Cluster Analysis of Northern Hemisphere Wintertime 500-hPa Flow Regimes during 1920-2014

Bao and Wallace, 2015

Cheng and Wallace, 1993

Machine Learning Journal Club – March 13, 2026

Introduction

- Studies on structure and dynamics of atmospheric variability takes 1 of 2 approaches:
  - Linear
    - Historically used regression analysis, empirical orthogonal functions (EOFs), other linear algebra based methods
  - Non-linear
    - Historically used cluster analysis (Ward's hierarchical clustering), some studies also used probability density function (PDF) in reduced phase space and focusing on leading EOFs

Methods compared in this paper for characterizing hemispheric regimes:

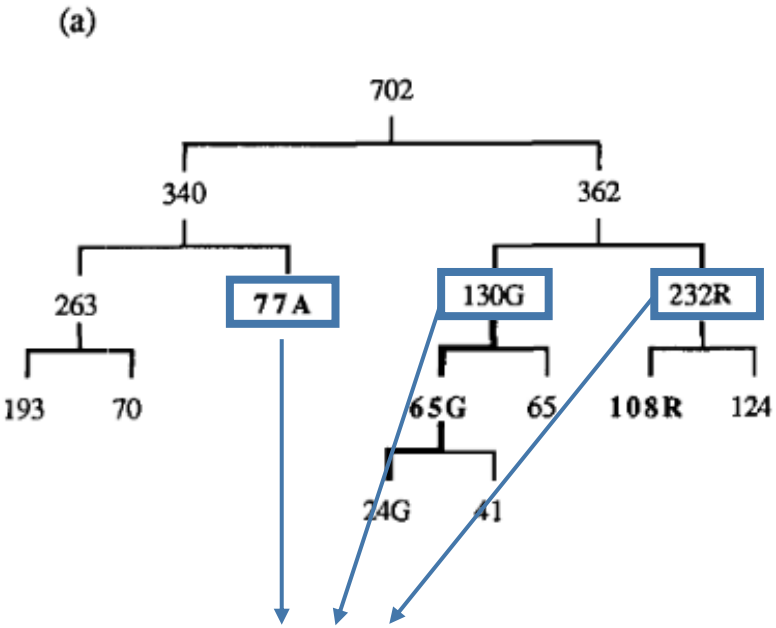
- Self-Organizing Maps (SOM) - this study
  - Clusters are "more distinctive and more robust"
- Ward's Hierarchical clustering – (Cheng and Wallace, 1993) referred to as CW
  - Repeated the methods used in this paper

Data

- Combined 3 different reanalysis datasets with different time spans:
    1) 4-times-daily Z500 geopotential height from ECMWF (ERA-40)
    2) 4-times-daily Z500 geopotential height from ECMWF interim (ERA-Interim)
    3) daily Z500 from NOAA 20th-century reanalysis (20CR)
- 2.5° lon x 2.5 ° lat horizontal resolution (or interpolated to this)

Cluster Analysis Methods Review

- Both Ward's and SOM use minimum Euclidean distance between maps
- Ward's Hierarchical clustering
  - Creates a cluster tree where each tree is a multilevel hierarchy in which the clusters at one level are merged and form the clusters at the next level based on Euclidean distance
  - User specifies the scale/level of clustering (i.e. how many clusters to retain)
  - Fig. 3a from CW showing number of member maps in the cluster. The 702 is the climatological mean map



- SOM
  - Mean squared distance between maps in clusters and original dataset is minimized
  - User specifies dimensions of the array of clusters in low dimensional array (usually 2 dimensions)
  - User prescribes the # of clusters

The "hemispheric regimes"
A – Alaska ridge
G – Greenland blocking
R – Rockies ridge

Cluster Analysis Assessment Methods

- Distinctiveness & Robustness vs Number of Clusters
  - Smaller more numerous clusters are more distinctive but more susceptible to unpredictability of sampling variability
- Variance Ratio (VR)
  - External Variance – the squared distance between cluster centroid and centroid of dataset
  - Total Variance – the mean squared distance between individual maps in that cluster and centroid of dataset
  - VR = External/Total
  - VR increases as large clusters are divided into smaller tighter clusters and it reaches its limiting value of unity when each cluster only has 1 member
- Countervailing Robustness Metric aka Reproducibility Parameter (RP)
  - Measures how much the clusters change in response to small perturbations in the input data
  - Perturbations in input data are large enough to tell the dependence of RP on clustering protocol and observe RP decline as # of SOM clusters increases

Results

The 1st three patterns correspond with G', A', R' regimes, respectively

→ Primed regimes have seasonally varying climatological mean removed (from CW)

"SOM patterns are reminiscent of the most reproducible clusters obtained using Ward's method, shown in Fig. 1, but they exhibit much higher values of RP"

Note:

Ward's clusters – contain ½ the days

SOM cluster – contain all the days

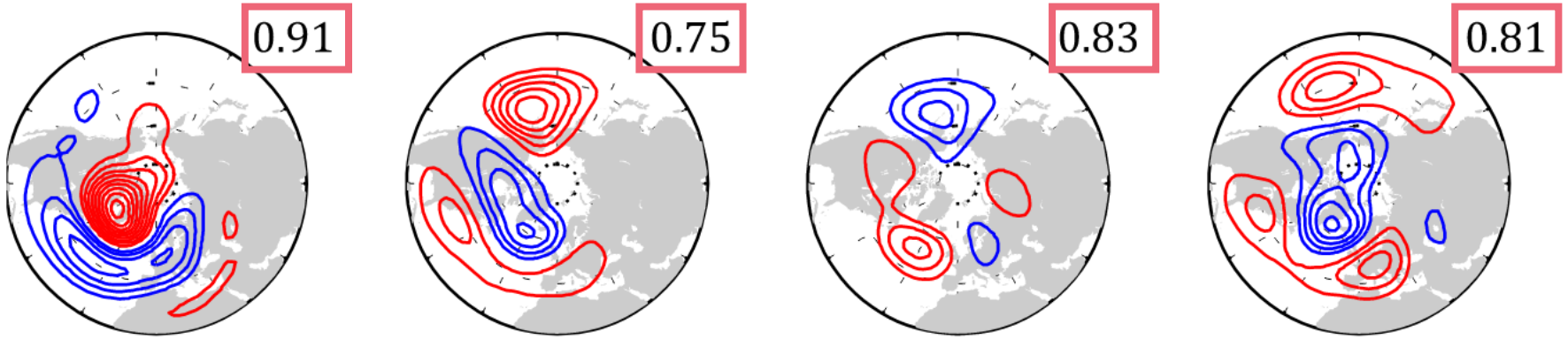


FIG. 1. Composite 500-hPa height anomaly maps of the cluster derived from Ward's method. Contour interval: 25 m; red (blue) contours denote positive (negative) values; the zero contours are omitted. The value of RP is shown at the top right of each panel.

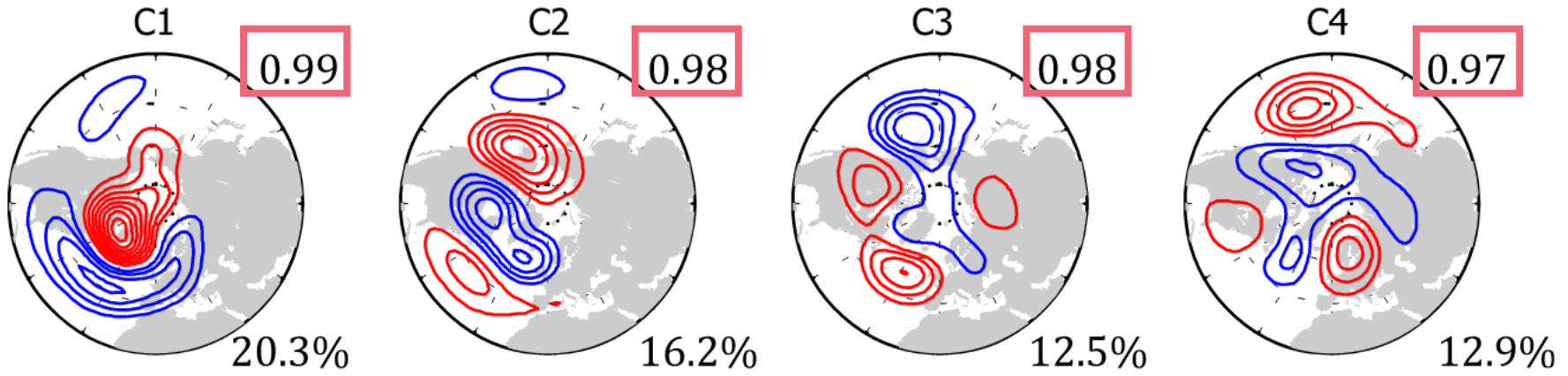


FIG. 2. Composite 500-hPa height anomaly maps of the cluster derived from SOM. Contour interval: 25 m; red (blue) contours denote positive (negative) values; the zero contours are omitted. The value of the variance ratio is shown at the bottom right of each panel, and the value of RP is shown at the top right.

Intepreting Patterns

- 4 clusters (2x2 and 1x4 arrays) optimizes reproducibility (RP)
- Collinearity shows up as symmetries:
  - Leading EOF – North Atlantic Oscillation (NAO)
    - C1 and C4 in Ward's present with opposing polarity
  - 2nd EOF – Pacific-North America (PNA) pattern
    - C3 & C4 in Ward's present with opposing polarity

TABLE 2. RP for the clusters derived from each of the array configurations arranged in descending order. Boldface denotes a value higher than 0.80 and an asterisk represents a value less than 0.70.

$m \times n$	Average of 10 correlation coefficients in descending order								
1×2	**0.88**	**0.88**	—	—	—	—	—	—	—
1×3	**0.99**	**0.98**	**0.97**	—	—	—	—	—	—
1×4	**0.99**	**0.98**	**0.98**	**0.97**	—	—	—	—	—
2×2	**0.99**	**0.98**	**0.98**	**0.97**	—	—	—	—	—
1×5	**0.98**	**0.90**	**0.86**	0.78	*	—	—	—	—
1×6	**0.93**	**0.92**	**0.88**	**0.85**	0.75	*	—	—	—
2×3	**0.95**	**0.93**	**0.83**	0.78	0.76	*	—	—	—
1×7	**0.92**	**0.89**	**0.86**	**0.84**	0.77	0.71	*	—	—
1×8	**0.92**	**0.91**	**0.84**	**0.82**	0.78	0.78	*	*	—
2×4	**0.94**	**0.88**	**0.87**	**0.83**	0.74	0.73	*	*	—
1×9	**0.94**	**0.85**	**0.85**	**0.83**	**0.82**	**0.82**	0.79	*	*
3×3	**0.94**	**0.91**	**0.91**	**0.90**	**0.88**	**0.87**	**0.86**	**0.82**	*

Null or Reserve Cluster

Z500 maps for days closest to climatology lead to almost equidistant maps from centroid of 2 different clusters (i.e. cluster membership ambiguity)

- Null cluster with Ward's (from CW paper)
  - Exclude from cluster membership and put in null cluster days where:

    $[d_{\text{Z500 map from centroids of nearest cluster}} - d_{\text{Z500 map from centroid of next nearest cluster}}] < d_0$

    d = distance and d_0 = 14m
  - Lead to 2/3 of days in dataset went to null cluster
- Null & reserve clusters with SOM
  - Composite Z500 maps for "reserve cluster" and for days that formerly were in that cluster but were put into null cluster

Null or Reserve Cluster

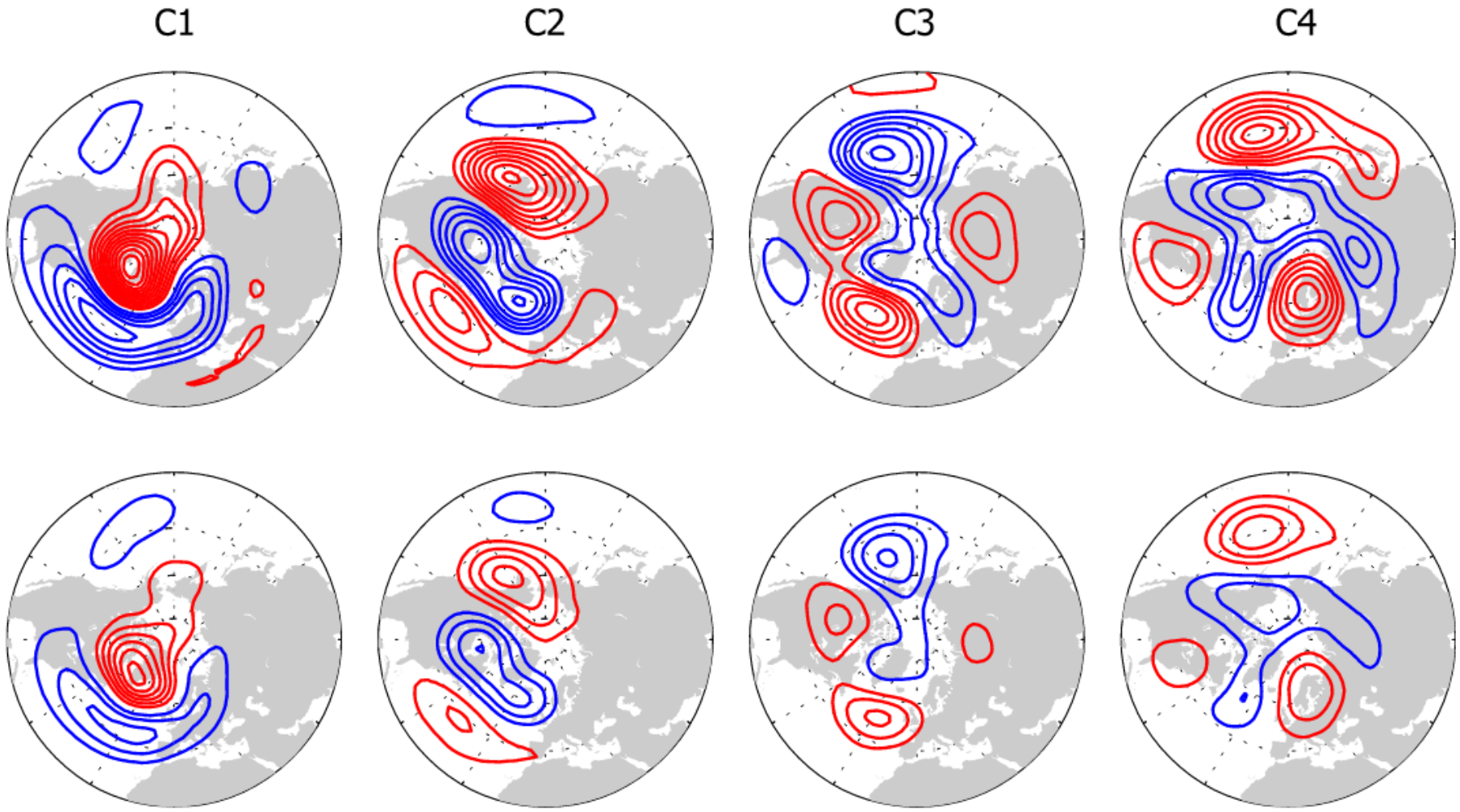


FIG. 3. As in Fig. 2, but composite of the (top) reserve members and (bottom) members relegated to the null cluster. Contour interval: 25 m.

Sectoral Clusters

- Clusters derived from analysis of full **N Hemisphere domain** tend to be centered over/close to **N. America**
- 180° sectors centered at various longitudes used to investigate potential other flow regimes over **Eurasia**
  - "admittedly cursory survey"

Sectors?? Happy almost pi day

Or cursed pies since it's Friday the 13th

Sectoral Clusters

- W hemisphere clusters are similar to those in the null clusters
  - WC1 – weakening of climatological-mean trough
  - WC2 & 4 – Eurasian wave trains
- RP not calculated for Eastern clusters, “but there are indications that they may be quite fragile”

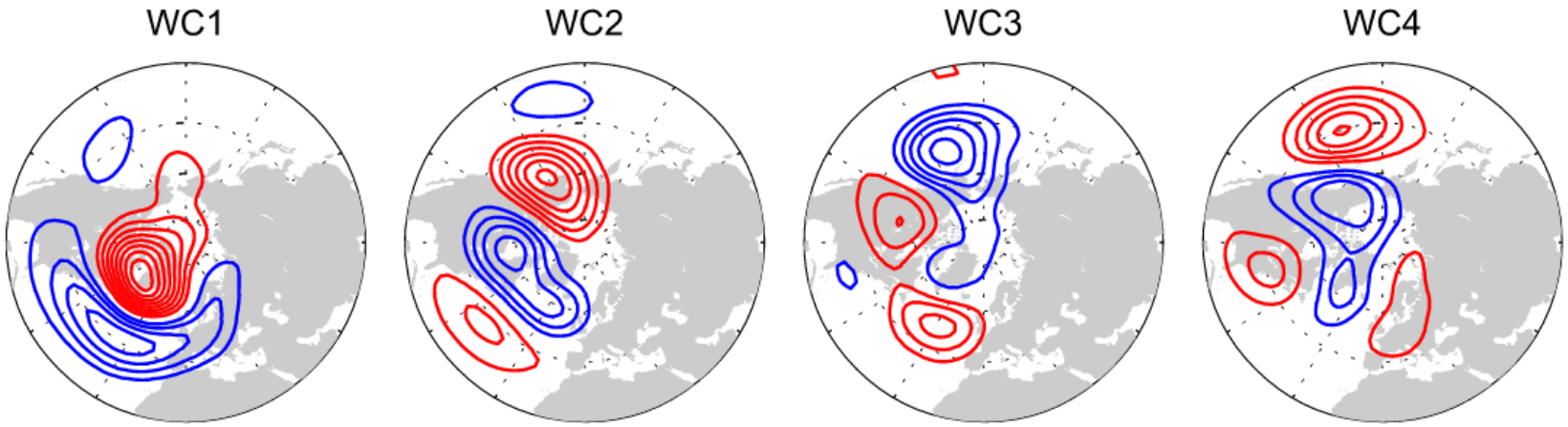


FIG. 4. Composite maps for the sectoral clusters of (top) the Western Hemisphere, extending from the date line, across North America, to the Greenwich meridian and (bottom) the Eastern Hemisphere from the Greenwich meridian, across Eurasia, to the date line. Contour interval: 25 m. The sectoral clustering is used to determine which dates belong to which clusters, but the composite maps are plotted for the full Northern Hemisphere domain.

Synoptic Interpretations

- SOM C1 – negative polarity of NAO, often observed in association with blocking over Greenland
- SOM C2 – observed in association with blocking over Gulf of Alaska (GOA) w/cold downstream trough east of Rockies and strong westerly jet across N. Atlantic
- SOM C3 – enhancement of climatological-mean stationary waves particularly in trough over GOA and downstream ridge over Rockies
- SOM C4 – not interesting for synoptic climatology

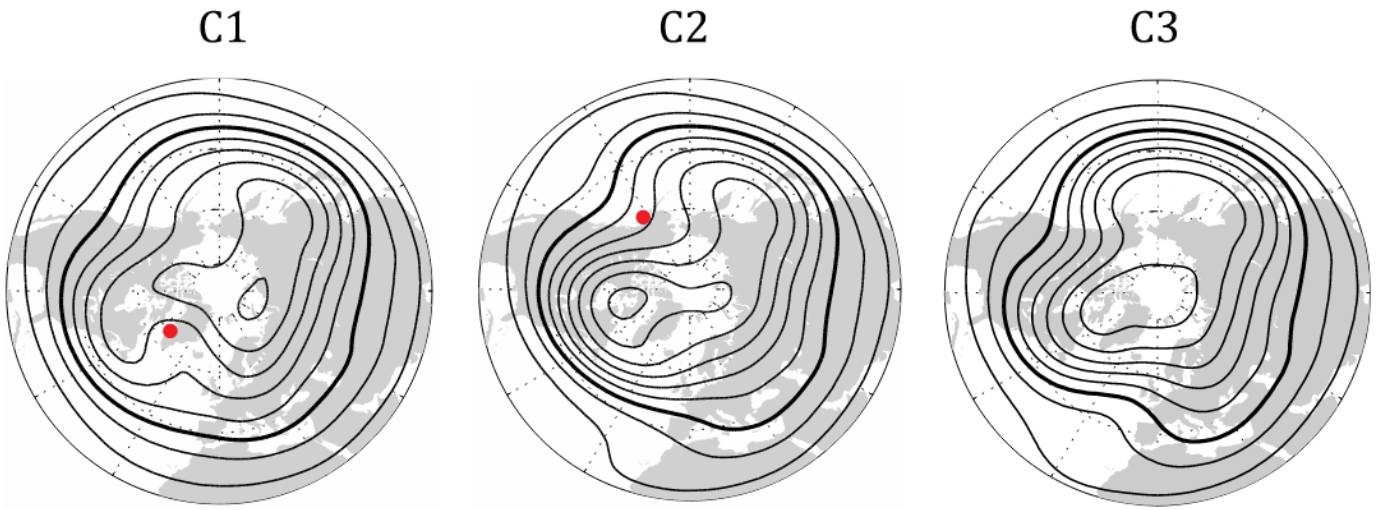


FIG. 5. Composite total Z500 fields for the first three SOM clusters. Contour interval: 100 m. The 5600-m contour is bold. The red dots mark the primary positive centers in the Z500 anomalies for (left) C1 and (middle) C2, transcribed from Fig. 2.

Summary & Discussions

- 4 reproducible clusters
- SOM is better than Ward's because:
  - Able to cluster with entire dataset vs half
  - More robust w.r.t. small perturbations to input data
  - Spatial patterns less collinear and less dominated by 2 leading EOFs
  - External-to-Total Variance ratios are as large as Ward's clusters even though more data is used in SOM
- "The reason why the SOM clusters are more distinctive and more robust than the clusters derived from Ward's method remains to be resolved"
  - Speculation: related to use of all days vs half of days

Summary & Discussions

- Need to analyze E. and W. of N. Hemisphere separately → what could be causing this?
- Basically, SOM is attractive for organizing maps in 2D arrays but "we **have not found it useful for this application**, because the number of reproducible clusters is not large enough to take advantage of it"
- Studies have tried larger matrices for SOM
  → unsure if it helps
- Cross validation seems necessary to ensure array size ~ # of statistical degrees of freedom