in the library scikit-learn) shows the inner compactness of the cluster. When the number of clusters becomes bigger than the number of real clusters, the inertia will continue to decrease but more slowly, showing

the typical elbow shape in the curve [17]. In our case, the inertia vs the number of cluster k is reported in the first upper panel in Figure 2. As expected, the decrease of the inertia is evident, but it is monotonous, and the identification of the "elbow" is not so obvious. This is due to the nature of the meteorological data, which are almost continuous and show only small variations between observations.

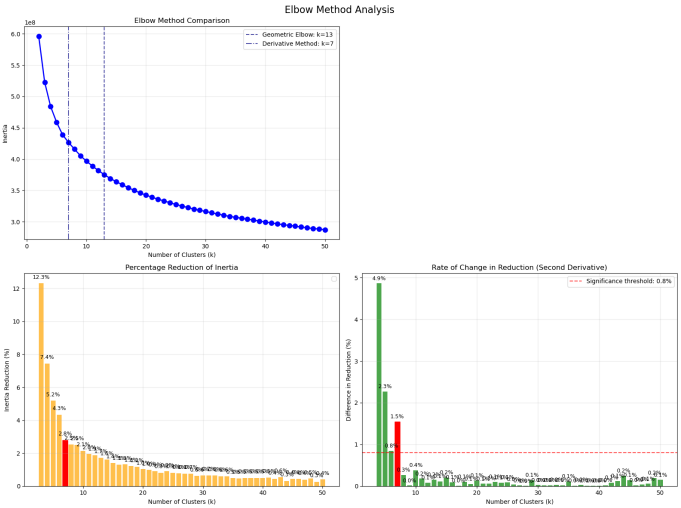

FIGURE 2: The curve of the inertia is plotted in the first upper panel in function of the number of cluster k (from 2 to 50) for the 60-year dataset. The "elbow" of the curve computed with the derivative method shows the point of transition from significant reduction to marginal improvement of the inertia, it suggest the optimal value for the number of cluster k. The lower panels show the analysis of the second derivative that confirms the identification of the point of maximum curvature.

To address this limitation, we applied a derivative-based method to objectively estimate the optimal number of clusters, which evaluates changes in the slope of the inertia curve. By detecting where the rate of decrease drops significantly, this method provides a more objective criterion for identifying the elbow and selecting the optimal number of clusters. In Figure 2 the blue dashed dotted line identifies the best number of clusters in $k = 7$.

The lower panels in Figure 2 display the analysis of the first and second derivative that confirms $k = 7$ as the point of maximum curvature. The left one shows the first derivative of the inertia curve, which evaluates the inertia reduction rate for each increment of k. The mathematical interpretation is that it measures the local slope of the inertia curve, giving information about the speed of clustering improvement. The first few columns have high values in modulus because they show a drastic reduction of inertia, and adding a cluster produces a relevant improvement. After $k = 7$, the values tend to reach zero, as the addition of more groups gives marginal improvements. The right panel visualizes the second derivative of the inertia curve and represents the acceleration of the change in the rate of inertia reduction. Peaks well defined in the right graph indicate a clear clustering structure, while plateaux or oscillations suggest ambiguity in the definition of the best k.
Those approaches provide a precise numeric method for the selection of k, reducing the subjective interpretation. From an interpretative point of view, $k = 7$ represents the best balance between the ability to identify distinct atmospheric patterns and the avoiding excessive fragmentation that could lead to incorrect classification and noise.

4.2 Silhouette Anlysis

The silhouette coefficient represents a measure of inner validation that quantifies the quality of clustering. It compares and evaluates the internal cohesion of the clusters (compactness) and simultaneously the separation between different clusters (resolution) [17]. For each point, the silhouette coefficient is in the range $-1 - 1$, and is defined as:

$$s = \frac{b - a}{max(a, b)},$$

where a is the mean distance intra-cluster and b is the mean distance from the nearest cluster. Positive values close to 1 indicate well-defined and well-separated clusters, while negative values suggest probable wrong assignments or overlaps between clusters.

The test has been performed for $k = 4, 5, 6, 7, 8, 9, 10, 11, 12$. It can be seen from the graphs displayed in Figure 3 that the mean silhouette values for all of them are relatively far from 1, yet positive. This is expected because meteorological data are continuous and complex, such that clusters, are not sharply defined. Even though, in our example, the key is to make sure that the silhouette values for chosen k should be higher than those corresponding to larger values of k, and it indeed is for $k = 7$.

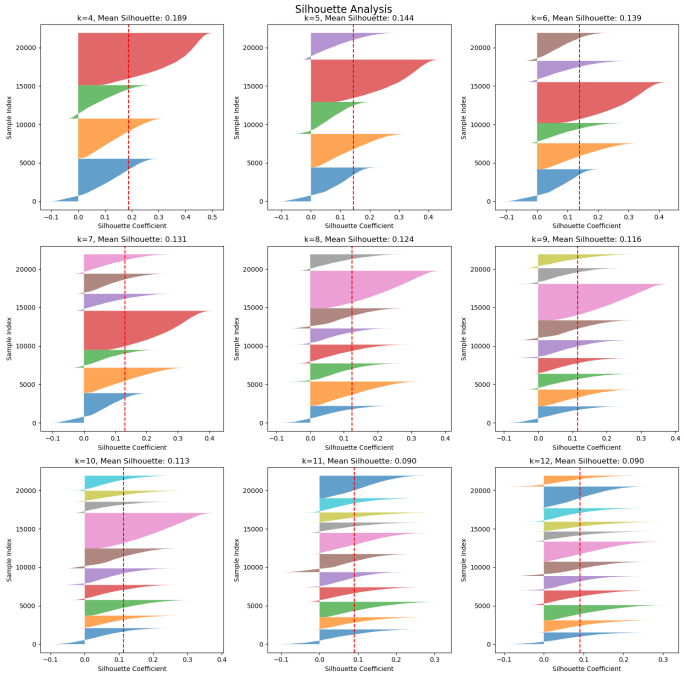

FIGURE 3: Mean silhouette coefficients for different values of k, the range is $(-1 - 1)$. Higher values represent higher quality clusters, with well separated and internally coherent clusters.

4.3 Calinski–Harabasz Index

For a better valuation of the clustering, the Calinski-Harabasz (CH) index has been computed. It measures the ratio between the variance inter and intra-cluster. Higher values highlight more compact groups and well separated, however, the index doesn't have an absolute threshold to be considered acceptable, but it needs to be compared with the other values for different k.
In the plot shown in Figure 4, the CH index reaches the highest value for k = 4 and then progressively decreases as k increases. This suggests that solutions with a small number of clusters are statistically more solid. Despite the fact that for k = 7, the CH index remains higher with respect to the values obtained with k larger than 7. This suggests that, even though the choice does not coincide with the theoretical best of the CH, raising the number of clusters from this point lowers quality in terms of this internal measure, confirming that k = 7 is a reasonable choice as a compromise.

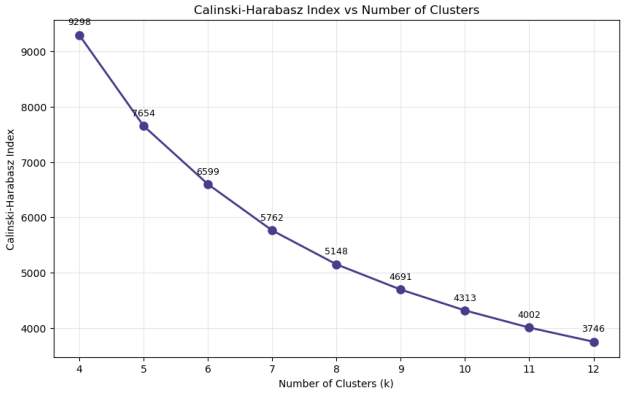

FIGURE 4: The plot shows the Calinski–Harabasz Index values for different numbers of clusters. The quality of clustering is higher for higher values of the index.

4.4 Final K choice

According to previous analysis, although the Silhouette and Calinski-Harabasz indices show that the separation of the clusters is relatively weak and maximized around $k = 4$, we have finally decided to choose $k = 7$. It is strongly supported by the elbow method (Sec. 4.1), and it represents the best compromise between the inertia percentage reduction and the interpretability of the clusters in the applicative and interpretative context.

This choice represents a general classification that is consistent with the major atmospheric patterns, which are typically classified in 4 or 5 groups in the literature [2] [8].
Now that the final k has been determined, we save the centroid vectors for further analysis.

5 Cluster Analysis for k= 7

In this section, we will study and analyze the centroids obtained with the previous analysis (Sec. 5.1), and we will provide a physical interpretation.
Moreover, we will assign each point of the six decades to the nearest centroids and examine the decadal frequency and the cluster frequency distribution (Sec. 5.3).

5.1 Identification of atmospheric circulation patterns (k= 7)

We implemented a physical reconstruction methodology to transform the vectors of the centroids in the PCA space into interpretable meteorological patterns. This step allows the meteorological interpretation of each identified regime and provides physically meaningful atmospheric configurations for climate analysis.
We first reverse the PCA transformation using the pre-trained Incremental PCA model, converting the 20-dimensional centroids back to the original high-dimensional space (7, 129042). After that, we implemented separate de-standardization of temperature and geopotential variables, reversing the preprocessing approach. The destandardized geopotential is also converted from m^2/s^2 to decameters (dam) through division by the gravitational acceleration factor (9.80665) and an additional factor of 10, ensuring compatibility with standard meteorological conventions. This conversion is essential for a meaningful interpretation of the geopotential.
The linear feature vectors for each centroid are reshaped back to their original $201x321$ grid configuration, reconstructing the bidimensional map over Europe.

Now, we show the physical map of each centroid associated with the monthly frequencies of occurrence for each atmospheric regime (cluster) over the six decades under study.

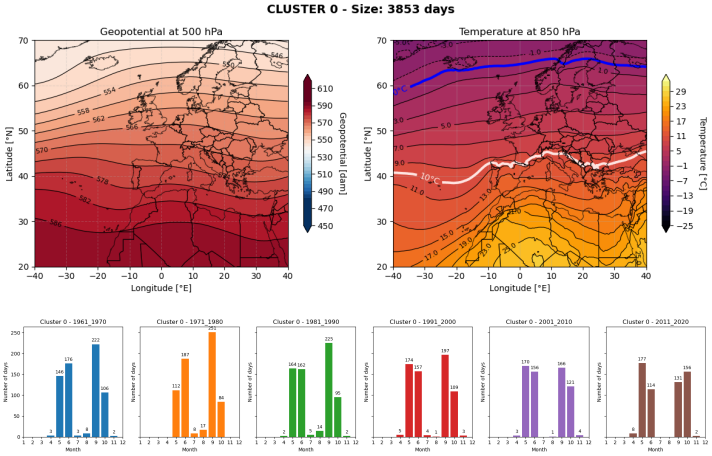

FIGURE 5: a) Centroid number 0, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

The meteorological maps that illustrate the *Cluster*0 in Figure 5 are associated with warm phases typical of the transitional months and with occasional occurrences at the edge of summer. It is characterized by a zonal flow and generally moderate to warm conditions. Over time, its occurrence shows a moderate general decline, as it is progressively dominated by warmer summer regimes from June to September. It is, however, more frequently observed in May and October.

The configurations of *Cluster*1, represented in Figure 6, are associated with winter conditions and with colder periods of the transitional seasons. It is characterized by a zonal flow and generally cool climate, often accompanied by rapid Atlantic disturbances. Over time, no significant trend of change is observed in its frequency.

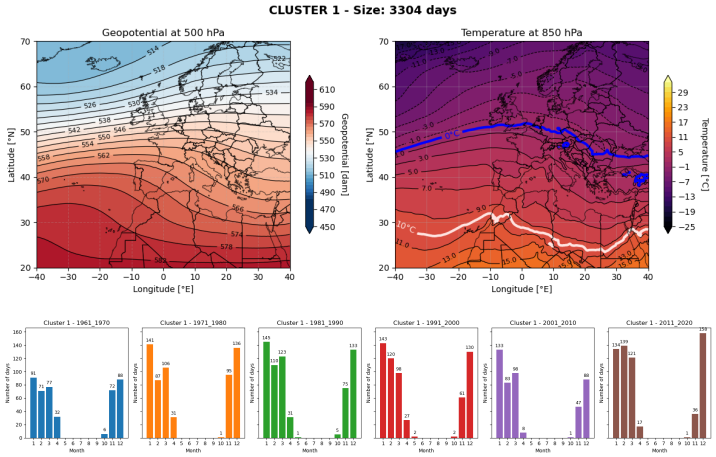

FIGURE 6: a) Centroid number 1, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

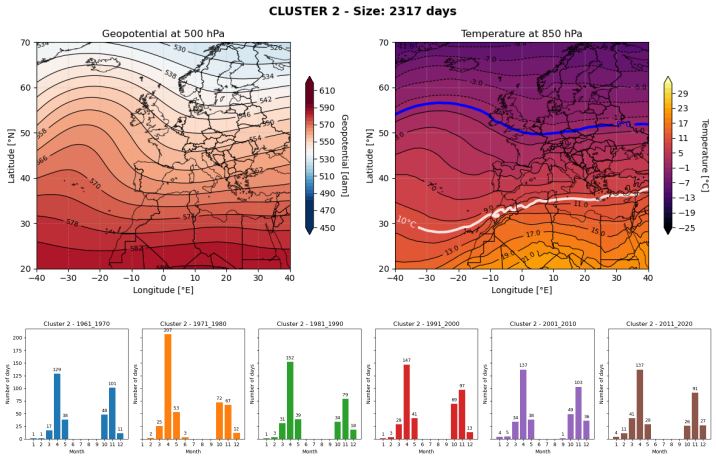

FIGURE 7: a) Centroid number 2, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

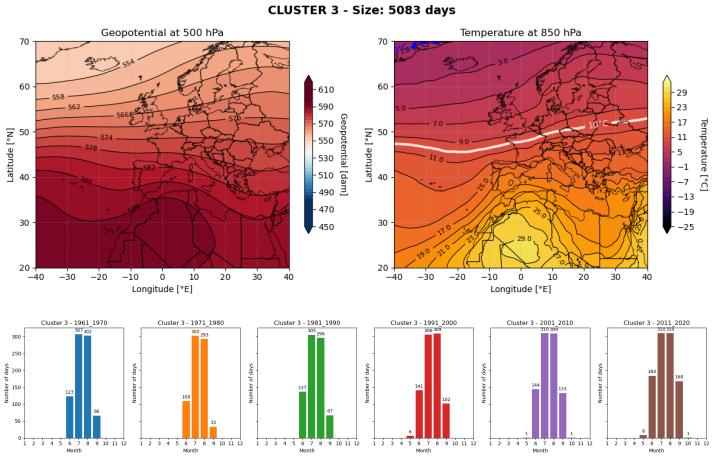

FIGURE 8: a) Centroid number 3, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

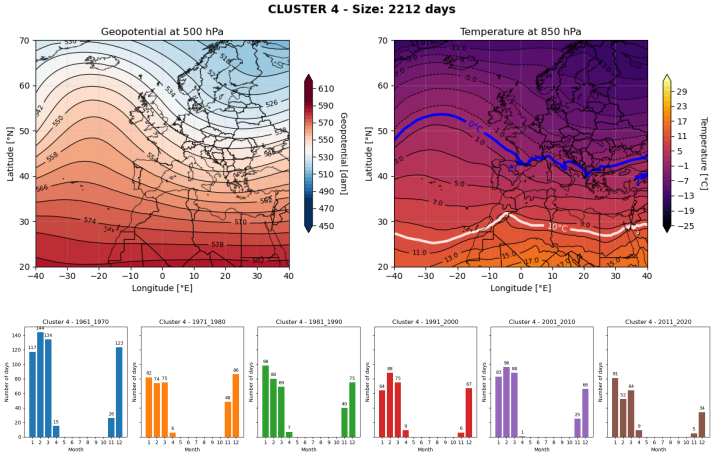

FIGURE 9: a) Centroid number 4, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

The weather maps of *Cluster*2 present in Figure 7 correspond to transitional-season conditions, with disturbances approaching Europe from the north and northwest. It is linked to a generally cool climate and has no visible long-term trend, but does have a tendency to happen more frequently in winter. From the 2000s, there has been an increase in interannual variability.

In Figure 8 (*Cluster*3) the configuration is obviously summer-like with hot and stable conditions. It is dominated by African anticyclones, with a core temperature field of around 29 °C over the central–western Sahara. Over time, it has increasingly extended into the late-summer months, becoming much more frequent.

For the *Cluster*4 in Figure 9 instead, the synoptic charts correspond to winter conditions, with disturbances entering Europe from the north and northwest and bringing cold to very cold air outbreaks. It represents a typical winter ATR (Atlantic Ridge) pattern. Its frequency undergoes a significant decline, being significantly less frequent at the beginning of winter, especially in December, and is restricted considerably to the mid and late-winter periods. This is consistent with the significant decline of snowfall observed in most regions.

The configuration present in Figure 10, referring to *Cluster*5, is typical of the transitional seasons, with disturbances reaching Europe from the west or northwest. It is associated with a generally temperate climate and increased rainfall over central–western Europe. Its frequency has risen markedly in November, while the already rare cases observed in September have nearly disappeared.

In the end, for *Cluster*6 (Figure 11), the weather charts can be associated with winter conditions, characterized by an Atlantic ridge extending into western Europe. It often includes Scandinavian blocking episodes and more eastward extensions of the Siberian High. Its frequency shows a slight decline, which appears only weakly related to large-scale circulation changes.

5.2 Temporal evolution of cluster compactness

A quick analysis of the compactness of the various clusters in each decade was also conducted on. The comparison between the various clusters is displayed in Figure 12. The

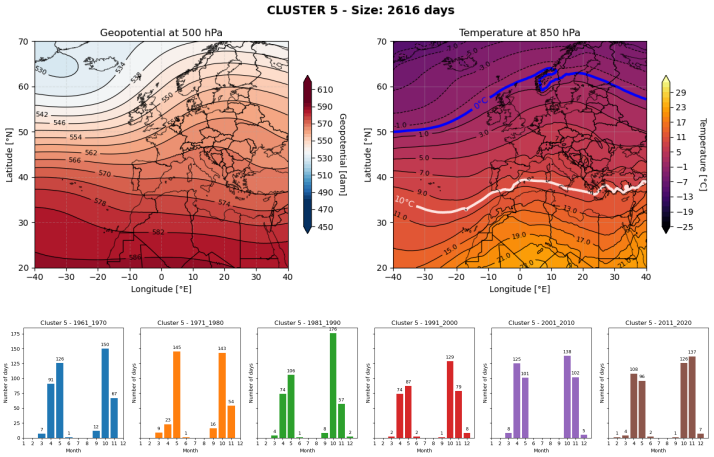

FIGURE 10: a) Centroid number 5, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

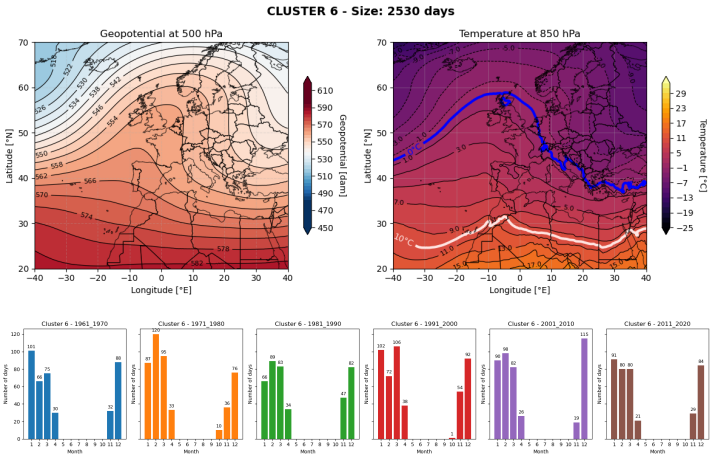

FIGURE 11: a) Centroid number 5, on the left panel, is plotted the geopotential (dam), on the right one is plotted the temperature (°C). b) Monthly frequency of occurrence for each cluster over the six decades

vertical axis represents the average distance from the centroid; the lower the distance, the greater the compactness. On the horizontal axis, the various decades are indicated.

What is immediately evident is that the *Cluster*2 is the most compact along the decades. This result is expected, as the cluster corresponds to summer configurations, which are typically well-defined and stable. Such conditions exhibit lower variability, ensuring greater consistency in the identification of these patterns. Consequently, the cluster appears more compact. Other clusters have larger and more variable values of mean distance, showing a higher internal dispersion

5.3 Temporal evolution of cluster frequencies (k= 7)

To interpret the results, we conducted a comprehensive temporal analysis to investigate how the frequency and characteristics of these regimes evolved across different decades. This analysis is important to understand long-term variability in the atmosphere and potential signals of climate change signals in European weather patterns.

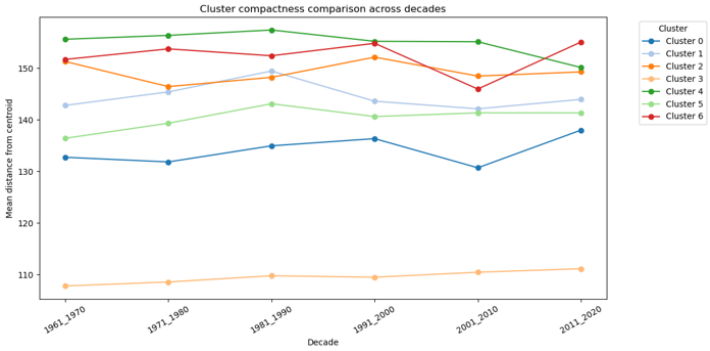

FIGURE 12: Comparison of cluster compactness during the six decades

To study and compare the way the number of days in each cluster vary over decades and from year to year, we need to map each day for each decade (1961-1970, 1971-1980, 1981-1990, 1991-2000, 2001-2010, 2011-2020) to the $k = 7$ centroids obtained with the use of the kmeans algorithm (Sec. 3.3).
For each day in every decade, we computed all the Euclidean distances to cluster centroids. We assign each observation to the nearest centroid and keep the regime assignment temporally consistent. Having saved those counts, they can be utilized as a basis for our work to result in the graph in Figure 13. The data indicate that some clusters have increased in frequency over the last decades and with increasing numbers in more recent decades; by comparison, some other clusters show a significant drop in frequency over the same decades. Particularly, *Cluster*3 is remarkable since it increased significantly, while *Cluster*4 decreased in the past 60 years.

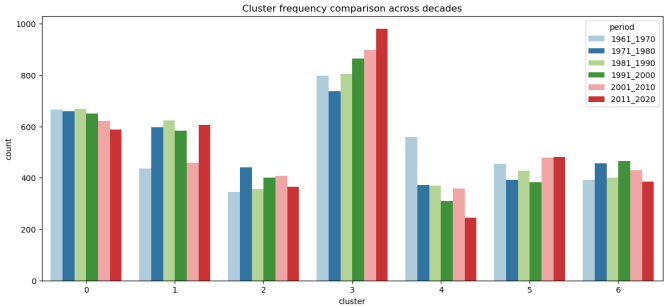

FIGURE 13: Bar charts showing the absolute occurrence of each cluster across decades.

To better identify these trends, we initially plotted the cluster counts per decade (Figure 14). However, collecting the data by decade tends to smooth out short-term variations, resulting in a too approximate analysis of the trend in the occurrence of the atmospheric configuration throughout the decades. Therefore, we also produced a graph with the annual counts for each cluster Figure 15. The temporal evolution of each atmospheric regime is quantified through linear regression analysis applied to the annual frequency counts over the 60-year period. The slope coefficient represents the linear rate of change in regime frequency per year, and it is computed using ordinary least squares regression. The overall change is calculated as the

relative difference between the predicted values at the end and beginning of the time series. This percentage represents the total relative change in regime frequency over the entire 60-year period, providing a normalized long-term climate trend that is independent of the absolute frequency values.
If the label present in the graph is green, it indicates increasing regime frequency, while if it is red, it represents decreasing trends.

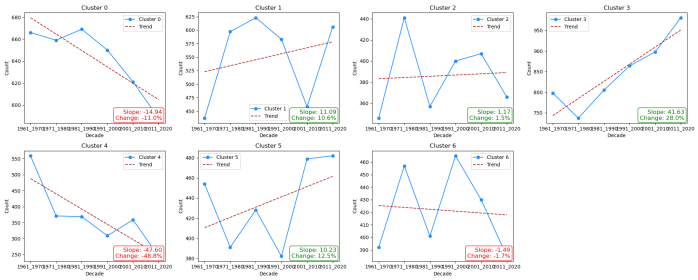

FIGURE 14: Cluster frequencies across the six decades. The linear regression analysis is represented by the red line.

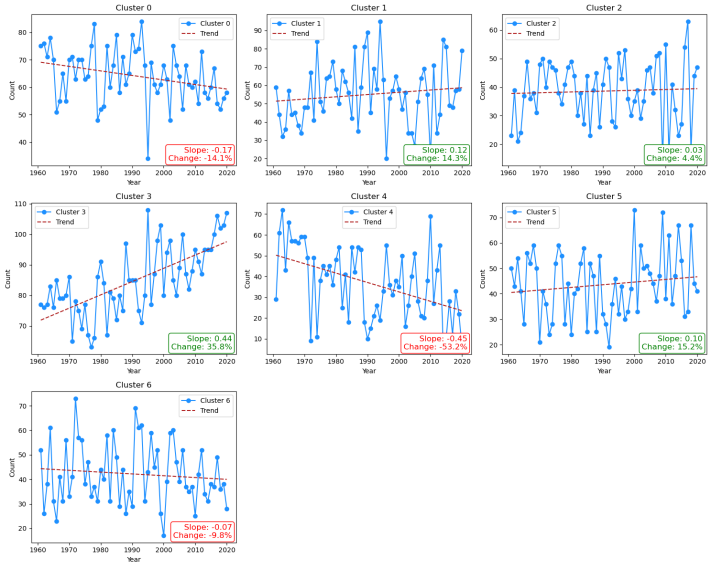

FIGURE 15: Cluster frequencies across the years. The linear regression analysis is represented by the red line.

5.3.1 Statistical Robustness: p-value

It can be observed from Figure 15 that the frequency of meteorological patterns varies over the 60 years, but it needs to be checked if the variations are statistically significant or only a result of interannual variability.

For testing the statistical robustness, we used the p-value test, which tests the null hypothesis that the regime frequency does not have a linear trend. The result is provided in Figure 16.

Values of $p < 0.05$ indicate statistically significant trends, and we can reject the null hypothesis of no change with 95% confidence. The analysis indicates that *Cluster*0, *Cluster*3, and *Cluster*4 contain significant trends ($p < 0.05$). These regimes exhibit high evidence of systematic temporal change, most probably caused by climate variability or long-term climate change signals. Other clusters possess stable frequencies ($p > 0.05$), signifying that their occurrence pattern

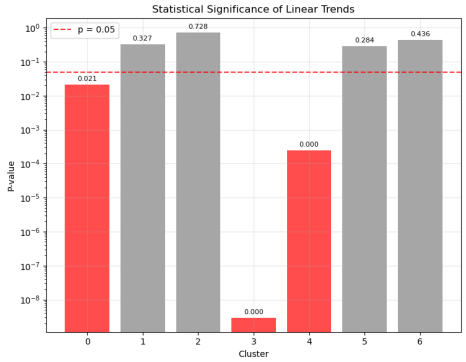

FIGURE 16: P-value test for annual frequency.

have remained unchanged with the natural variability of climate.
This statistical framework distinguishes natural climate trends from arbitrary fluctuations and offers a good foundation upon which to comprehend how atmospheric circulation patterns are evolving in reaction to varying climate conditions over the past six decades.

So, analyzing the three most robust variabilities over 60 years, it is evident that *Cluster*0, the warm phases at the edge of summer, has a moderate, but statistically relevant, negative trend. It is decreasing by a factor equal to -14.1%. *Cluster*3 has an excellent level of significance, and the African Anticyclone regime shows a $+35.8\%$ increase in frequency, underlying a marked rise in the occurrence of situations in which Europe is affected by African heat.
*Cluster*4, instead, presents a -53.2% decrease in frequency, a very marked reduction. This typical winter ATR regime is progressively dying out; it is becoming less and less frequent.
The other regimes, corresponding to *Cluster*1, *Cluster*2, *Cluster*5, and *Cluster*6, do not show a significant trend; their frequencies oscillate around constant mean values.

6 Conclusions

The classification of atmospheric circulation patterns is a specific research area within synoptic climatology [14]. The presented method provides a robust framework for identifying typical structures and climate change signals in European atmospheric circulation patterns, enabling the detection of statistically significant trends in the frequency of the weather regime that may reflect broader changes in the climate system.

In our approach, differential normalization of the data set (Sec.3.1), reduction of dimensionality through incremental PCA (Sec.3.2), and clustering with K mean have been proven to be a robust and effective method for identifying European atmospheric regimes. In particular, the normalization of the two separated halves of the dataset (temperature and geopotential) has guaranteed a balanced contribution of the two variables for the analysis, and has improved the overall quality of the physical reconstruction of the atmospheric structures.

The K-means algorithm (Sec. 3.3)has proven to be a valid and practical method for the clustering of multi-variable phase spaces: in addition to its simplicity and rapid conver-

gence, we have applied objective techniques for the choice of the final number of clusters k, like the elbow method, the silhouette and the Calinski-Harabasz index.
The clusters obtained correspond clearly and coherently to the well-known continental climatic configurations. The physical maps of the reconstructed centroids show characteristic structures like zonal regimes, blocks, cyclones, and anticyclones.

The decade-by-decade analysis highlights the desired signals: a strengthening of the frequency of very warm summer configurations (*Cluster*3), a marked decline in some winter configurations (*Cluster*4), and a seasonal shift of some intermediate classes toward averagely colder or warmer seasons (seasonal drift). Significance tests (p-values) show that the trends in some classes are statistically significant.

For future developments, we propose explicitly comparing two 30-year time series (or two selected 30-year time series) by performing separate clustering on each period and comparing the resulting centroids (rather than simply relying on frequency variation). This direct comparison of structures can reveal changes in the shape of the regimes as well as their occurrence. Furthermore, increasing the value of k, while requiring a more careful numerical justification, can increase the interpretability and allow the identification of less frequent or finer configurations, potentially useful for forecasting purposes or synoptic studies.

References

[1] Charu C. Aggarwal and Chandar K. Reddy. Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, 2014.

[2] European Centre for Medium-Range Weather Forecasts. Extended regime probabilities, 2025.

[3] Francesco Fantuzzo. Dalla brezza all'uragano: meteorologia moderna. ETS, Pisa, 1976.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2nd edition, 2009.

[5] James W. Hurrell and Clara Deser. North atlantic climate variability: The role of the north atlantic oscillation. Journal of Marine Systems, 78(1):28–41, 2009.

[6] Radan Huth, Christoph Beck, Andreas Philipp, Matthias Demuzere, Zbigniew Ustrnul, Dagmar Cahynová, Jan Kyselý, and Odd E. Tveito. Classifications of atmospheric circulation patterns: Recent advances and applications. Annals of the New York Academy of Sciences, 1146:105–152, 2008.

[7] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065):20150202, 2016.

[8] Johannes Max Müller, Oskar Andreas Landgren, and Dörthe Handorf. Summertime arctic and north atlantic-eurasian circulation regimes under climate change. Alfred-Wegener Institute for Polar and Marine Research, 2025.

[9] Gerald R. North, Thomas L. Bell, Robert F. Cahalan, and Fu-Jia Moeng. Sampling errors in the estimation of empirical orthogonal functions. Monthly Weather Review, 110(7):699–706, 1982.

[10] Andreas Philipp et al. Cost733cat – a database of weather and circulation type classifications. International Journal of Climatology, 30:176–183, 2010.

[11] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65, 1987.

[12] scikit-learn developers. Incremental pca example, 2025.

[13] scikit-learn developers. K-means clustering and evaluation resources, 2025. Includes K-Means API reference, evaluation metrics, and silhouette analysis example.

[14] Goran Stanojević. The classifications of atmospheric circulation. Journal of the Geographical Institute "Jovan Cvijić" SASA, 60(2):1–14, 2010.

[15] David M. Straus. Clustering techniques in climate analysis. In Oxford Research Encyclopedia of Climate Science. Oxford University Press, 2019.

[16] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering, 336:012017, 2018.

[17] Hongyue Wang and Mingzhou Song. Research on k-value selection method of k-means clustering algorithm. Journal of Information and Optimization Sciences, 40(1):1769–1786, 2019.

[18] Tim Woollings et al. Blocking and its response to climate change. Current Climate Change Reports, 4(3):287–300, 2018.