Revisiting the Identification of Wintertime Atmospheric Circulation Regimes in the Euro-Atlantic Sector

Swinda Falkena
Jana de Wiljes, Antje Weisheimer, Ted Shepherd

Contact: s.k.j.falkena@pgr.reading.ac.uk

May 5, 2020
Atmospheric Circulation Regimes

Atmospheric circulation regimes = Recurrent and persistent patterns

- Concept: Weather is a stochastic process with statistics conditioned on the circulation regime
- Many regions (NH, SH, Pacific Sector, ...) have been studied for the identification of circulation regimes
- Focus on the Euro-Atlantic sector in winter
  - Most studies identify four regimes

Figure: The four regimes based on the 500 hPa geopotential height (Cassou, 2008).
Contents

1 Methods
   • Standard Approach
   • k-means Clustering

2 Optimal Number of Regimes
   • Consistency of Clustering Method
   • Information Criteria
   • Six Regimes

3 Persistent Regimes
   • Including a Constraint in k-means

4 Conclusion and Discussion
Methods: Standard Approach
Standard Approach

Starting from the 1990s many approaches have been taken to identify circulation regimes. The most-common approaches are:

- **Data**
  - 500 hPa geopotential height
  - Project onto Empirical Orthogonal Functions (EOFs)
  - Remove the seasonal cycle
  - Apply a (10-day) low-pass filter to focus on persistent, low-frequency behaviour

- **Methods**
  - Analysis of the probability density function
  - Hierarchical clustering
  - \textit{k-means clustering}
  \[ \Rightarrow \text{This has become the most-used approach} \]

**Question:**
- Does filtering data (EOFs or time-filtering) before applying a clustering method yield different results than clustering raw data?
Our Approach

Question:
- Does filtering data (EOFs or time-filtering) before applying a clustering method yield different results than clustering raw data?

To answer this question we use:

- **Data**
  - ERA-Interim 500 hPa geopotential height
  - Euro-Atlantic sector (20-80°N, 90°W-30°E)
  - Daily data for December till March, 1979 - 2018
  - Deviations with respect to a fixed background state

- **Method**
  - *k*-means clustering

We compare regimes for
- Full field (raw) data
- EOF data (for 5 till 20 EOFs)
and enforce persistence of the regimes by using either
- Low-pass filtered data
- A constraint in the clustering algorithm
Methods: $k$-means Clustering
**K-means Clustering**

First, fix the number of clusters $k$, here $k = 3$.

Then, follow this procedure:

1. Pick the initial $k$ clusters, coloured dots
2. Assign each data point (black) to the closest cluster, within coloured lines
3. Compute the average over the data points assigned to each cluster, coloured stars
4. Repeat until the clusters converge
Formulated mathematically this means that, given a number of clusters $k$, with

- Dataset $\{x_t\}_{t \leq T}$
- Cluster parameters $\Theta = (\theta_1, ..., \theta_k)$
- Model distance functional $g(x_t, \theta_i)$, giving the distance between a data point $x_t$ and a cluster $\theta_i$
- Weights $\Gamma = (\gamma_1(t), ..., \gamma_k(t))$, indicating to which cluster a data point belongs (since in practice $\gamma_i(t)$ is either zero or one)

$k$-means clustering minimizes the **averaged clustering functional**:

$$L(\Theta, \Gamma) = \int_0^T \sum_{i=1}^k \gamma_i(t)g(x_t, \theta_i)dt$$

Note: Inclusion of $\Gamma$ is not required for $k$-means as described here, but is needed when a persistence constraint is included in the algorithm later on.
Optimal Number of Regimes: Consistency of the Clustering Method
An important question when using $k$-means clustering is:  
**What is the optimal number of clusters $k$?**

Mainly, studies look at how **consistent** the outcome of the clustering algorithm is when run for different initial conditions:

- Often-used is the classifiability index (Michelangeli et al., 1995)
- Significance of the clusters is verified against synthetic datasets

We run the $k$-means clustering algorithm 500 times for different initial conditions and look at:

- The clustering functional $L$, the lower its value, the better the result (the lowest value $L_{min}$ is selected as the ‘true’ clusters)
- The **data similarity** with the ‘true’ cluster = the number of data points assigned to the same clusters
Consistency of the Clustering Method: EOFs

Below are examples of the distributions of $L$ and the data similarity for 500 tests for EOF data (20 EOFs)

$\Rightarrow$ We want a measure indicating how consistent, or similar, a result is

Figure: The difference of the clustering functional $L$ with the lowest value $L_{\text{min}}$.

Figure: The data similarity.
Consistency of the Clustering Method: EOFs vs Full Field

We look at the distribution of the data similarity for $L$ close to the optimal result ($L_{i+1} - L_i < \epsilon$, for $L_i$ ordered), i.e. if all results with small $L$ also have large data similarity the result is consistent:

- A high mean indicates a good match
- The variance (corrected for the number of clusters $k$) gives a measure of how consistent the results are

Note that the values are expected to be smaller for higher $k$, as more clusters allow for more variability.

Note: For definite conclusions statistical significance tests on these measures are needed. When full field data is used these are difficult to establish.
We compare the results for EOF data with those for full field data to look into the optimal number of regimes.

- For EOF data $k = 4$ is found most consistent, which corresponds with results from literature.
- For full field data $k = 5$ and $k = 6$ are found to be more consistent than $k = 4$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\mu$</th>
<th>$\sigma^2/k$</th>
<th>#data</th>
<th>$\mu$</th>
<th>$\sigma^2/k$</th>
<th>#data</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4643</td>
<td>3649</td>
<td>254</td>
<td>4552</td>
<td>2109</td>
<td>485</td>
</tr>
<tr>
<td>4</td>
<td>4658</td>
<td>330</td>
<td>197</td>
<td>4607</td>
<td>1440</td>
<td>201</td>
</tr>
<tr>
<td>5</td>
<td>4509</td>
<td>978</td>
<td>265</td>
<td>4660</td>
<td>149</td>
<td>204</td>
</tr>
<tr>
<td>6</td>
<td>4571</td>
<td>1103</td>
<td>315</td>
<td>4581</td>
<td>790</td>
<td>316</td>
</tr>
</tbody>
</table>
Consistency of the Clustering Method: Odd and Even Years

A method often-used to check the performance of the clustering algorithm is to split the dataset in half. Therefore we look at the consistency of the clusters when clustering the odd and even years separately.

- Large differences between the odd and even years
- Both odd and even years are quite consistent for $k = 4$
- Is half the dataset of sufficient length to draw conclusions?

<table>
<thead>
<tr>
<th>$k$</th>
<th>Odd years</th>
<th>Even years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
<td>$\sigma^2/k$</td>
</tr>
<tr>
<td>3</td>
<td>2342</td>
<td>875</td>
</tr>
<tr>
<td>4</td>
<td>2359</td>
<td>452</td>
</tr>
<tr>
<td>5</td>
<td>2187</td>
<td>3562</td>
</tr>
<tr>
<td>6</td>
<td>2296</td>
<td>322</td>
</tr>
</tbody>
</table>

⇒ Can such a consistency argument be used to draw conclusions about the optimal number of regimes if the results are not coherent?
Optimal Number of Regimes: Information Criteria
Information Criteria

The consistency discussion did not yield coherent, conclusive, results. More general, one can ask the question whether a large spread in the clustering result disqualifies the suitability of the ‘true’ regimes? Therefore we turn to a different method for identifying the optimal number of regimes: Information Criteria.

Information criteria strike a balance between how well the clusters represent the data and the number of clusters used. Here we discuss the two most used criteria:

1. Akaike Information Criterion: \[ \text{AIC} = -2 \log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K \]

2. Bayesian Information Criterion: \[ \text{BIC} = -2 \log(\mathcal{L}(\hat{\theta}|\text{data})) + K \log(n) \]

where \( \mathcal{L}(\hat{\theta}|\text{data}) \) is the likelihood of the optimal clusters \( \hat{\theta} \) given the data, \( K \) the number of parameters needed to describe all clusters and \( n \) the sample size.
Information Criteria

The optimal number of clusters is found where the information criteria has its minimum.

\[
\text{AIC} = -2 \log(L(\hat{\theta}|\text{data})) + 2K \\
\text{BIC} = -2 \log(L(\hat{\theta}|\text{data})) + K \log(n)
\]

The first term in both criteria is the same and gives how well the clusters represent the data. The difference arises in the second term, often called the penalty term, which penalizes the use of many parameters to prevent over-fitting:

- The **BIC** is better suited for the **full field data** since the penalty term takes into account the sample size and therefore is stronger with respect to the number of parameters.
- The **AIC** is better suited for the **EOF data** since the penalty term of the BIC likely is too strong.
Information Criteria for Regimes

**EOF Data**

![Graph showing the AIC for different numbers of EOFs.](image)

**Figure:** The AIC for different numbers of EOFs.

- For 20 EOFs $k = 4$ is found to be optimal using the AIC.

**Full Field Data**

![Graph showing the AIC and BIC for the full field data.](image)

**Figure:** The AIC and BIC for the full field data.

- $k = 6$ is found to be optimal using the BIC.

**Note:** Changes in the AIC with the number of EOFs are due to more variability being neglected when less EOFs are used.
Optimal Number of Regimes: Six Regimes
What are the regimes for $k = 6$?

- The same as for $k = 4$ (NAO+, NAO-, Atlantic Ridge (AR), Scandinavian Blocking (SB))
- A low pressure area over the Atlantic (AR-)
- A low pressure area over Scandinavia (SB-)
Occurrence and Persistence

What are the occurrence rates and transition probabilities of a regime to itself (indicating its persistence)?

<table>
<thead>
<tr>
<th></th>
<th>AR+</th>
<th>SB+</th>
<th>NAO+</th>
<th>NAO-</th>
<th>AR-</th>
<th>SB-</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occurrence</td>
<td>21.3</td>
<td>26.8</td>
<td>31.5</td>
<td>20.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Trans. P.</td>
<td>0.756</td>
<td>0.792</td>
<td>0.850</td>
<td>0.849</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occurrence</td>
<td>15.6</td>
<td>19.6</td>
<td>16.9</td>
<td>15.5</td>
<td>16.3</td>
<td>16.1</td>
</tr>
<tr>
<td>Self-Trans. P.</td>
<td>0.712</td>
<td>0.748</td>
<td>0.751</td>
<td>0.847</td>
<td>0.787</td>
<td>0.730</td>
</tr>
</tbody>
</table>

Differences of the $k = 6$ regimes with $k = 4$:

- Two additional regimes identified by Northern low pressure
  - This explains the strong drop in occurrence of the NAO+ with respect to $k = 4$
- The NAO- remains as persistent as for $k = 4$, despite its drop in occurrence
Persistent Regimes
Drawbacks of $k$-means Clustering

$k$-means clustering has drawbacks that affect the occurrence rates and transition probabilities of the regimes:

- Every data point is assigned to a cluster, even if its distance to different clusters is comparable.
- Time is not taken into account, the data can be reshuffled randomly and the same clusters are found.

If a data point lies in between two clusters, to which one do you assign it?
- To the cluster it is (just) closest to?
- To the cluster of its neighbours even though it is (slightly) further away?

The standard approach to focus on persistent behaviour of the circulation is to apply a low-pass filter, here we use a new approach which includes a persistence constraint in the clustering algorithm.
Persistent Regimes: Including a Constraint in $k$-means
Including a Persistence Constraint in \( k \)-means

Recall the clustering functional

\[
L(\Theta, \Gamma) = \int_0^T \sum_{i=1}^k \gamma_i(t) g(x_t, \theta_i) \, dt.
\]

To enforce persistence of the clusters we put a constraint on the weights \( \Gamma \)

\[
\sum_{i=1}^k \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq C.
\]

This restricts the number of transitions between clusters that are allowed.

Minimization of \( L(\Theta, \Gamma) \) is done in two (iterated) steps:

1. Given \( \Theta \), minimize \( L \) for \( \Gamma \) \( \Rightarrow \) Linear programming
2. Given \( \Gamma \), minimize \( L \) for \( \Theta \) \( \Rightarrow \) \( k \)-means clustering

Each \( C \) has a corresponding average regime duration:

<table>
<thead>
<tr>
<th>( C )</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>600</td>
<td>15.8</td>
</tr>
<tr>
<td>800</td>
<td>11.8</td>
</tr>
<tr>
<td>1000</td>
<td>9.5</td>
</tr>
<tr>
<td>1200</td>
<td>7.9</td>
</tr>
<tr>
<td>1400</td>
<td>6.8</td>
</tr>
<tr>
<td>1600</td>
<td>5.9</td>
</tr>
<tr>
<td>1800</td>
<td>5.3</td>
</tr>
<tr>
<td>2000</td>
<td>4.7</td>
</tr>
<tr>
<td>2200</td>
<td>4.3</td>
</tr>
</tbody>
</table>
Example of the Effect of a Persistence Constraint

Consider a simple 2D example

3 clusters (blue, green, purple) with transitions determined by a transition matrix.

- Orange: wrongly assigned by standard $k$-means
- Red: still wrongly assigned with a constraint

Points on the boundary between clusters switch with the incorporation of the constraint

Short back-and-forth transitions in the clustering result are reduced

Closer to the ‘true’ persistence

Figure: The transition sequence between clusters.
Effect on Persistence and Occurrence

To see what the effect is of the persistence constraint on the occurrence and self-transition probabilities we run the algorithm for different values of $C$. The results are shown on the next slide, together with the results for applying a 5- or 10-day low-pass filter to the data (LP5, LP10) and the unconstrained algorithm (Field).

- The persistence constraint starts to affect the persistence for $C$ below either 1800 ($k = 4$) or 2200 ($k = 6$), from then the increase in self-transition probability is approximately linear with $C$ for all regimes.
- The occurrence rate does not change until $C$ becomes very small, corresponding to unrealistically large average regime durations (over 9 days for $k = 4$ and over 8 days for $k = 6$).
- Applying a low-pass filter does affect the occurrence rates of the regimes and thus introduces a possible bias in the found regimes.

A realistic range of $C$ is indicated by the gray bands in the figures on the next slide.
Persistent Regimes

Including a Constraint in $k$-means

Effect on Persistence and Occurrence

Swinda Falkena
Circulation Regimes Revisited
May 7, 2020
**Optimal Constraint Value**

The final question to discuss is: What is the optimal constraint value $C$? There are two possible ways to determine this:

- Information criterion (BIC)
- Look for which $C$ the occurrence rates start to be affected

Both point to an optimal $C$ of approximately 1400 – 1500, corresponding to an average regime duration of 6-7 days.

We find

- Persistence beyond the synoptic timescale
- The optimal regime duration differs less for different $k$, than for the unconstrained results
- The constraint helps to identify the physical signal
Conclusion and Discussion
Conclusion and Discussion

- Using full field data, six circulation regimes is found to be optimal
  - This introduces a symmetry in the clusters
  - Four regimes is the standard in literature (using EOFs)
- Including a persistence constraint in the clustering method increases the persistence without changing the occurrence of the regimes
  - “In between” data points are forced towards the cluster of their neighbours
  - Time-filtering does affect the occurrence

The persistence constraint helps to identify the physical signal of the persistent regimes. More generally, care needs to be taken with filtering data before applying a clustering method.


