Self-organizing maps (SOMs) and k-means clustering: Part 2

Steven Feldstein

The Pennsylvania State University

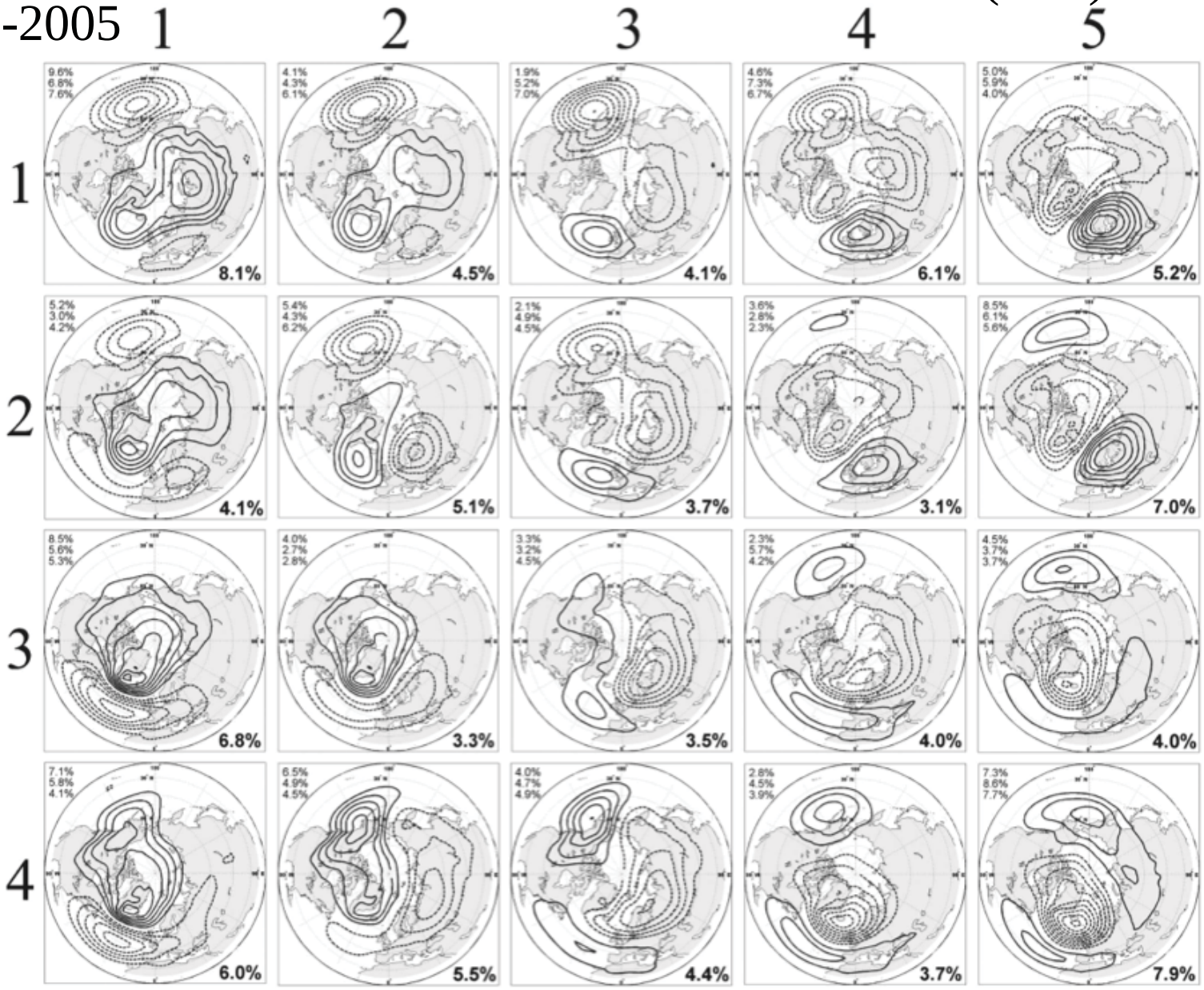
Collaborators: Sukyoung Lee, Nat Johnson

Trieste, Italy, October 21, 2013

A SOM Example

P1=1958-1977
P2= 1978-1997
P3=1998-2005

North Pacific Sea Level Pressure (SLP)



Number of SOM patterns

- The number of SOM patterns that we choose is a balance between:

  resolution (capturing essential details) & convenience (not too many panels for interpretation)

- Choice on number of SOM also depends upon particular application

Testing of SOM Results

- Check for reasonably large mean daily pattern correlations for each SOM

TABLE 1. Pattern correlations and Euclidean distances (hPa) between the SOM-derived hemispheric SLP anomaly field and the actual hemispheric SLP anomaly fields for each of the three periods considered: in each cell the pattern correlation lies above the Euclidean distance.

Number of SOM patterns	P1	P2	P3
20 (4 × 5)	0.42	0.83	0.76
	35.1	27.4	35.5
35 (5 × 7)	0.64	0.88	0.74
	30.3	25.9	36.3
96 (8 × 12)	0.79	0.92	0.78
	25.4	21.2	33.2
300 (15 × 20)	0.84	0.92	0.85
	24.1	19.8	29.2

Testing of SOM Results

- Compare composites of daily data with the SOM pattern itself

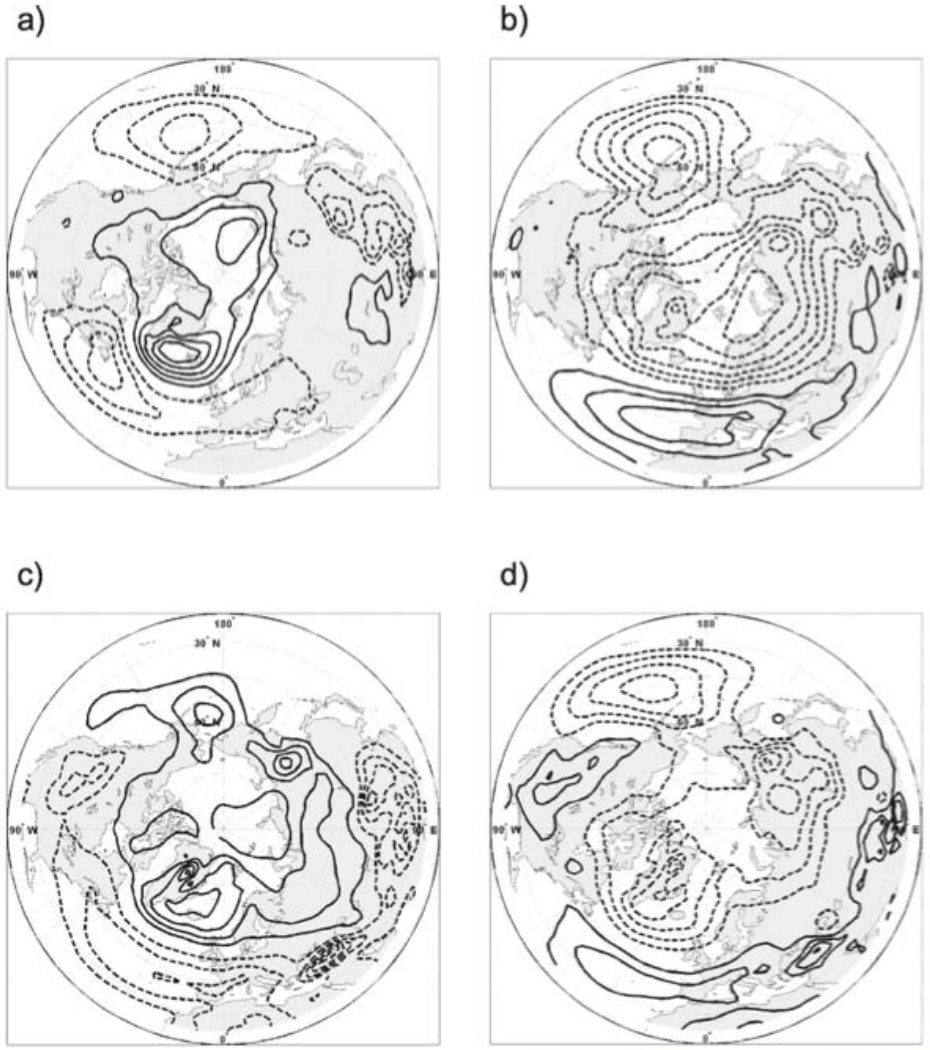


FIG. 3. SOM-derived composite SLP anomalies for (a) P1 and (b) P2 and the actual SLP anomaly fields for (c) P1 and (d) P2: contour interval is 0.2 hPa for (a) and (b) and 0.5 hPa for (c) and (d) with the same contouring conventions as in Fig. 1.

Impact of number of SOMs

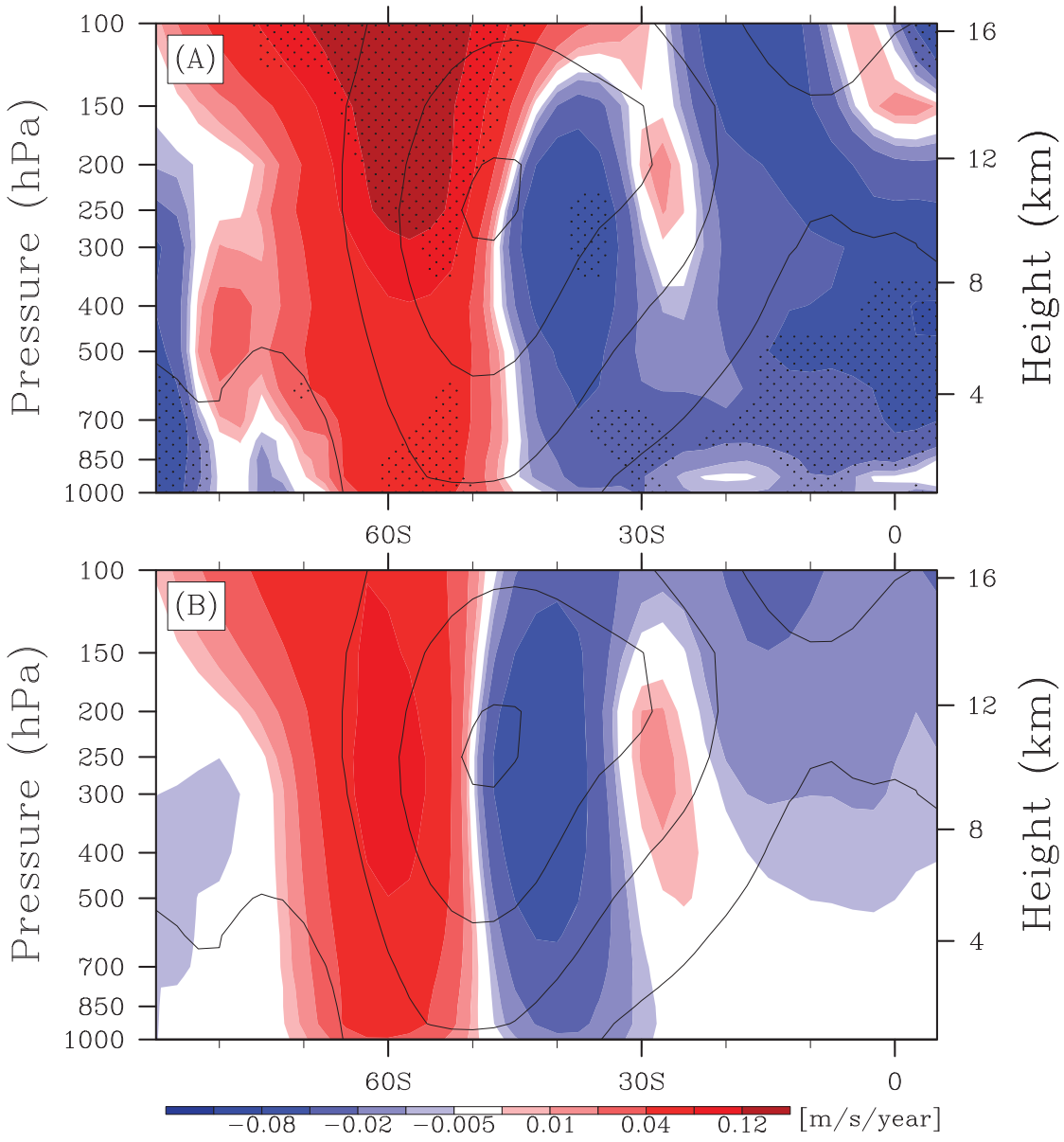
- A relatively small number of SOM patterns favours low-frequency, slowly-propagating, large-scale patterns.

- To resolve rapidly propagating synoptic-scale patterns, i.e., weather, require more SOM patterns

Poleward Jet Shift in the Southern Hemisphere

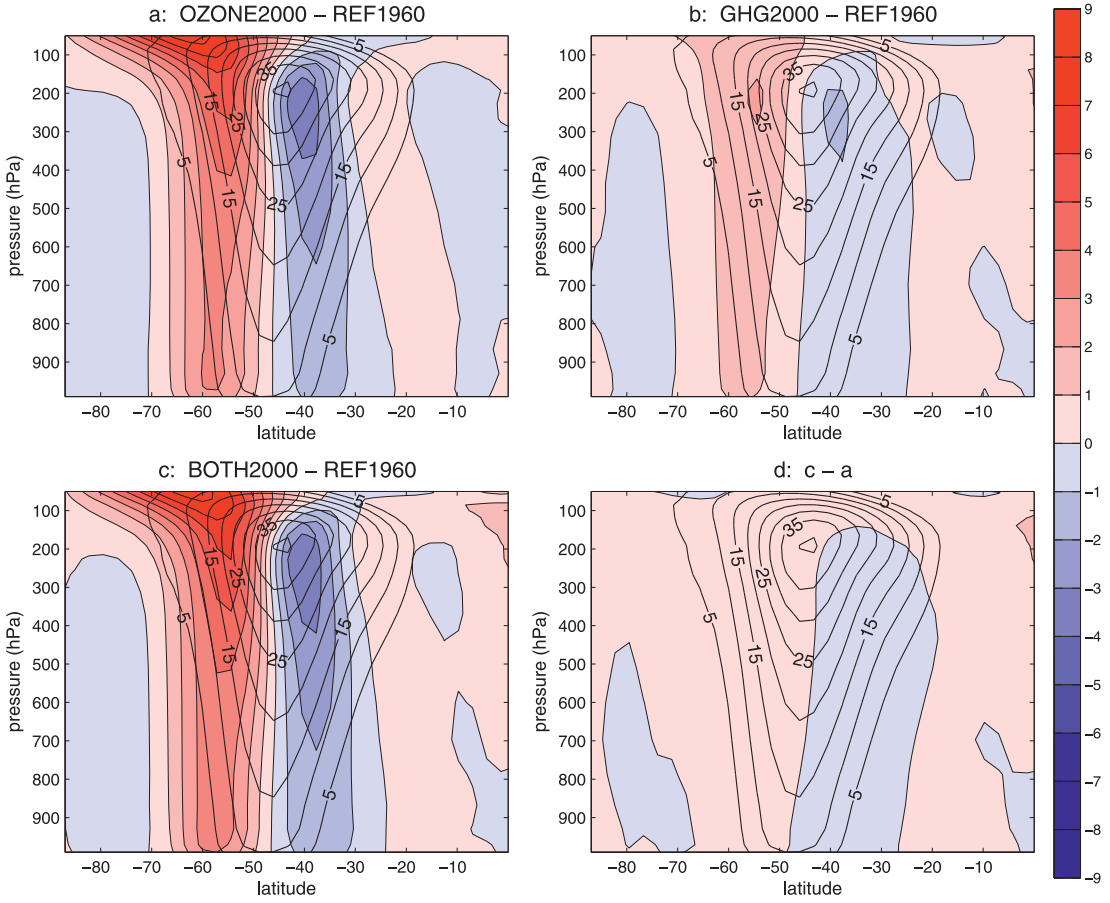
Observed trend

SOM trend



ERA-Interim Data (1979-2008)

Zonal wind trend associated with GHG driving and ozone

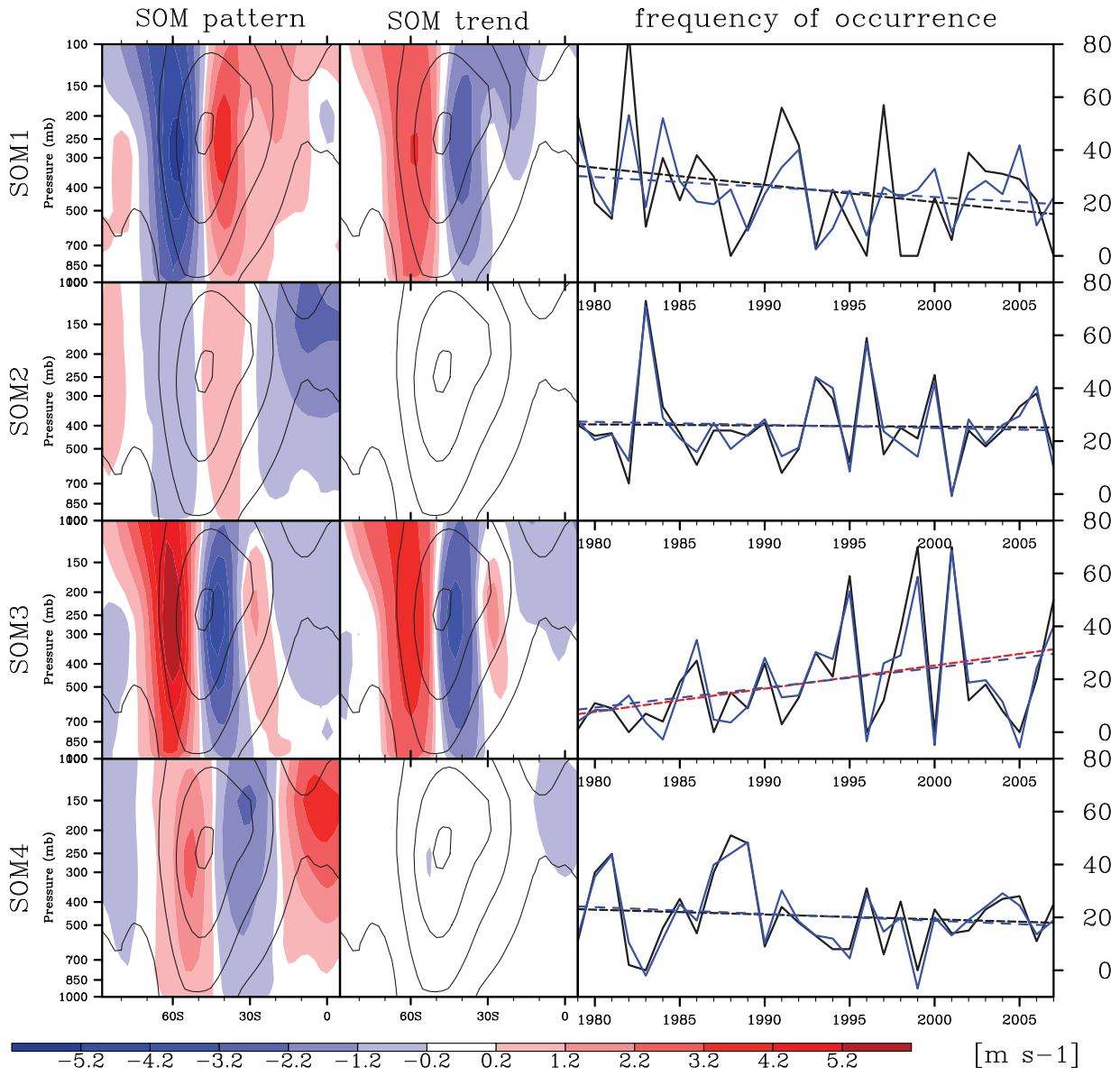


From Polvani et al. (Journal of Climate, 2011)

Self-organizing map (SOM) patterns, trend, and frequency of occurrence

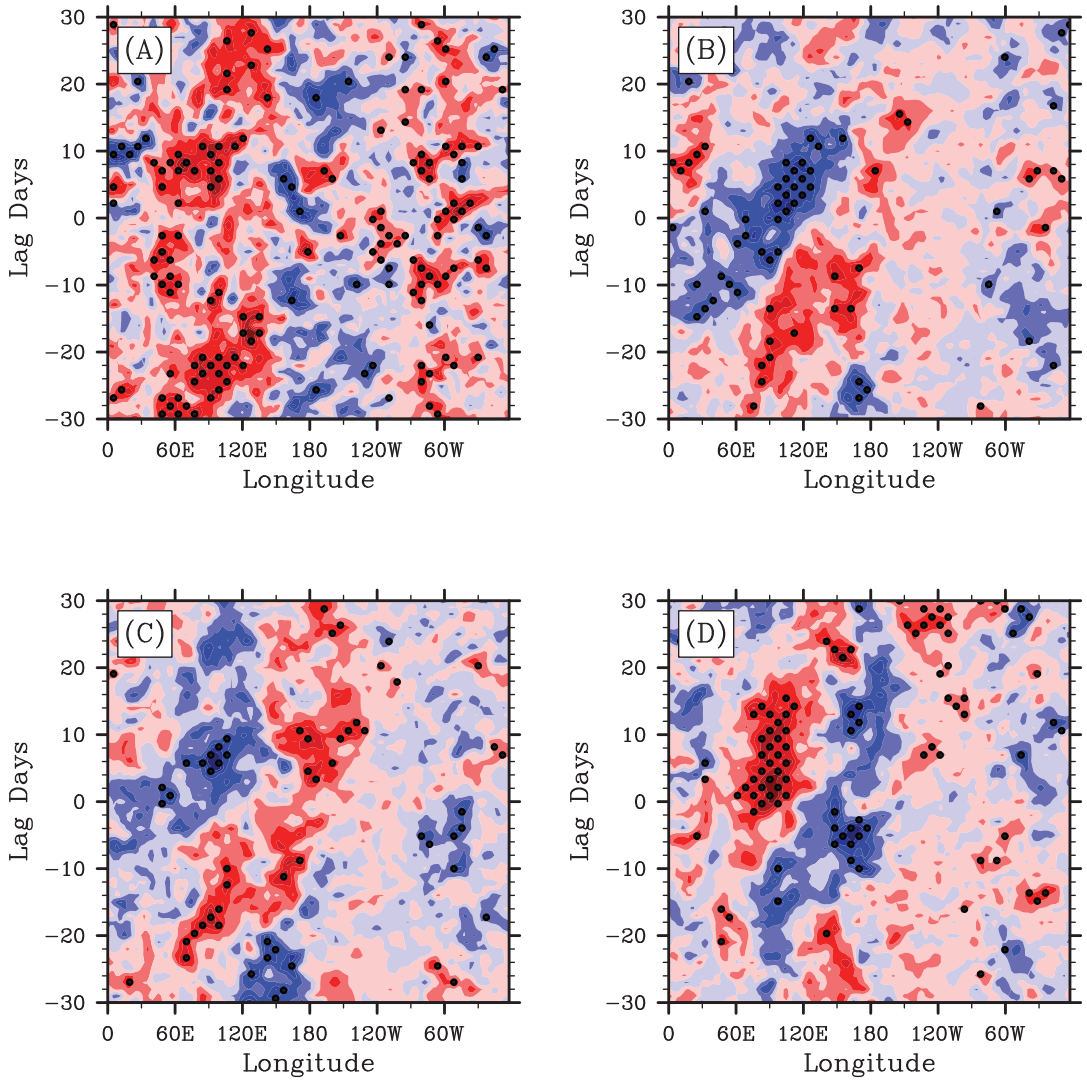
GHG

Ozone



7-11 day timescale for patterns

Anomalous OLR associated with all 4 SOM patterns

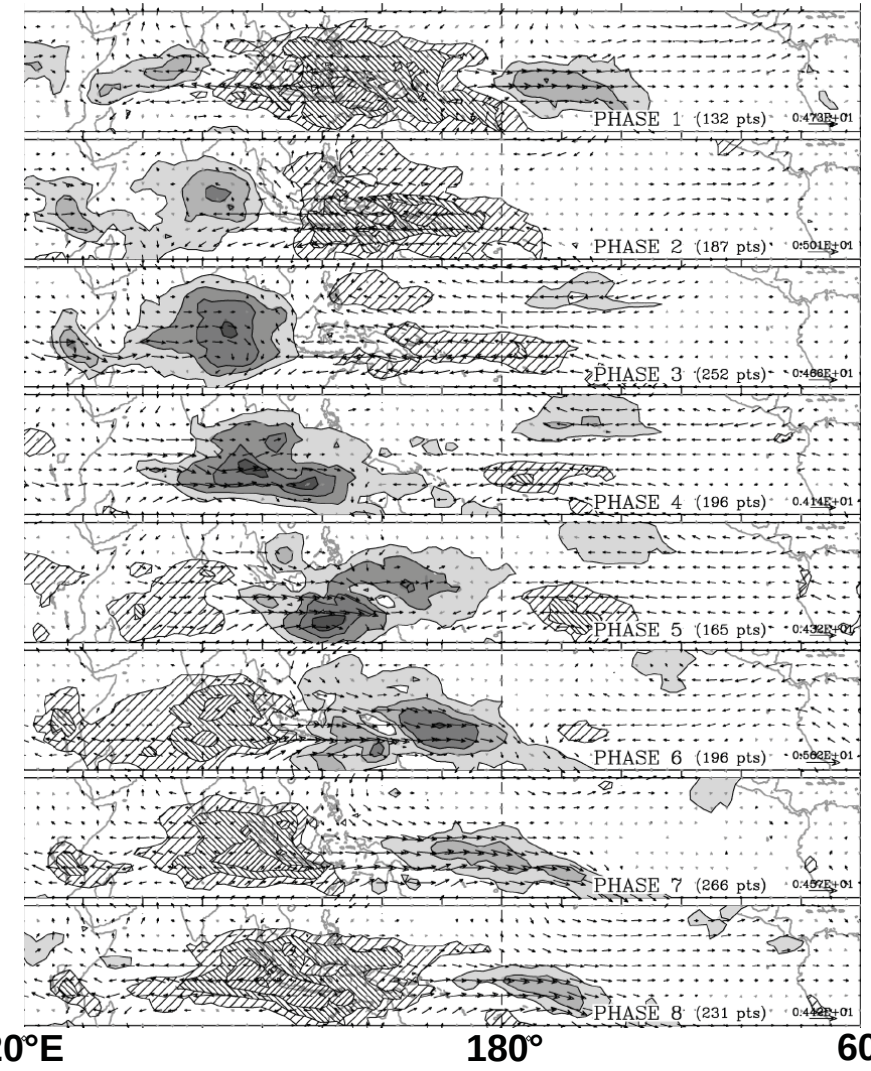


Tropical Convection Associated with the Madden-Julian Oscillation (MJO)

- **Dominant** intraseaonal oscillation in tropics
- MJO cycle: **30-60 days**

Negative outgoing longwave radiation (OLR) Anomalies (shaded) → Enhanced Tropical Convection

Time between Phases ~ 6 days



Phase 1

Phase 2

Phase 3

Phase 4

Phase 5

Phase 6

Phase 7

Phase 8

From Wheeler and Hendon (2004)

k-means clustering

- Unlike for SOMs, the cluster patterns are not organized on a two-dimensional grid.
- As with SOMs, cluster patterns arise from the minimization of Euclidean distance, i.e., to minimize J

$$J = \sum_{c=1}^{K} \sum_{\mathbf{z} \in S_c} |\mathbf{z} - \mathbf{m}_c|^2,$$

k-means clustering (additional points)

- May wish to repeat calculation several times and choose result with smallest J
- Sometimes convenient for visualization if combine with Linear Uni-dimensional Scaling

North Pacific SLP k-means cluster patterns

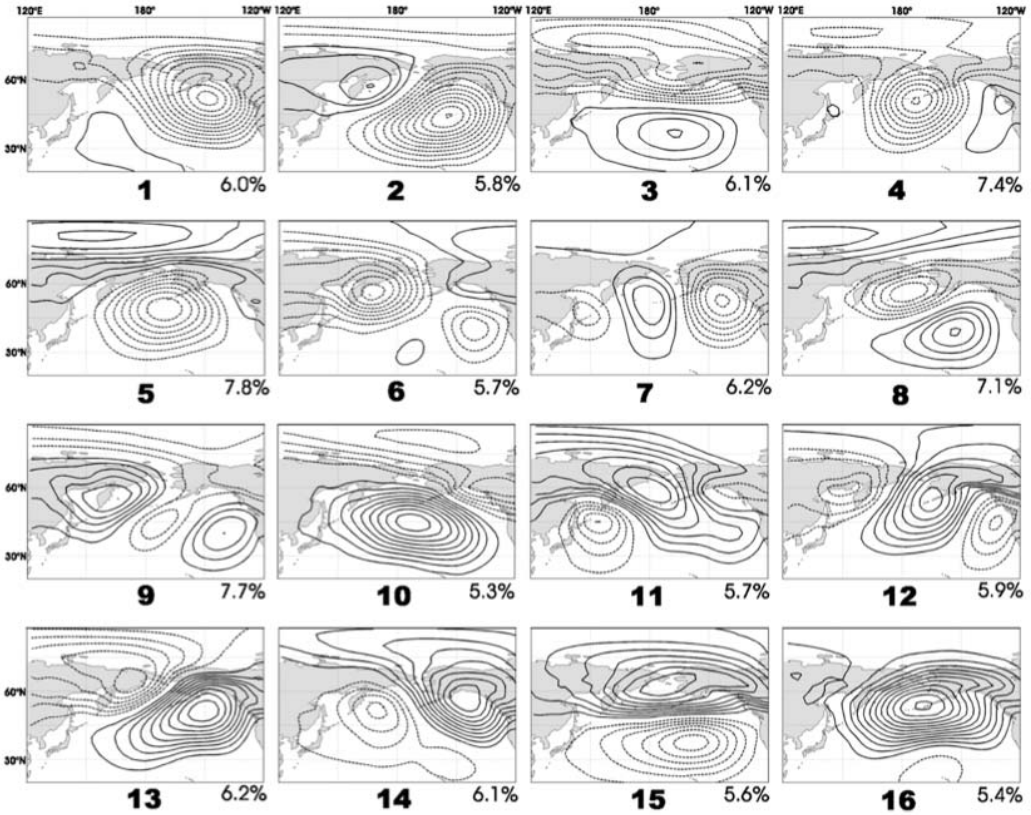


FIG. 1. The SLP anomaly patterns obtained by *k*-means cluster analysis, contoured at intervals of 2 hPa. Solid (dashed) lines depict positive (negative) values, with the zero contour omitted. The pattern number is displayed in bold below each pattern, and the percentages at the bottom right of each map describe the pattern frequency for the period 1958–2005.

Relationship between North Pacific SLP k-means cluster Patterns and the MJO (a weather forecast)

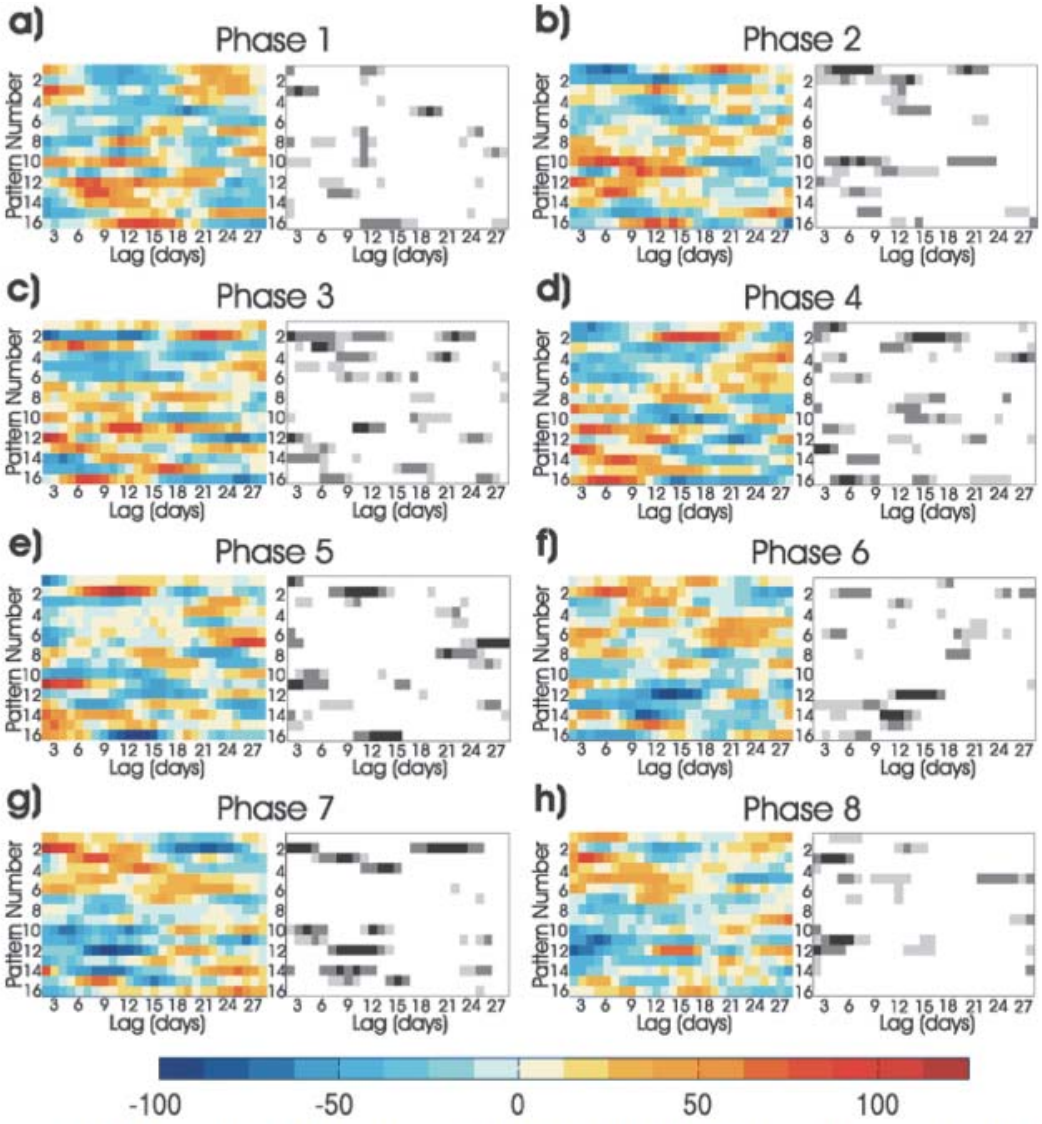


FIG. 4. The anomalous frequencies of occurrence for the SLP patterns in Fig. 1 for each of the eight phases of the MJO. In each panel, the colored plot corresponds with the anomalous frequency of occurrence for each pattern (numbered as in Fig. 1 on the *y* axis) as a function of lag (days) with respect to onset day: the plot on right side of each panel depicts the patterns and lags for which the frequency anomalies are statistically significant above the 90% (light gray), 95% (medium gray), and 99% (black) significance levels.

k-means clusters and climate variability

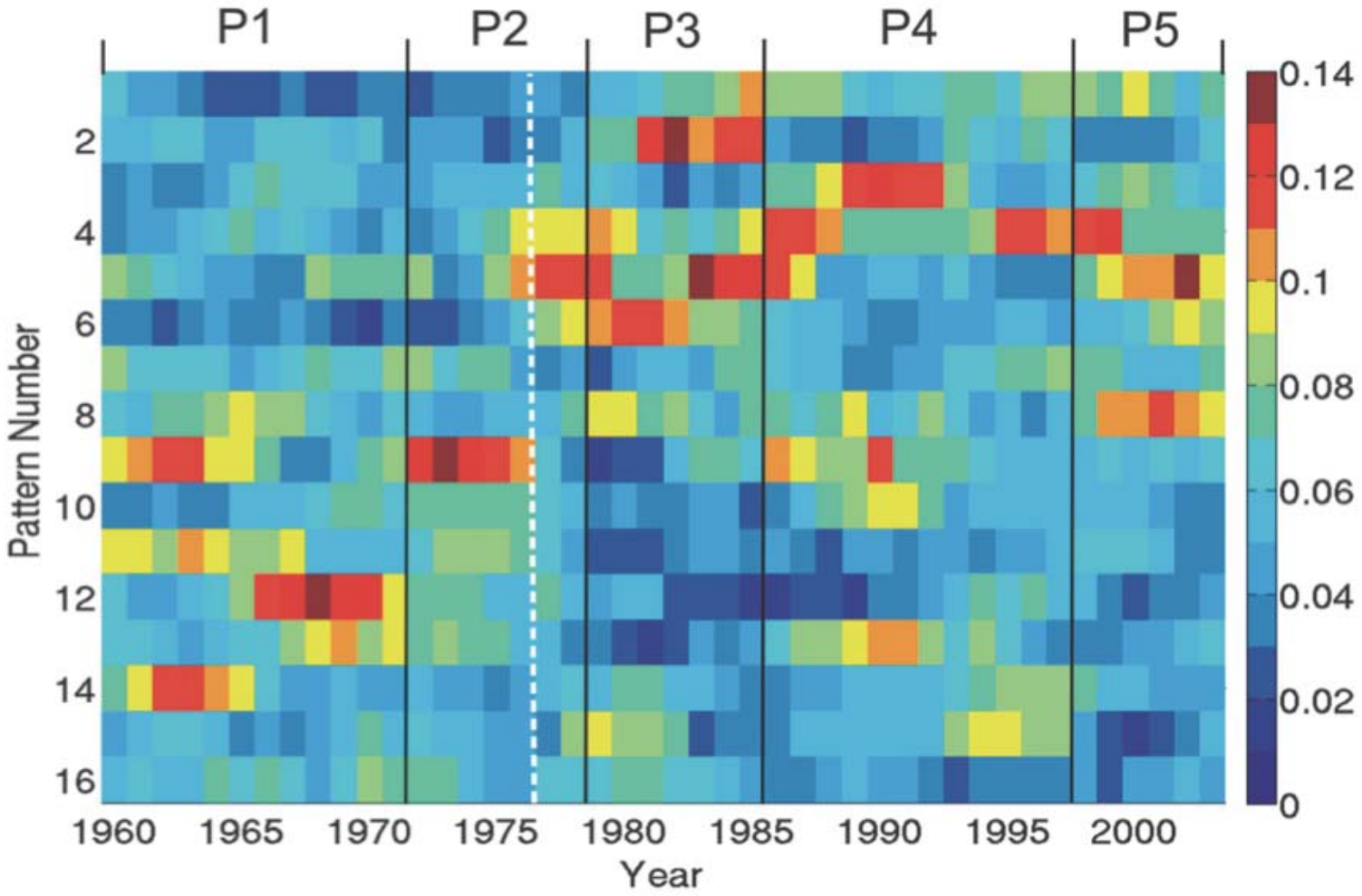


FIG. 5. The wintertime pattern frequency time series corresponding to each pattern depicted in Fig. 1. The colors describe the frequency of occurrence of each pattern for each winter season between 1958 and 2005: a 5-yr moving average has been applied to the frequency time series of each pattern. Five separate periods, P1–P5, are identified as periods between the thin vertical black lines, and the dashed white line corresponds to the Pacific regime shift at the beginning of 1977.

Examples of what can be learned with SOMs

A new explanation of the eastward shift of the NAO

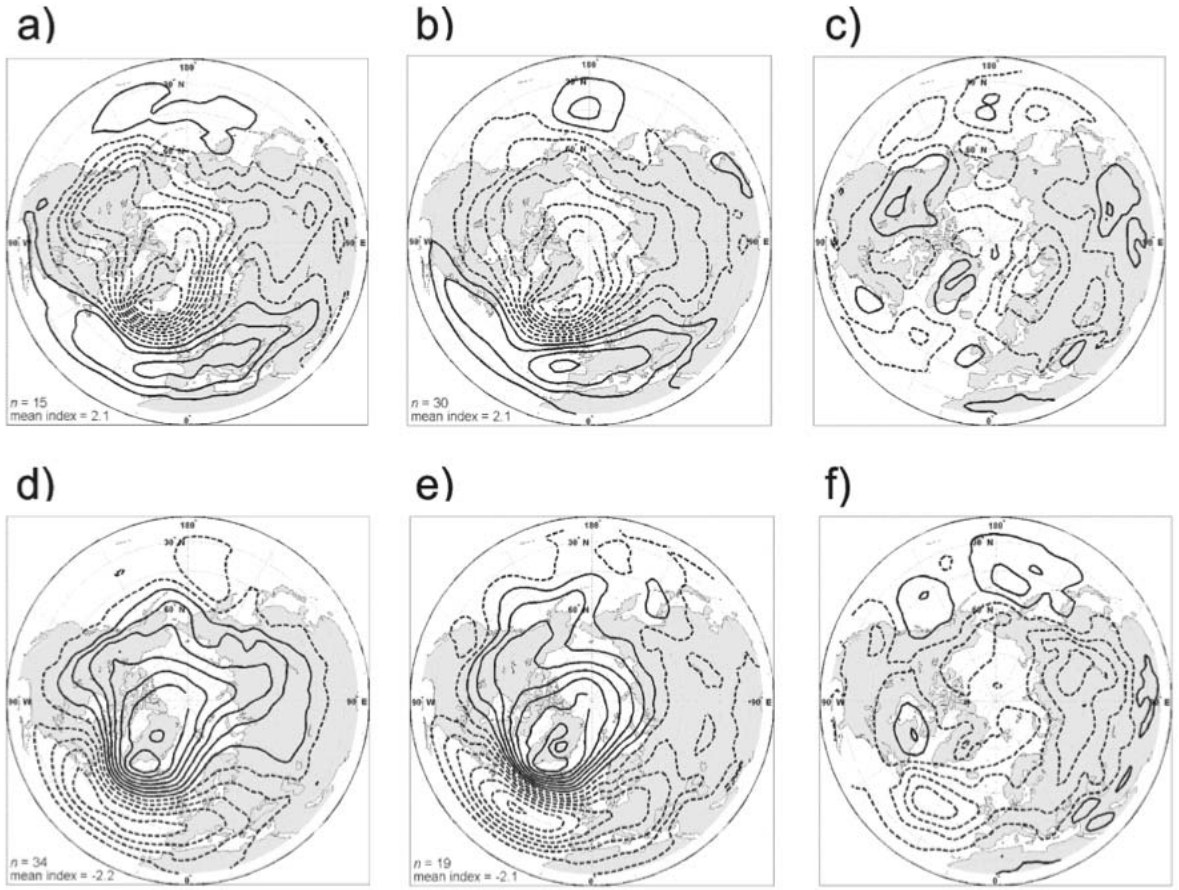


FIG. 6. The (a) SLP anomaly composites for positive NAO events during P1 and (b) P2 and for negative NAO events during (d) P1 and (e) P2. The difference between composites (P2 − P1) for (c) positive events and (f) negative: contour interval 3 hPa and zero contour omitted in all plots.

Matlab Code

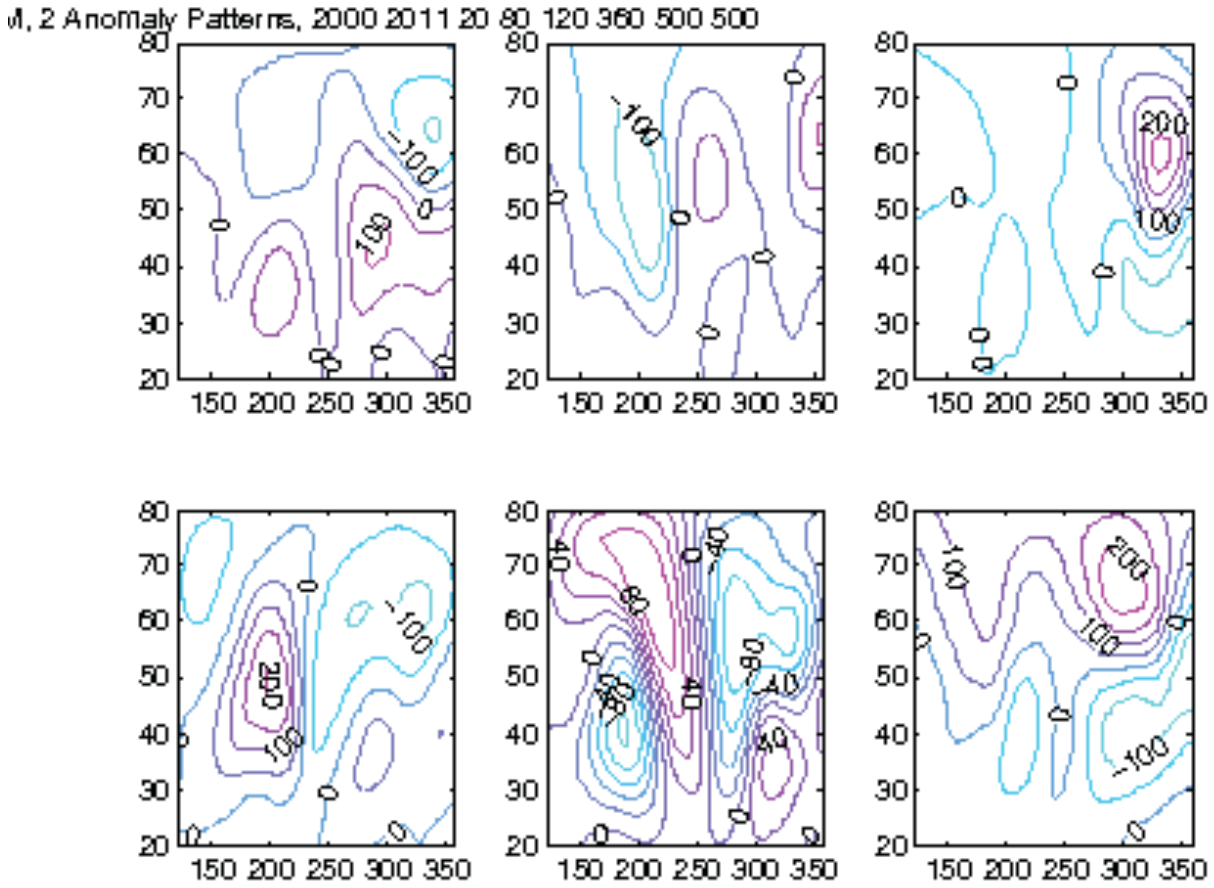
- addpath '/home/meteo/sbf1/cluster.analysis/somtoolbox/' (make sure to add SOM Toolbox)
- dir = '/kookaburra/s0/grayvireo/s3/sbf/hgt.nc (to access the Reanalysis data)
- year_begin = 2000; % First year in analysis
- year_end = 2011; % Last year in analysis
- season_days = [(1:60)'; (336:366)']; (Dec, Jan, Feb)
- latmax = 80; % 87.5 N
- latmin = 20; % 20 N
- lonmin = 120; % easternmost longitude (0 to 360)
- lonmax = 360; % westernmost longitude (0 to 360)
- smooth_cutoff = 21; % for removing seasonal cycle.

Matlab Code (continued)

- K = 6; % number of clusters
- num_rows = 2; % number of rows in SOM
- num_cols = 3; % number of columns in SOM
- cluster_method = 'som'; % choice is 'kmeans' or 'som'
- Calculate anomaly (deviation from seasonal cycle)
- SOM or k-means calculation performed
- Plotting and writing output
- serid = netcdf.defVar(ncid,'timeseries','double',[daydimid eledimid]); (create netCDF variable)
- netcdf.putVar(ncid,serid,[0 0],[num_day 4],timeseries);

  (writing data to netCDF variable)
- netcdf.putAtt(ncid,serid,'four columns of timeseries', ... 'year day best_matching_pattern rms_err'); (write netCDF attributes)

2X3 SOM map (500-hPa height)



2X3 SOM map (SLP)