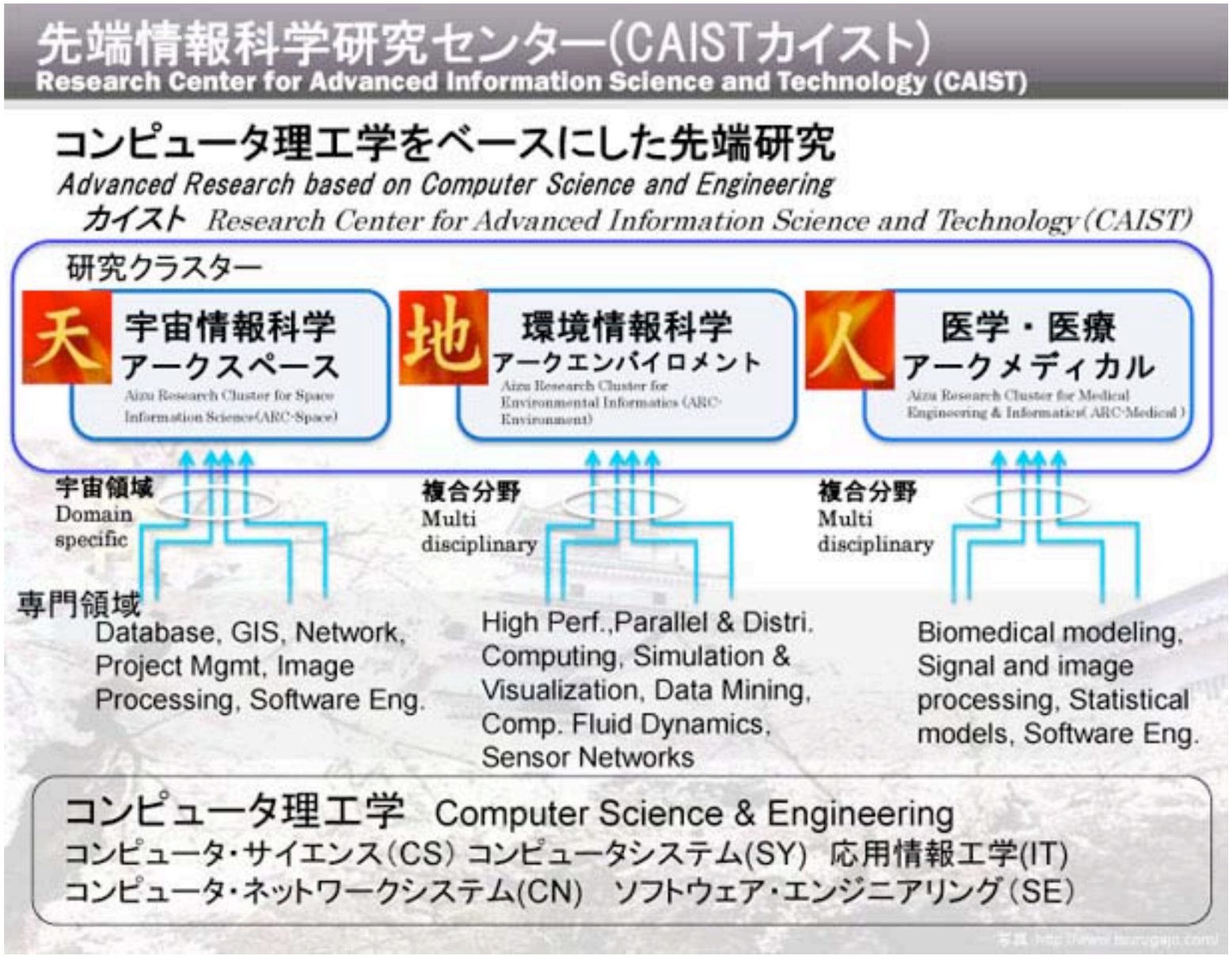
Self Organizing Maps - an alternative method for studying climate variability

Saji N Hameed
CAIST/ARC-ENV
University of Aizu, Fukushima, Japan

CAIST

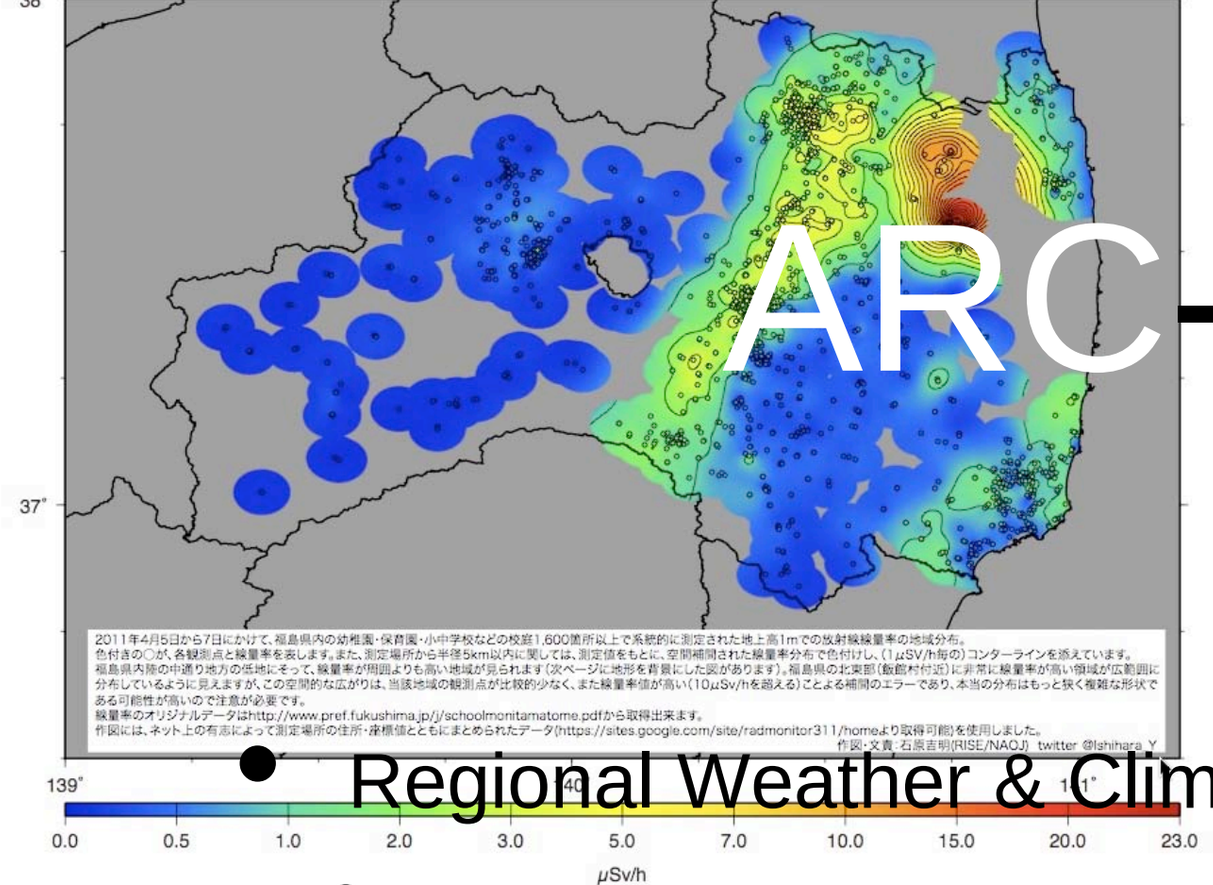

先端情報科学研究センター(CAISTカイスト)
Research Center for Advanced Information Science and Technology (CAIST)
コンピュータ理工学をベースにした先端研究
Advanced Research based on Computer Science and Engineering
カイスト Research Center for Advanced Information Science and Technology (CAIST)
研究クラスター
天
宇宙情報科学
アークスペース
Aizu Research Cluster for Space Information Science (ARC-Space)
地
環境情報科学
アークエンバイロメント
Aizu Research Cluster for Environmental Informatics (ARC-Environment)
人
医学・医療
アークメディカル
Aizu Research Cluster for Medical Engineering & Informatics (ARC-Medical)
宇宙領域
Domain specific
複合分野
Multi disciplinary
複合分野
Multi disciplinary
専門領域
Database, GIS, Network, Project Mgmt, Image Processing, Software Eng.
High Perf.,Parallel & Distri. Computing, Simulation & Visualization, Data Mining, Comp. Fluid Dynamics, Sensor Networks
Biomedical modeling, Signal and image processing, Statistical models, Software Eng.
コンピュータ理工学 Computer Science & Engineering
コンピュータ・サイエンス(CS) コンピュータシステム(SY) 応用情報工学(IT)
コンピュータ・ネットワークシステム(CN) ソフトウェア・エンジニアリング(SE)

The Clusters

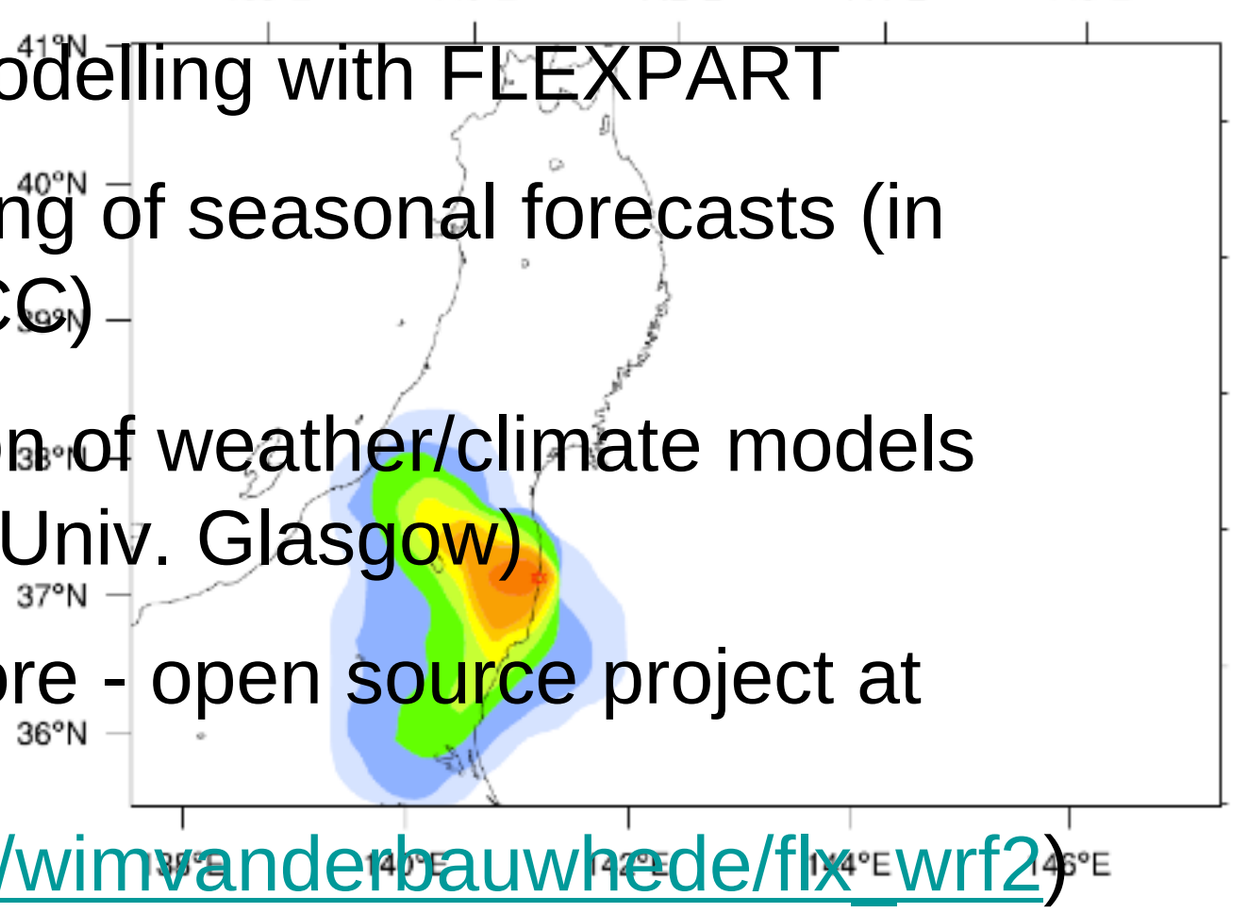
Members

Research Cluster	Name
Aizu Research Cluster for Space Science (ARC-Space)	Hirohide Demura, Associate Professor
	Naru Hirata, Assistant Professor
	Yoshiko Ogawa, Assistant Professor
	Chikatoshi Honda, Assistant Professor
	Kohei Kitazato, Assistant Professor
	Junya Terazono, Assistant Lecturer
Aizu Research Cluster for Local Environment and Informatics (ARC-Environment)	Haruo Terasaka, Professor
	Hameed Saji N., Associate Professor
	Takeaki Sanpe, Assistant Professor
Aizu Research Cluster for Medical Engineering and Informatics (ARC-Medical)	Shinya Oku, Professor

ARC-ENV



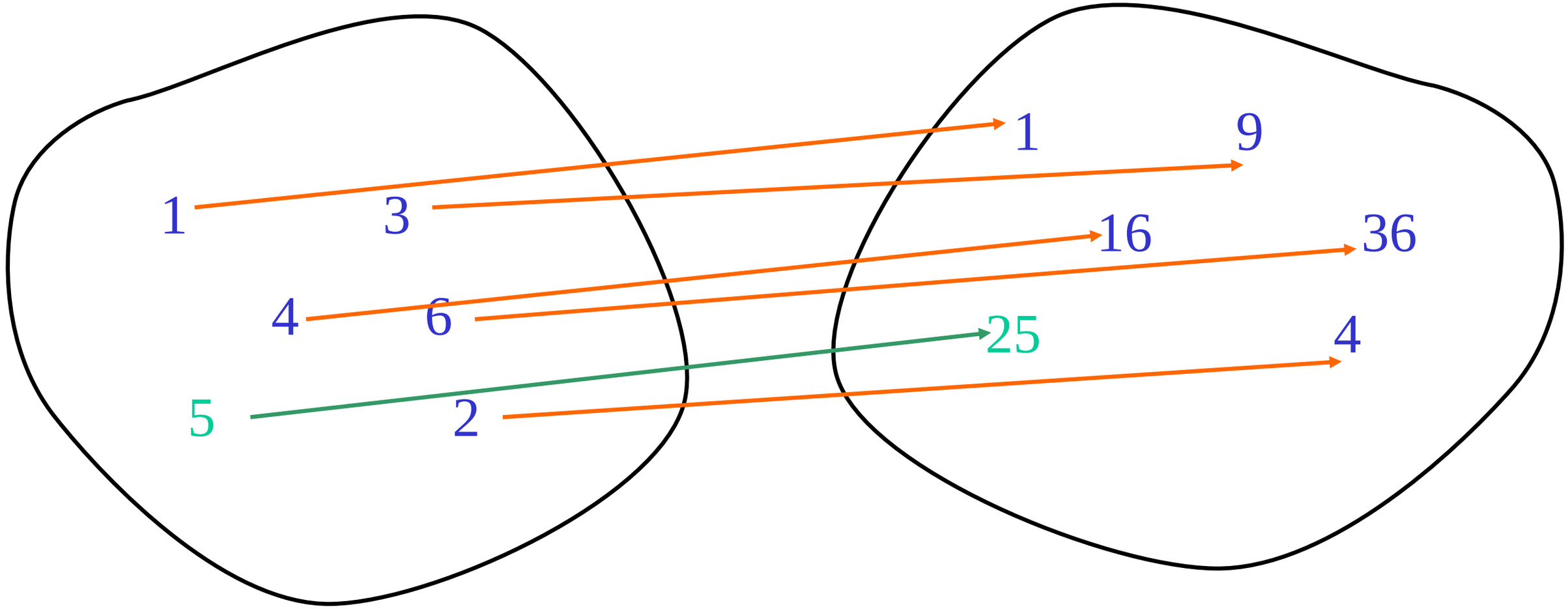
- Regional Weather & Climate
- WRF-based operational weather forecast
- Particle dispersion modelling with FLEXPART
- Dynamical downscaling of seasonal forecasts (in cooperation with APCC)
- Multi-core acceleration of weather/climate models (in collaboration with Univ. Glasgow)
  - Flexpart on multicore - open source project at github (https://github.com/wimvanderbauwhede/flx_wrf2)



Outline

- Clustering + EOF analysis
- A high level overview of SOM
- Advantages/Disadvantages
- Examples
- Software

Supervised/unsupervised Learning



Supervised Learning

- When a set of targets of interest is provided by an external teacher

  we say that the learning is Supervised

- The targets usually are in the form of an input output mapping that the net should learn

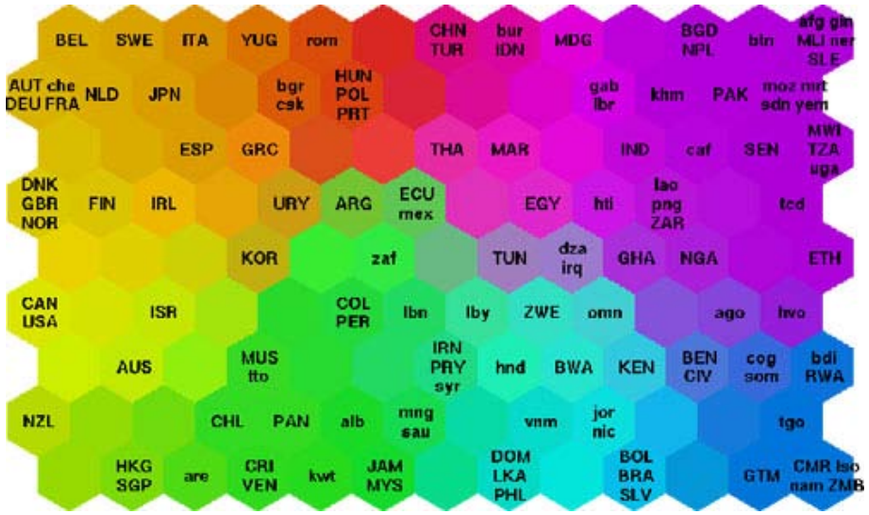
Unsupervised learning

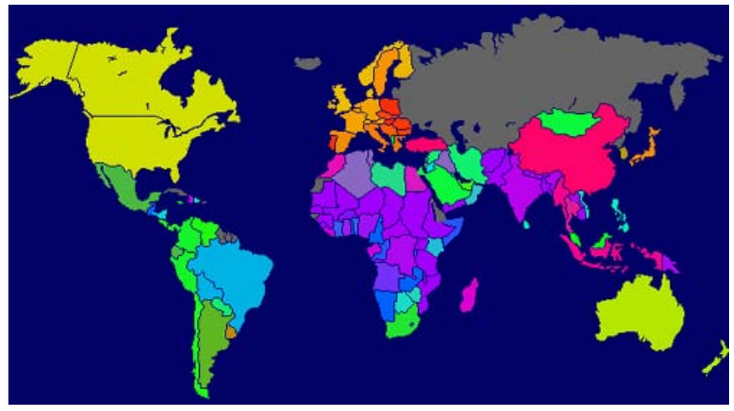
- Many times there is no “teacher” to tell us how to do things
  - A baby that learns how to walk
  - Grouping of events into a meaningful scene (making sense of the world)
  - Development of ocular dominance and orientation selectivity in our visual system

SOM - at a glance

- A Self-Organizing Map (SOM) is a way to represent higher dimensional data in an usually 2-D or 3-D manner, such that similar data is grouped together.
- It runs unsupervised and performs the grouping on its own.
- Once the SOM converges, it can only classify new data. It is unlike traditional neural nets which are continuously learning and adapting.
- SOMs run in two phases:
  - Training phase: map is built, network organizes using a competitive process, it is trained using large numbers of inputs (or the same input vectors can be administered multiple times).
  - Mapping phase: new vectors are quickly given a location on the converged map, easily classifying or categorizing the new data.

SOM - example 1





- Example: Data sets for poverty levels in different countries.
  - Data sets have many different statistics for each country.
  - SOM does not show poverty levels, rather it shows how similar the poverty sets for different countries are to each other. (Similar color = similar data sets).

SOM - example 2

```
13 rect 8 5 gaussian
1 0 0 1 0 0 0 0 1 0 0 1 0 Dove
1 0 0 1 0 0 0 0 1 0 0 0 0 Chick
1 0 0 1 0 0 0 0 1 0 0 0 1 Duck
1 0 0 1 0 0 0 0 1 0 0 1 1 Goose
1 0 0 1 0 0 0 0 1 1 0 1 0 Owl
1 0 0 1 0 0 0 0 1 1 0 1 0 Hawk
0 1 0 1 0 0 0 0 1 1 0 1 0 Eagle
0 1 0 0 1 1 0 0 0 1 0 0 0 Fox
0 1 0 0 1 1 0 0 0 0 1 0 0 Dog
0 1 0 0 1 1 0 1 0 1 1 0 0 Wolf
1 0 0 0 1 1 0 0 0 1 0 0 0 Cat
0 0 1 0 1 1 0 0 0 1 1 0 0 Tiger
0 0 1 0 1 1 0 1 0 1 1 0 0 Lion
0 0 1 0 1 1 1 1 0 0 1 0 0 Horse
0 0 1 0 1 1 1 1 0 0 1 0 0 Zebra
0 0 1 0 1 1 1 0 0 0 0 0 0 Cow
```

```
Hameed-Sajis-MacBook-Pro:SOMPAK_4R sajis$ ruby test_s
   1/   1 sec. .....................................
alpha = 0.700000
```

Lion Horse Zebra			Owl Hawk Eagle
	Cow		
Tiger			Dove
Fox Dog Wolf	Cat		Chicken Duck Goose

Topology preserving

- In the human cortex, multi-dimensional sensory input spaces (e.g., visual input, tactile input) are represented by two-dimensional maps.
- The projection from sensory inputs onto such maps is topology conserving.
- This means that neighboring areas in these maps represent neighboring areas in the sensory input space.
- For example, neighboring areas in the sensory cortex are responsible for the arm and hand regions.

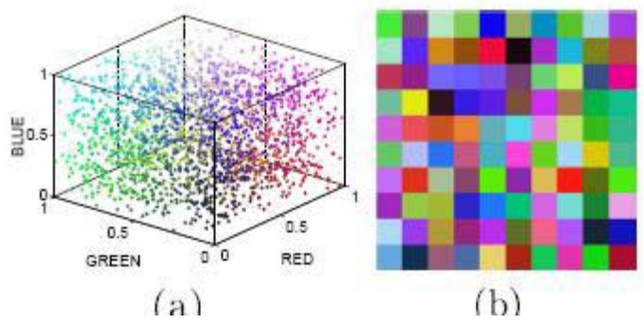
Self Organizing networks

- Discover significant patterns or features in the input data
- Discovery is done without a teacher
- Synaptic weights are changed according to local rules
- The changes affect a neuron's immediate environment

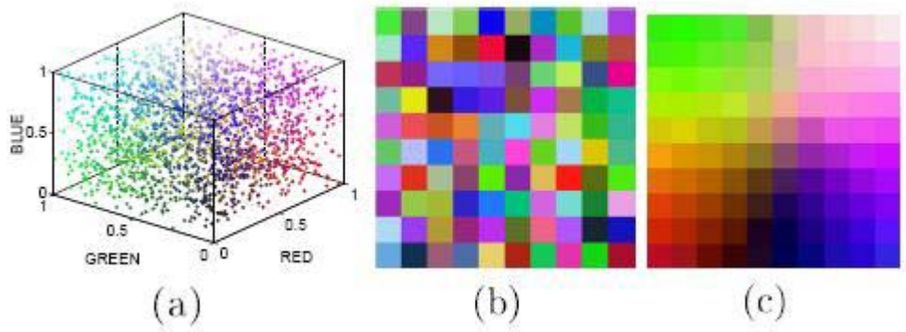
  until a final configuration develops

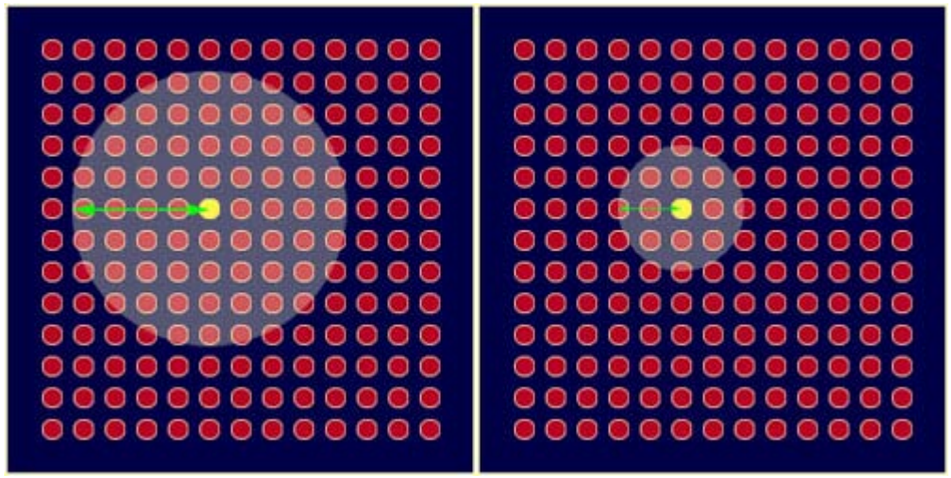
The algorithm

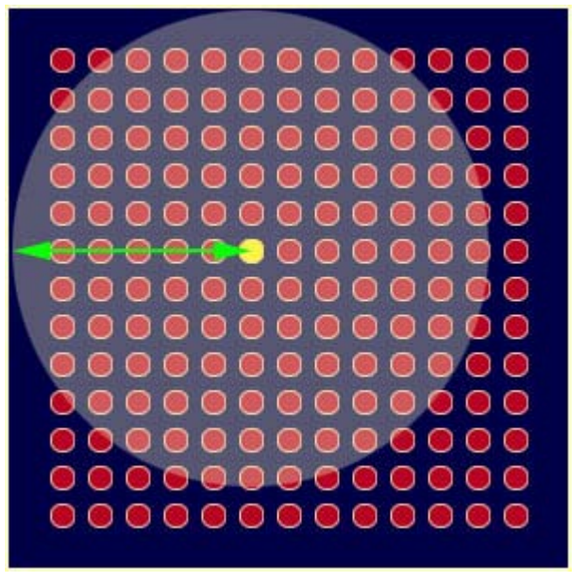
1) Initialize each node's weights.
2) Choose a random vector from training data and present it to the SOM.
3) Every node is examined to find the Best Matching Unit (BMU).
4) The radius of the neighborhood around the BMU is calculated. The size of the neighborhood decreases with each iteration.
5) Each node in the BMU's neighborhood has its weights adjusted to become more like the BMU. Nodes closest to the BMU are altered more than the nodes furthest away in the neighborhood.
6) Repeat from step 2 for enough iterations for convergence.



SOM algorithm







$$Dist = \sqrt{\sum_{i=0}^{i=n} (V_i - W_i)^2}$$

$$W(t+1) = W(t) + L(t)(V(t) - W(t))$$

$$L(t) = L_0 \exp\left(-\frac{t}{\lambda}\right) \qquad t = 1, 2, 3\ldots$$

BMU

- Calculating the BMU is done according to the Euclidean distance among the node's weights (W_1, W_2, … , W_n) and the input vector's values (V_1, V_2, … , V_n).
  - This gives a good measurement of how similar the two sets of data are to each other.

$$Dist = \sqrt{\sum_{i=0}^{i=n}(V_i - W_i)^2}$$

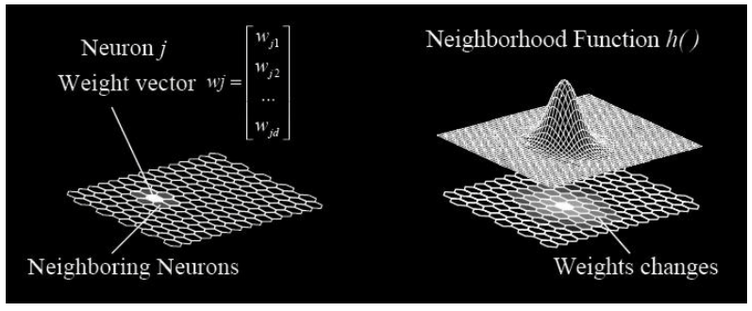
Determining BMU

- Size of the neighborhood: We use an exponential decay function that shrinks on each iteration until eventually the neighborhood is just the BMU itself.

$$\sigma(t) = \sigma_0 \; exp\left(-\frac{t}{\lambda}\right)$$

- Effect of location within the neighborhood: The neighborhood is defined by a gaussian curve so that nodes that are closer are influenced more than farther nodes.

$$\Theta(t) = exp\left(-\frac{dist^2}{2\sigma^2(t)}\right)$$



modifying nodes

- The new weight for a node is the old weight, plus a fraction (L) of the difference between the old weight and the input vector… adjusted (theta) based on distance from the BMU.

$$W(t+1) = W(t) + \Theta(t)L(t)\,(V(t) - W(t))$$

- The learning rate, L, is also an exponential decay function.
  - This ensures that the SOM will converge.

$$L(t) = L_0\ exp\left(-\frac{t}{\lambda}\right)$$

- The lambda represents a time constant, and t is the time step

Self Organization

- Network organization takes place at 2 levels that interact with each other:
  - Activity: certain activity patterns are produced by a given network in response to input signals
  - Connectivity: synaptic weights are modified in response to neuronal signals in the activity patterns
- Self Organization is achieved if there is positive feedback between changes in synaptic weights and activity patterns

Principles of Self Organization

Modifications in synaptic weights tend to self amplify

Limitation of resources lead to competition among synapses

Modifications in synaptic weights tend to cooperate

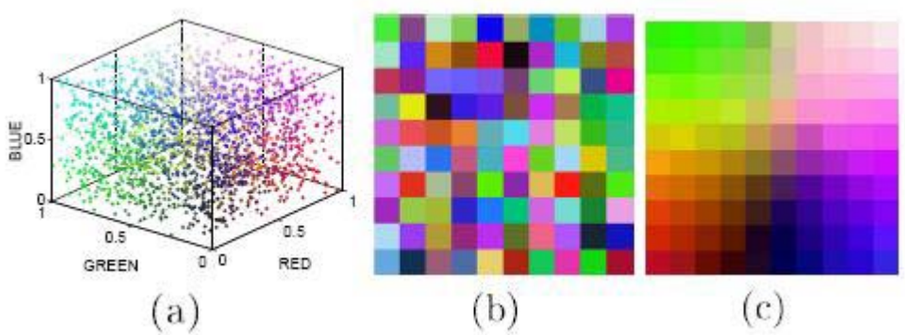
Order and structure in activation patterns represent redundant information that is transformed into knowledge by the network

Redundancy

- Unsupervised learning depends on redundancy in the data
- Learning is based on finding patterns and extracting features from the data

SOM map of RGB color space

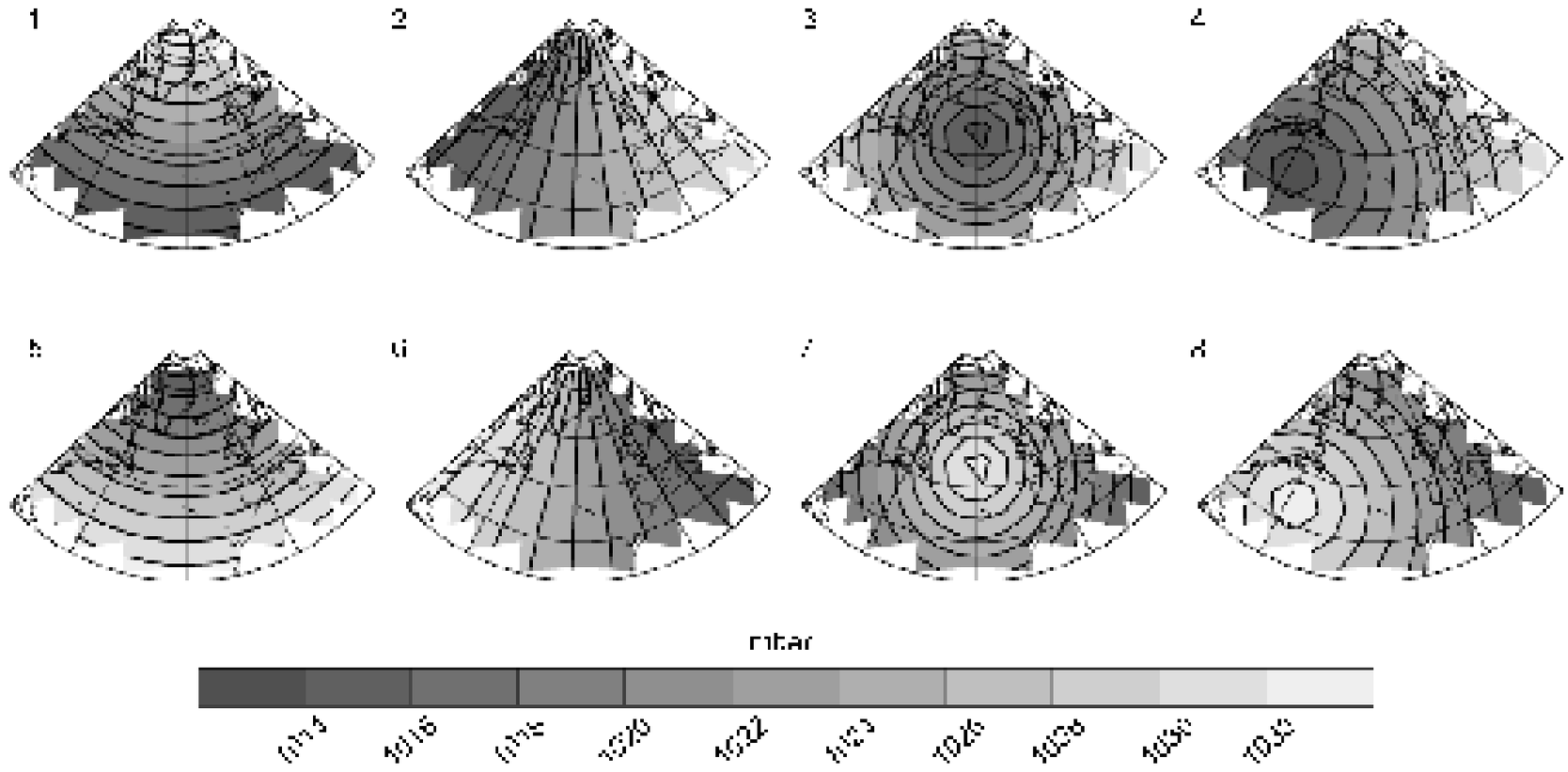
- 2-D square grid of nodes.
- Inputs are colors.
- SOM converges so that similar colors are grouped together.
- Program with source code and pre-compiled Win32 binary: http://www.ai-junkie.com/files/SOMDemo.zip or mirror.



a) Input space
b) Initial weights
c) Final weights

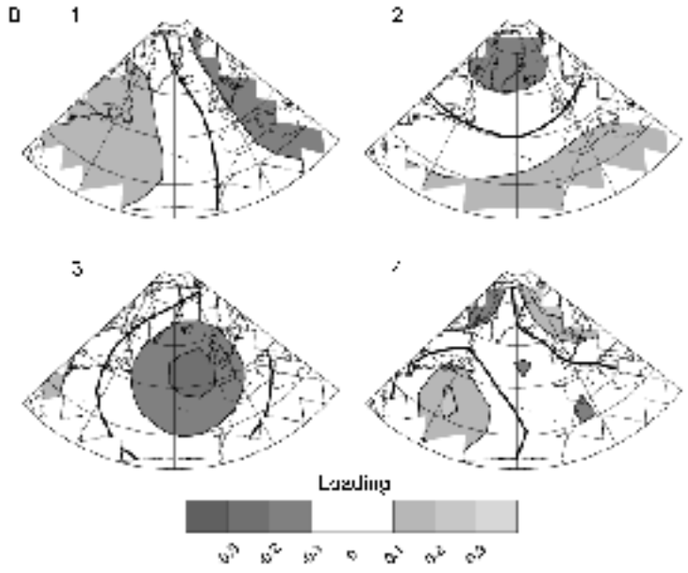
SOM as an alternative to EOF analysis

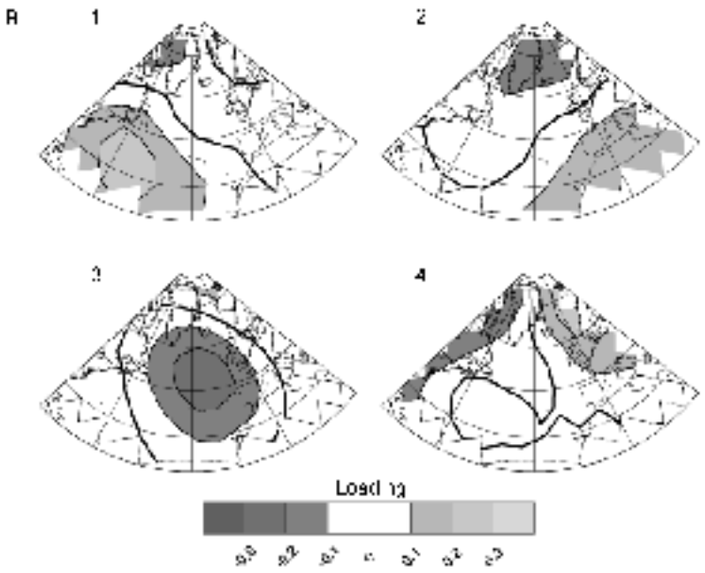
Synthetic data with artificial patterns

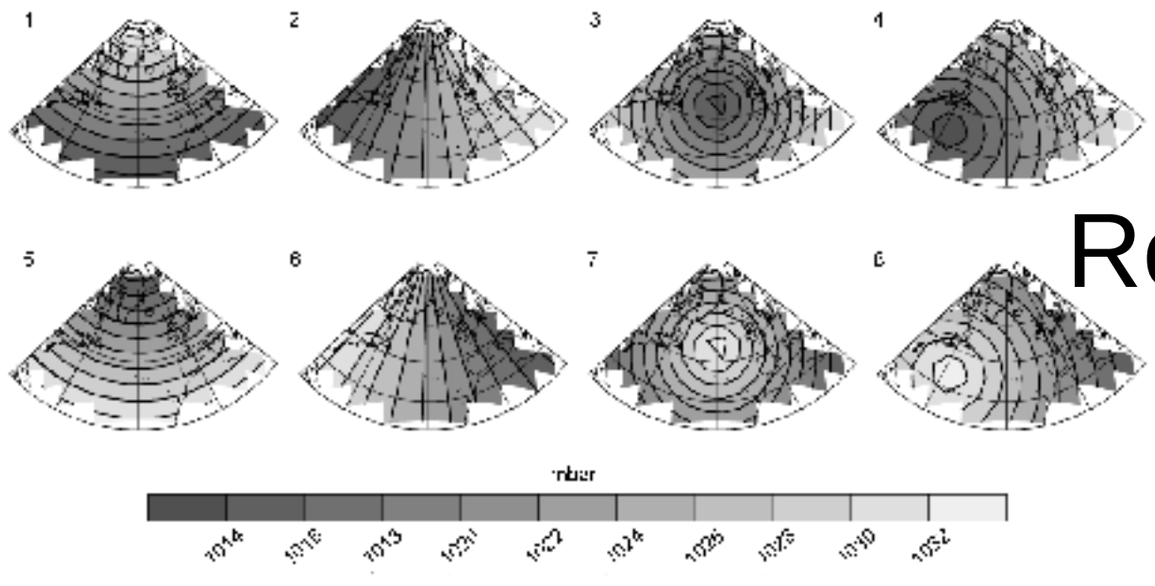


Reusch et al 2005

Rotated and unrotated EOFs of synthetic data







Reusch et al 2005

Variance distribution

TABLE 2

Principal Component Analysis Results[a]

PC	Dataset 1			Dataset 2			Dataset 3		
	EigVal	Var[b]	Tot[b]	EigVal	Var	Tot	EigVal	Var	Tot
1	40.3	44.3	44.3	38.8	42.7	42.7	35.0	38.5	38.5
2	30.2	33.2	77.5	28.9	31.8	74.5	25.8	28.4	66.8
3	19.3	21.2	98.7	18.3	20.1	94.6	16.0	17.5	84.4
4	1.2	1.3	100.0	1.1	1.3	95.9	1.2.0	1.3	85.7
Residual[c]			0			4.1[c]			14.3[d]

TABLE 3

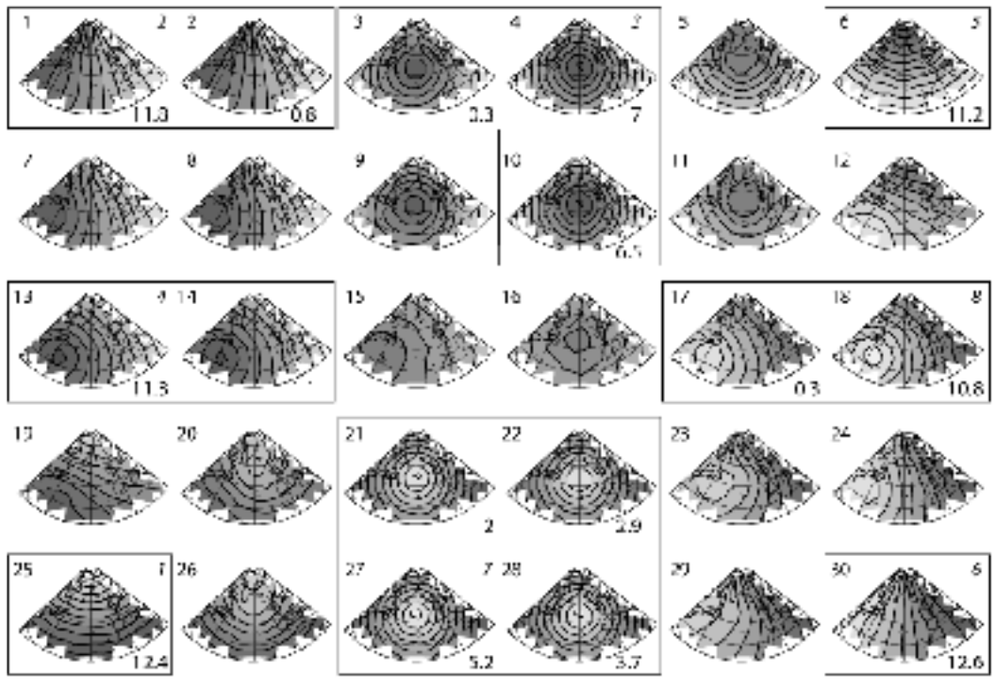
Rotated Principal Component Analysis Results[a]

PC	Dataset 1			Dataset 2			Dataset 3		
	EigVal	Var	Tot	EigVal	Var	Tot	EigVal	Var	Tot
1	22.7	24.9	24.9	22.4	25.7	25.7	20.4	26.2	26.2
2	26.9	29.6	54.5	25.7	29.4	55.1	23.2	29.7	55.9
3	18.7	20.6	75.1	17.9	20.5	75.6	15.5	19.8	75.7
4	22.6	24.9	100.0	21.2	24.3	99.9	18.9	24.2	99.9

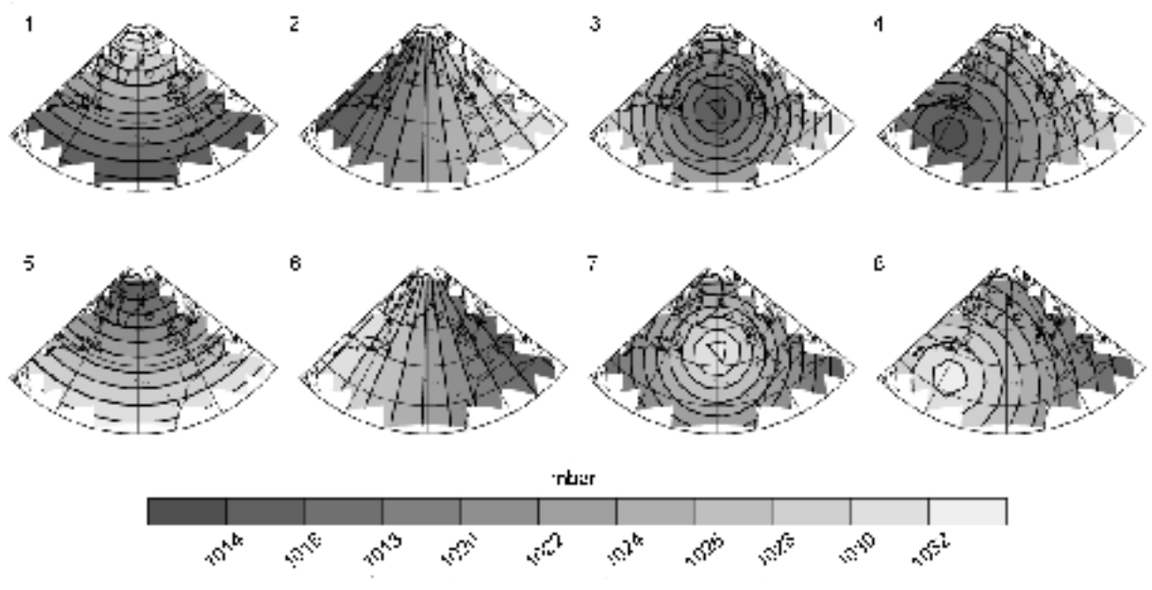
[a]PC = principal component; EigVal = eigenvalue, estimated from the variance of the rotated PC scores; Var = variance (percent); tot = total variance (percent).

Reusch et al 2005

SOM classification



Reusch et al 2005



Estimating relative skill of SOM

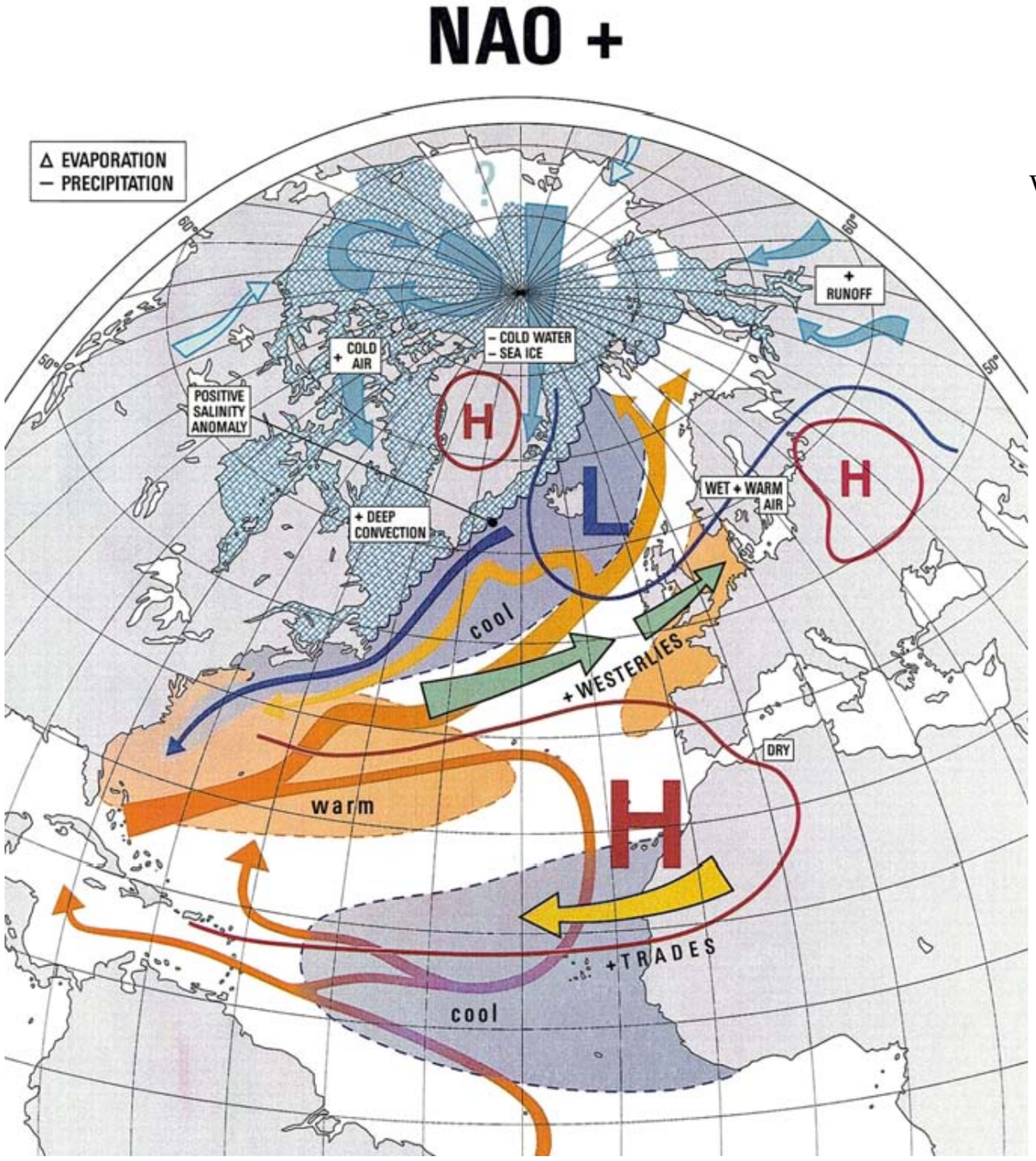
Relative Skills of SOMs for Dataset 3 Expressed as a Percentage Error Reduction[a]

Predefined pattern	Linear vs. random				Trained vs. random				Trained vs. linear			
	4 × 3	5 × 3	5 × 4	6 × 5	4 × 3	5 × 3	5 × 4	6 × 5	4 × 3	5 × 3	5 × 4	6 × 5
1	85	86	79	72	78	97	98	98	-47	79	88	93
2	75	76	76	75	82	80	91	95	26	16	62	80
3	60	60	58	61	69	95	98	98	23	86	95	95
4	85	83	81	84	80	81	91	95	-26	-14	53	66
5	85	84	78	71	66	96	97	98	-123	73	86	92
6	76	79	75	76	83	82	92	98	32	17	69	92
7	60	57	59	60	81	95	98	98	52	87	96	96
8	85	82	79	83	83	79	91	98	-11	-18	56	88

Reusch et al 2005

SOM for climate variability

SOM analysis of NAO variability



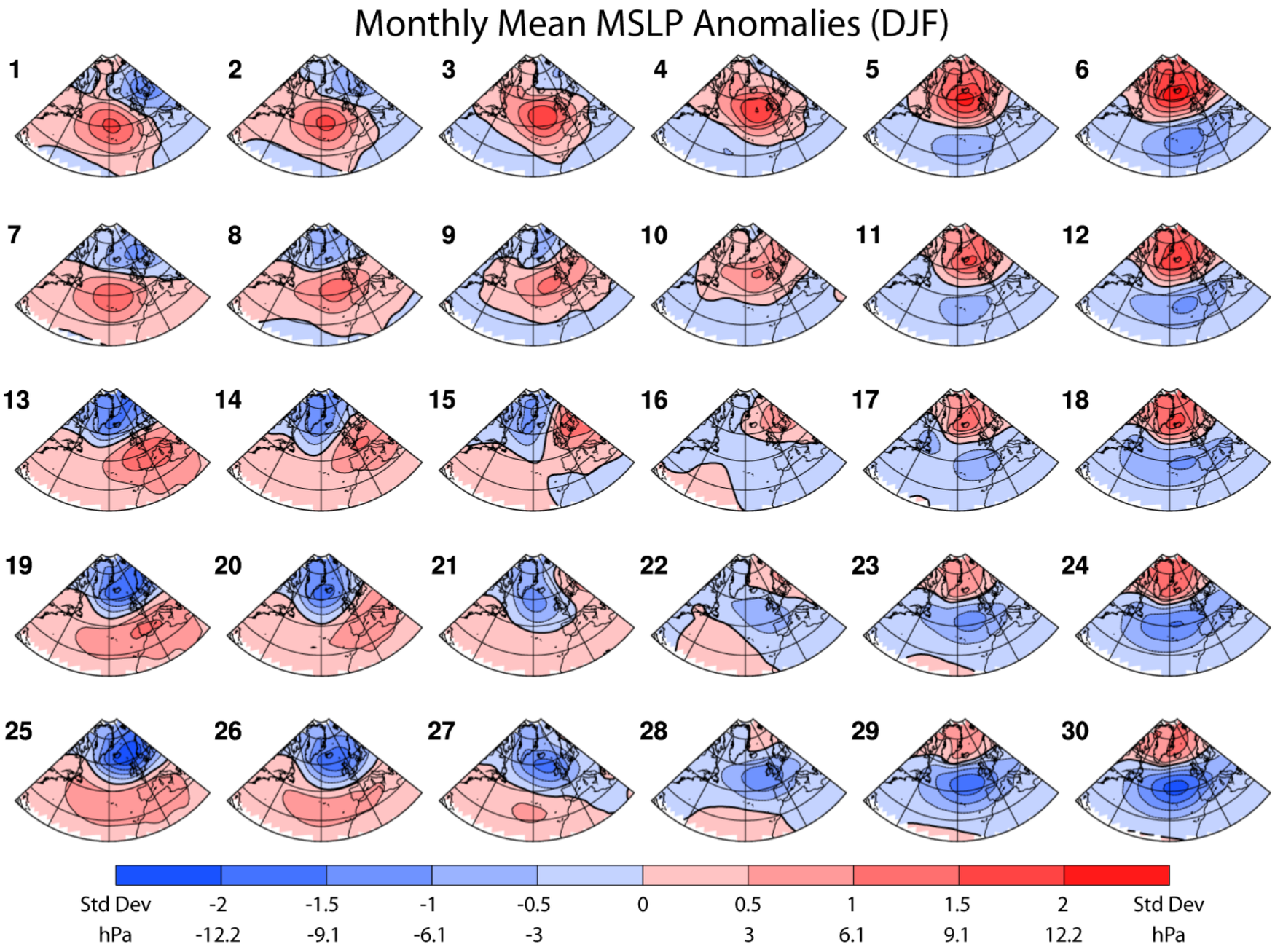
Wanner et al, Surveys in Geophysics, 22: 321-382, 2001.

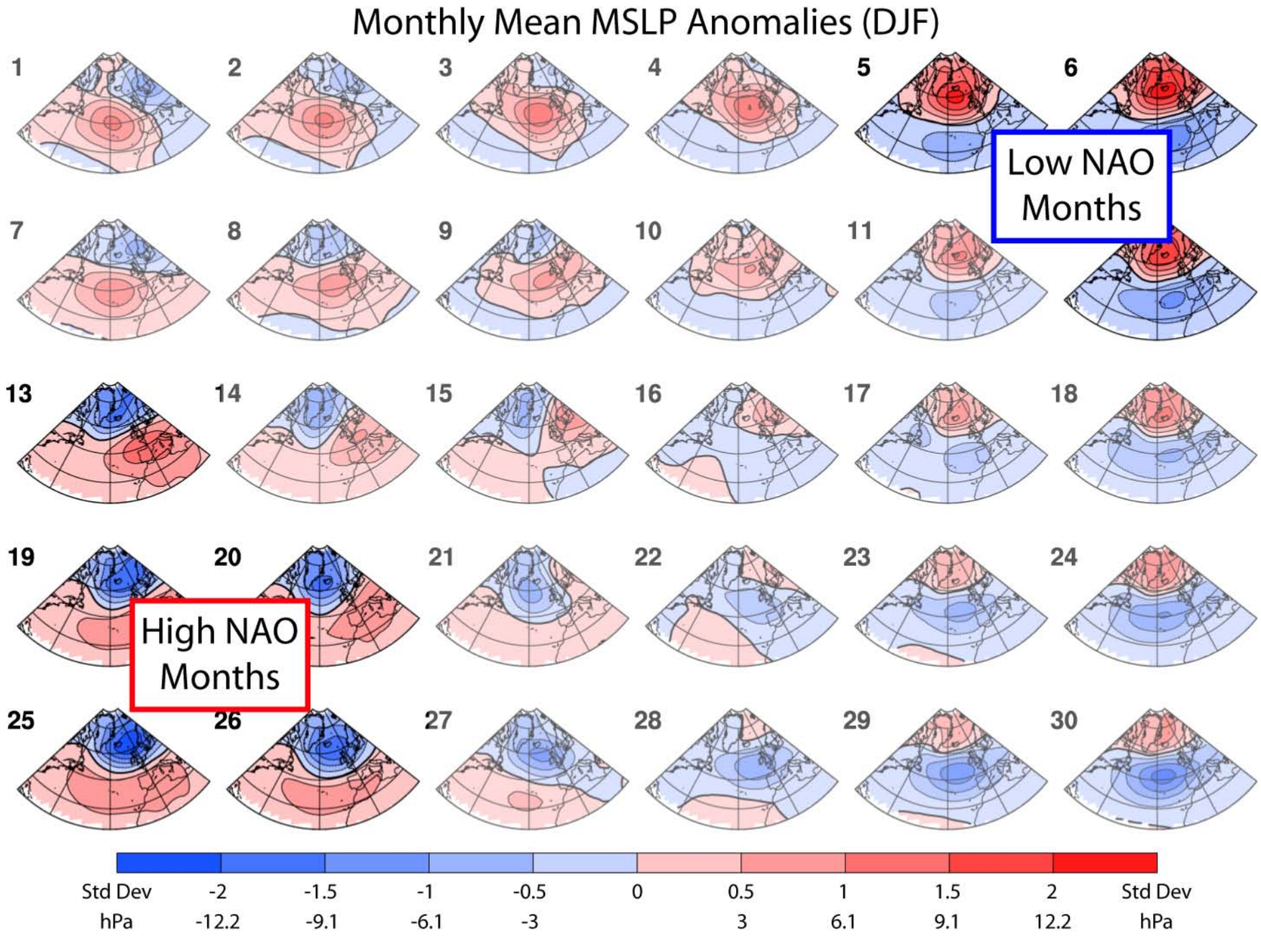
SOM analysis of NAO

- Focus on NAO (North Atlantic Oscillation)
  Dec-Jan-Feb monthly data, 1957-2001
  ECMWF 45 yr reanalysis (ERA-40)
  20-85° N, 80° W - 25° E
  Anomalies from 1971-2000 baseline

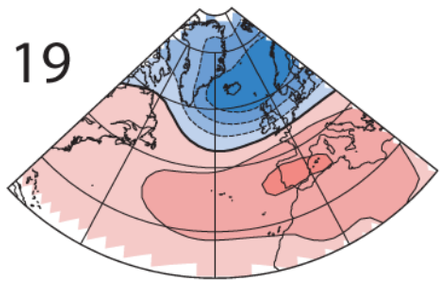
Monthly mean, standard deviation for MSLP, T-2m, $(U, V, Z)_{500}$

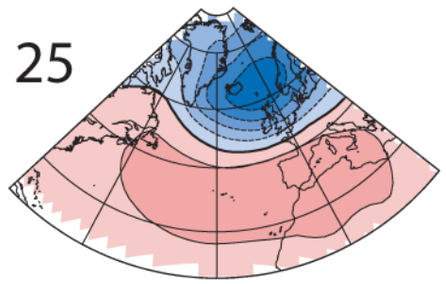
SOM Map of NAO

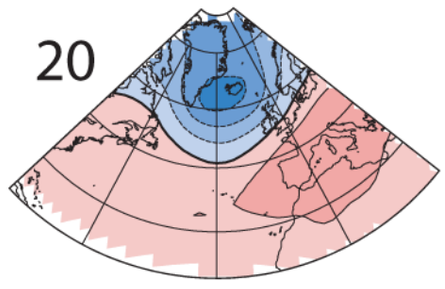


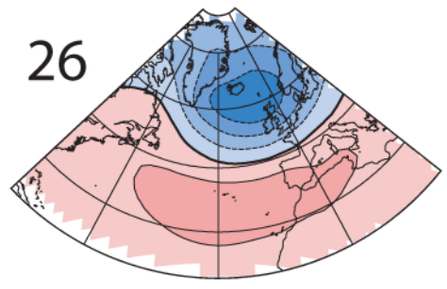

Monthly Mean MSLP Anomalies (DJF)
1
2
3
4
5
6
Low NAO Months
7
8
9
10
11
13
14
15
16
17
18
19
20
High NAO Months
21
22
23
24
25
26
27
28
29
30
Std Dev -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 Std Dev
hPa -12.2 -9.1 -6.1 -3 3 6.1 9.1 12.2 hPa

A Set of Positive NAO Patterns

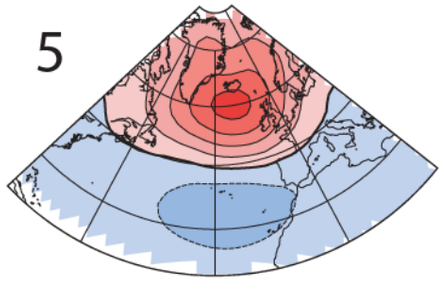


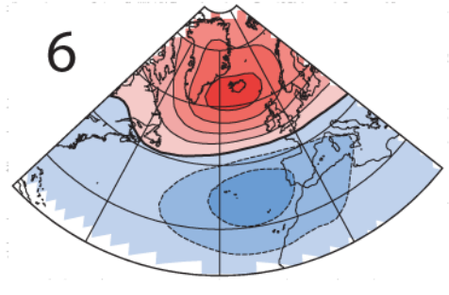


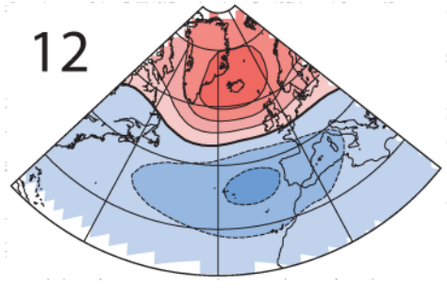


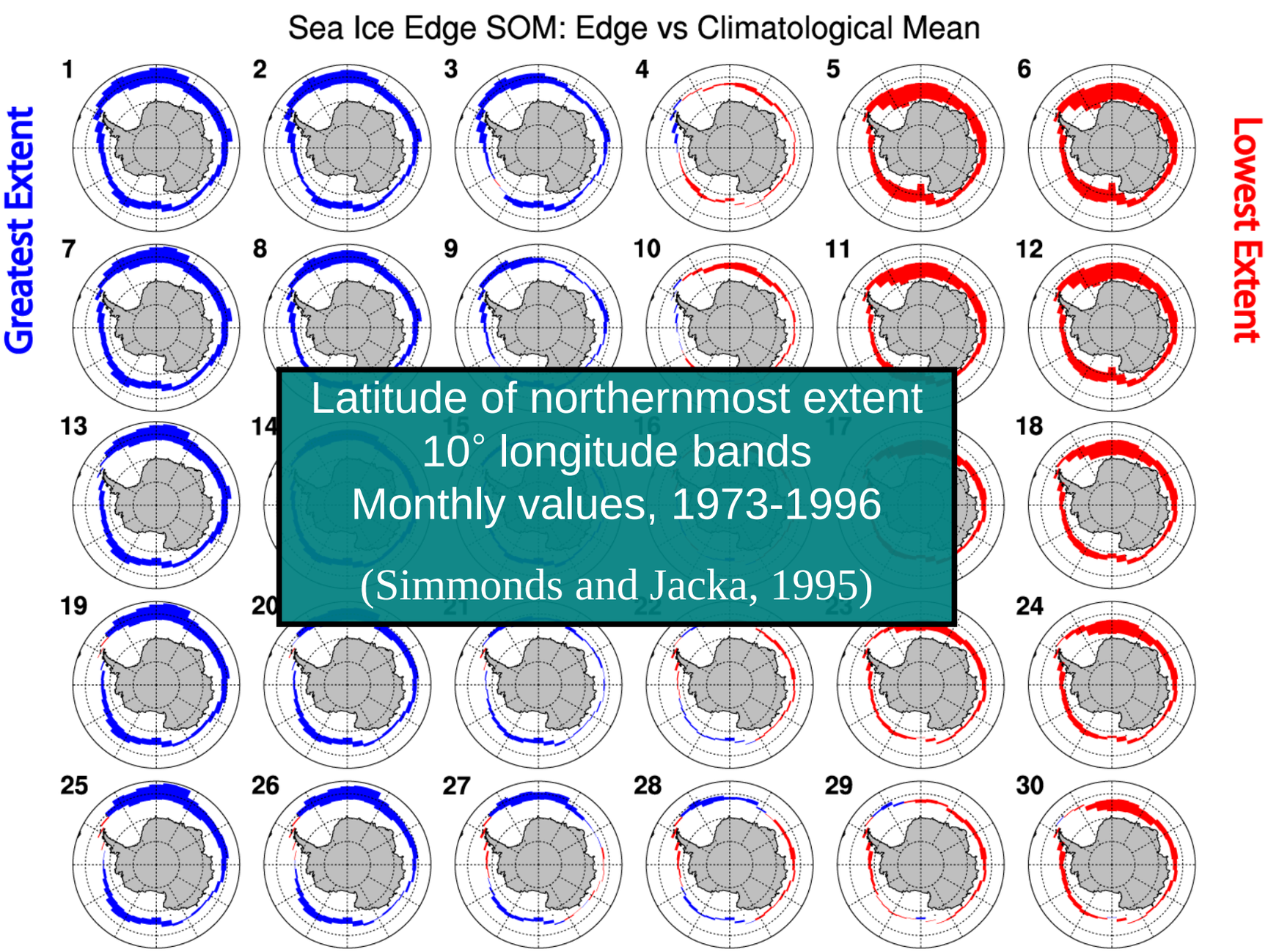


A Set of Negative NAO Patterns

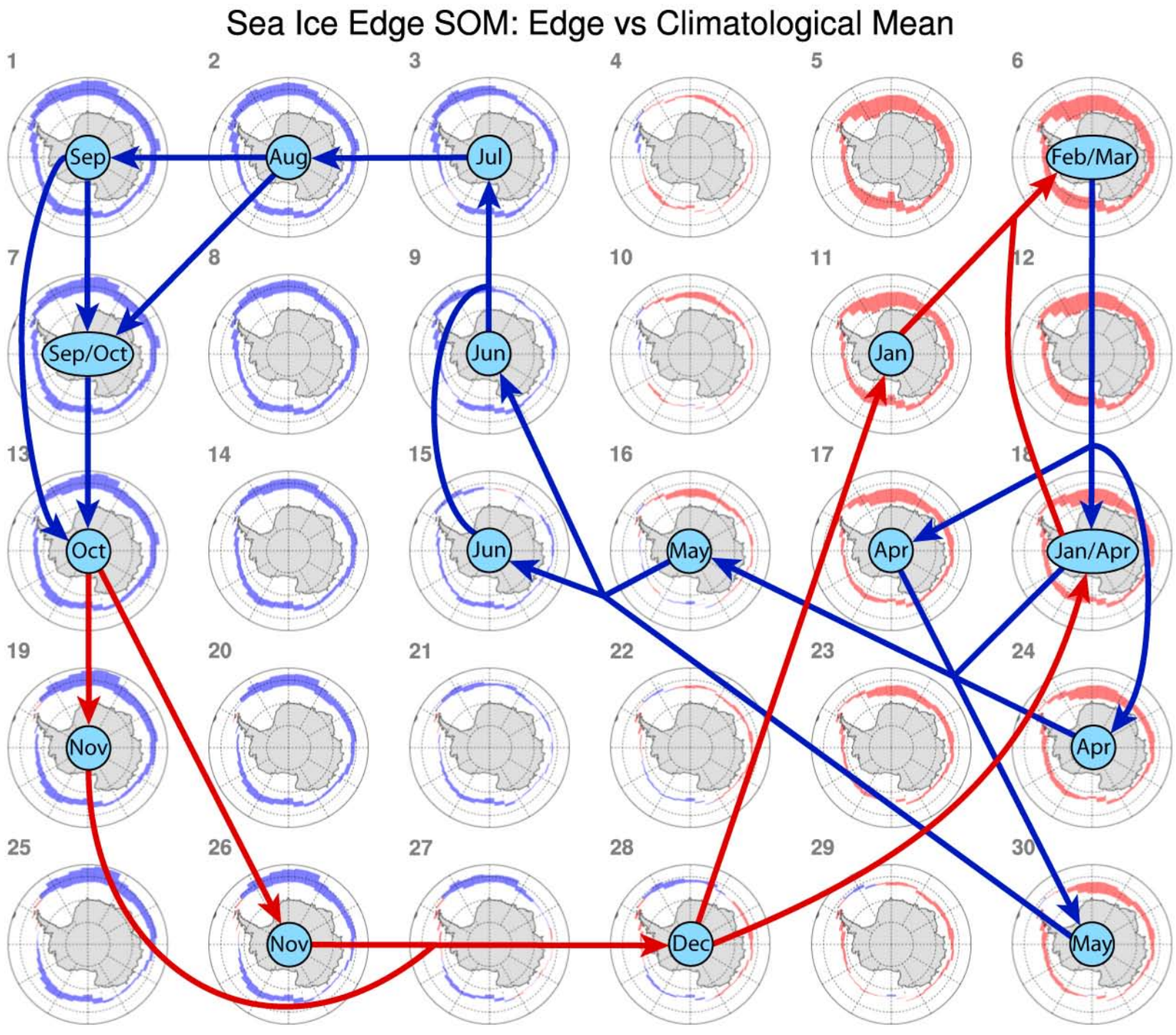




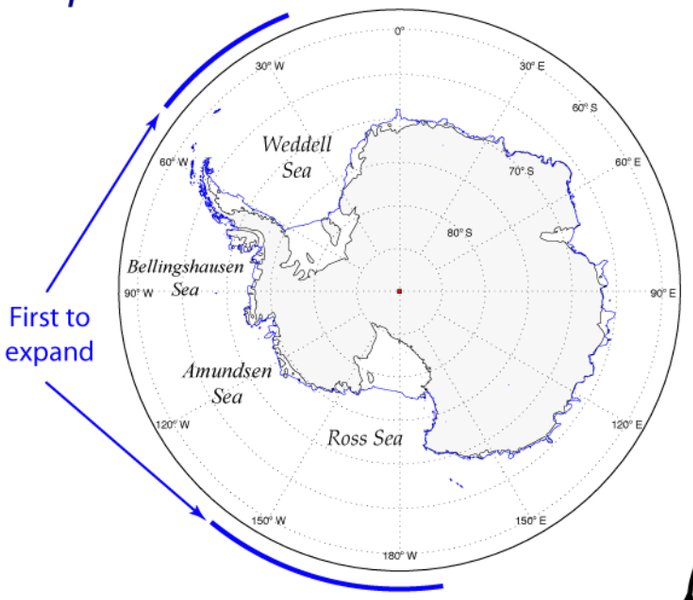



Sea Ice Edge SOM: Edge vs Climatological Mean
Greatest Extent
Lowest Extent
1
2
3
4
5
6
7
8
9
10
11
12
13
14
18
19
20
24
25
26
27
28
29
30
Latitude of northernmost extent
10° longitude bands
Monthly values, 1973-1996
(Simmonds and Jacka, 1995)

Sea Ice edge SOM

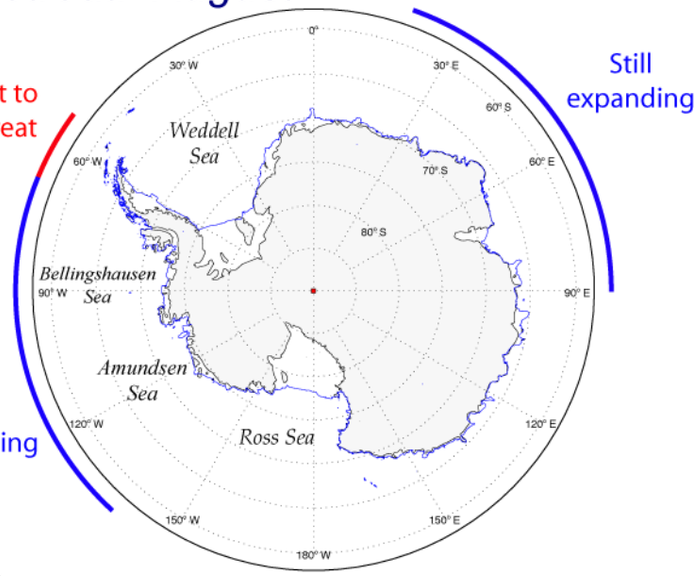


Expansion: March to June



First to expand

Retreat: August

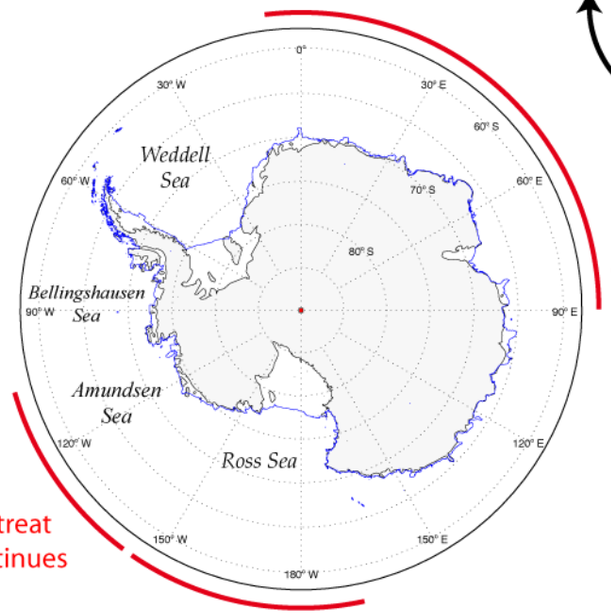


First to retreat

Still expanding

Still expanding

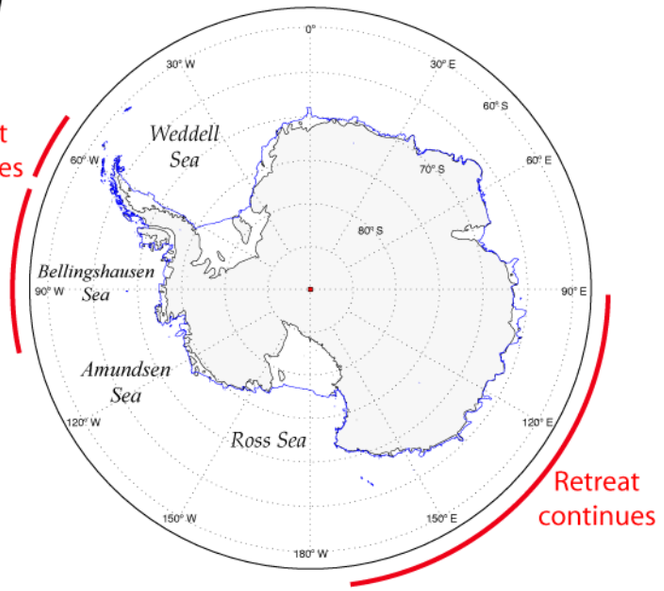
Annual Cycle



Retreat continues

Retreat continues

Retreat: December to February



Retreat continues

Retreat continues

Retreat: September to November

Object

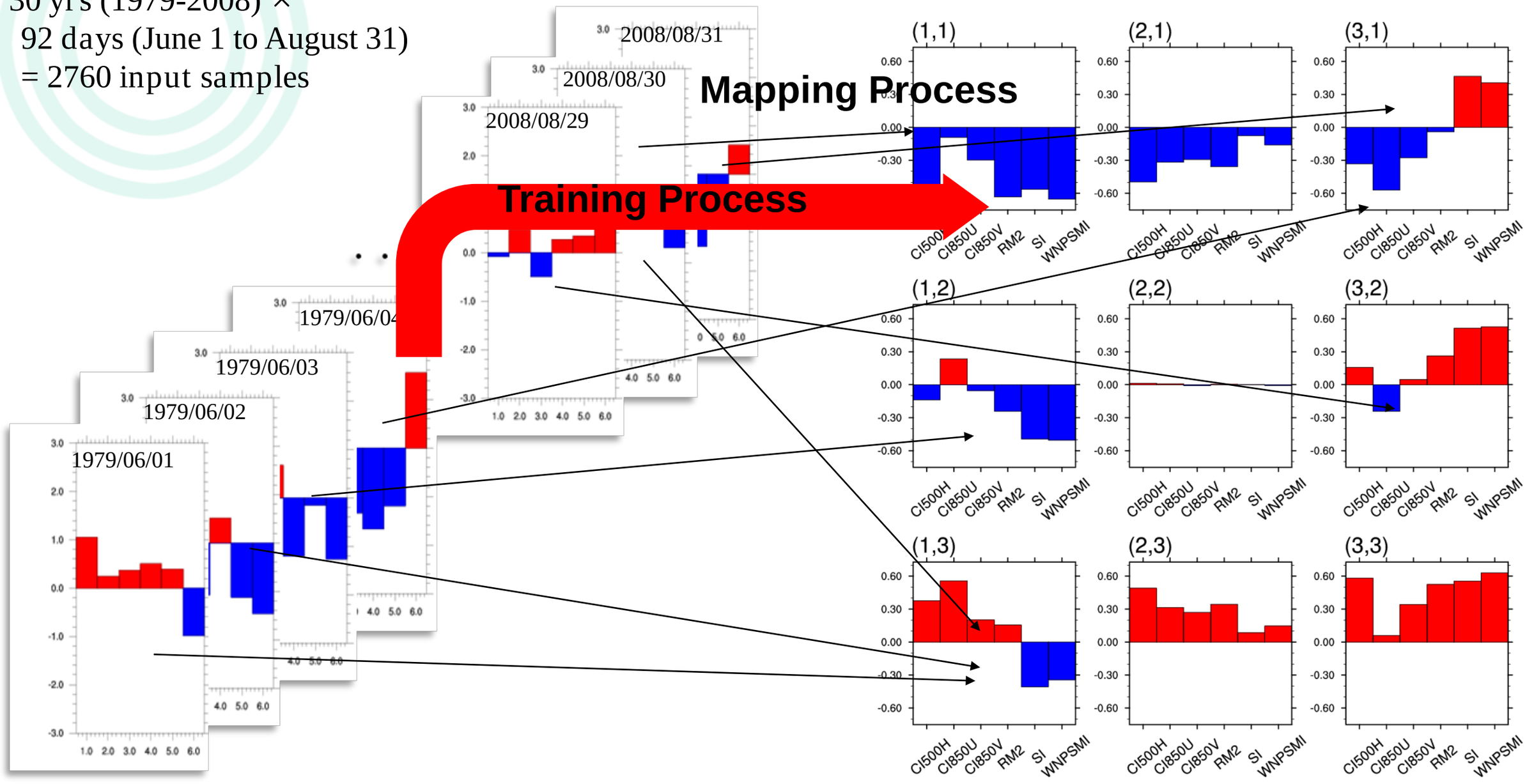
- To identify the nonlinear phases of monsoon ISO using self-organizing map (SOM).
- To provide the dynamical interpretation of precipitation ISO phases by examining large-scale circulation patterns.

< EASM indices >

indices	definition	remarks (ref.)
CI500H	Z500 [25°N～35°N, 135°E～152.5°E]	North Pacific High (Ha et al., 2005)
CI850U	U850 [32.5°N～37.5°N, 127.5°E～147.5°E]	Low-level westerly (Ha et al., 2005)
CI850V	V850 [32.5°N～37.5°N, 127.5°E～147.5°E]	Low-level southerly (Ha et al., 2005)
RM2	U200 [40°N-50°N,110°E-150°E] – U200 [25°N-35°N,110°E-150°E]	Upper level vorticity (Lau et al., 2000)
SI	U850 [5°N-15°N,90°E-130°E] - U200 [5°N-15°N,90°E-130°E]	Vertical shear of zonal wind (Wang, 1998)
WNPMI	U850 [5°N-15°N, 100°E-130°E] - U850 [20°N-30°N, 110°E-140°E]	Western North Pacific monsoon (Wang et al., 2001)

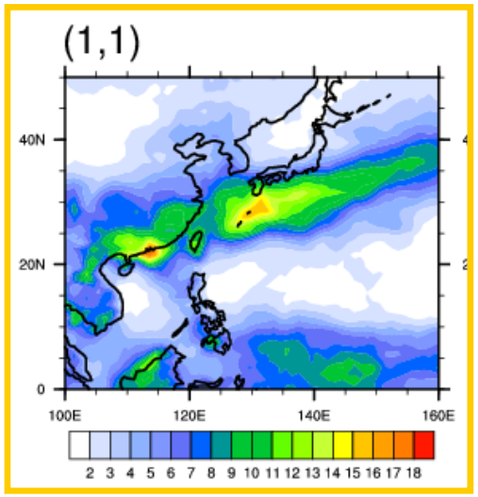
Application of SOM on this study

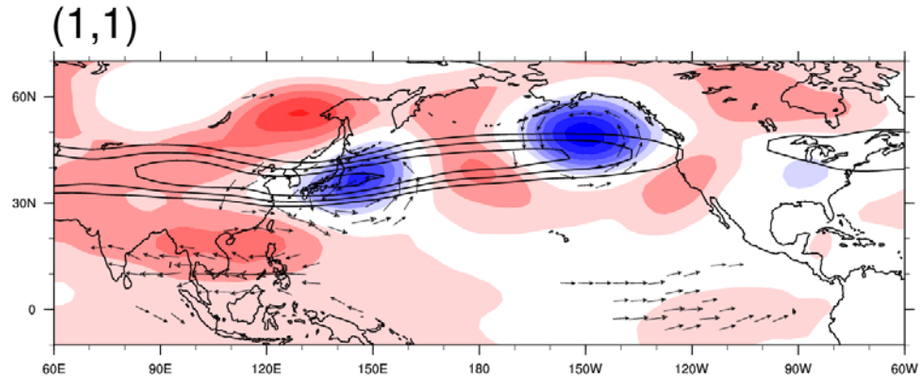
- 6 EASM indices
- 30 yrs (1979-2008) ×
  92 days (June 1 to August 31)
  = 2760 input samples



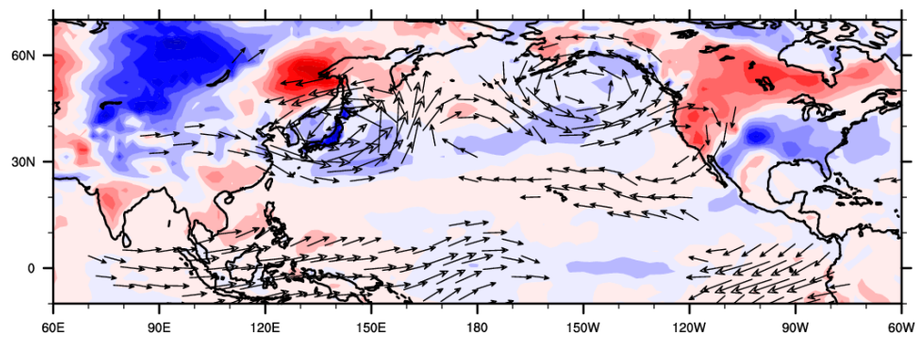
Input data set

2-D Map (3×3)





Contour : U200
Vector : Wind 850 a
Shading : Z500 a

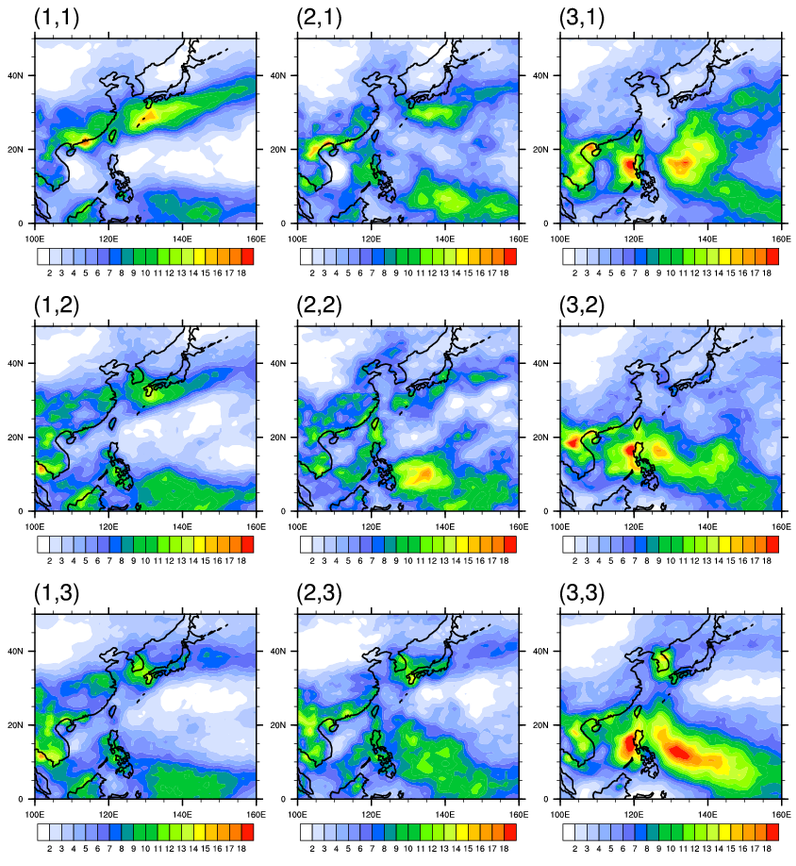


Vector : Wind 200 a
Shading : SKT a

- Zonally elongated (extended) jet stream
- Cyclonic circulation over downstream region of Korea
- Bay of Bengal, SCS warm & Heat induced High → Meridional thermal gradient → enhanced jet stream

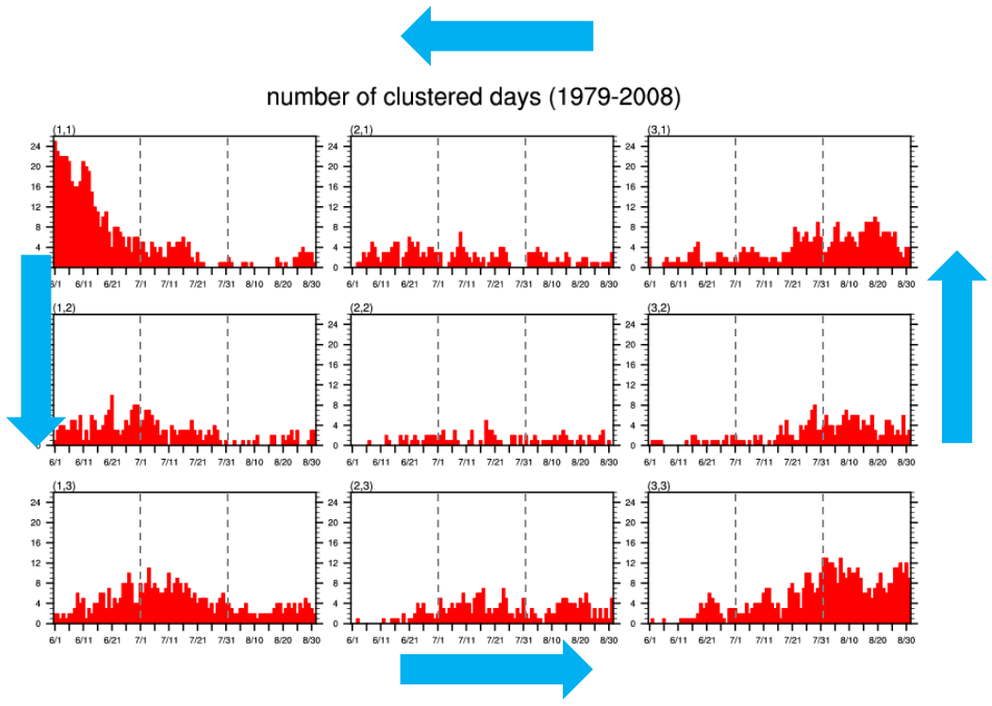
Meiyu-Baiu Mode

Dry-spell Mode

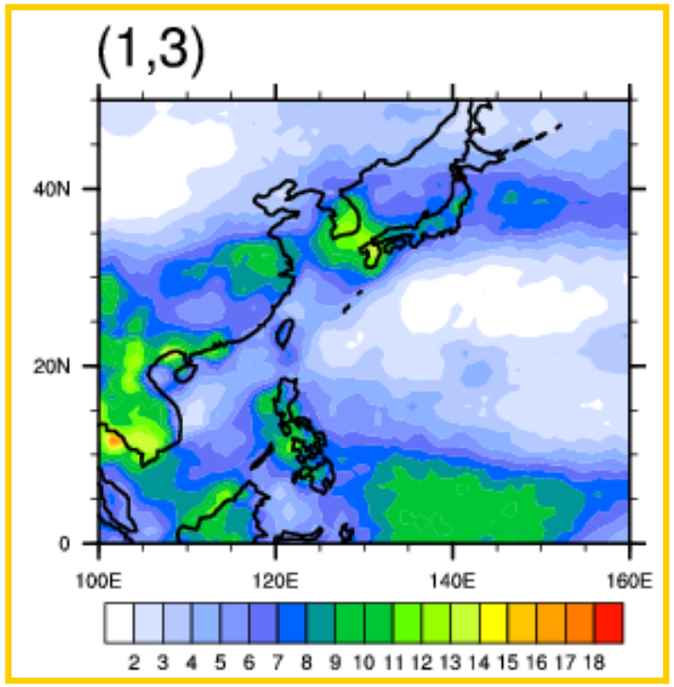


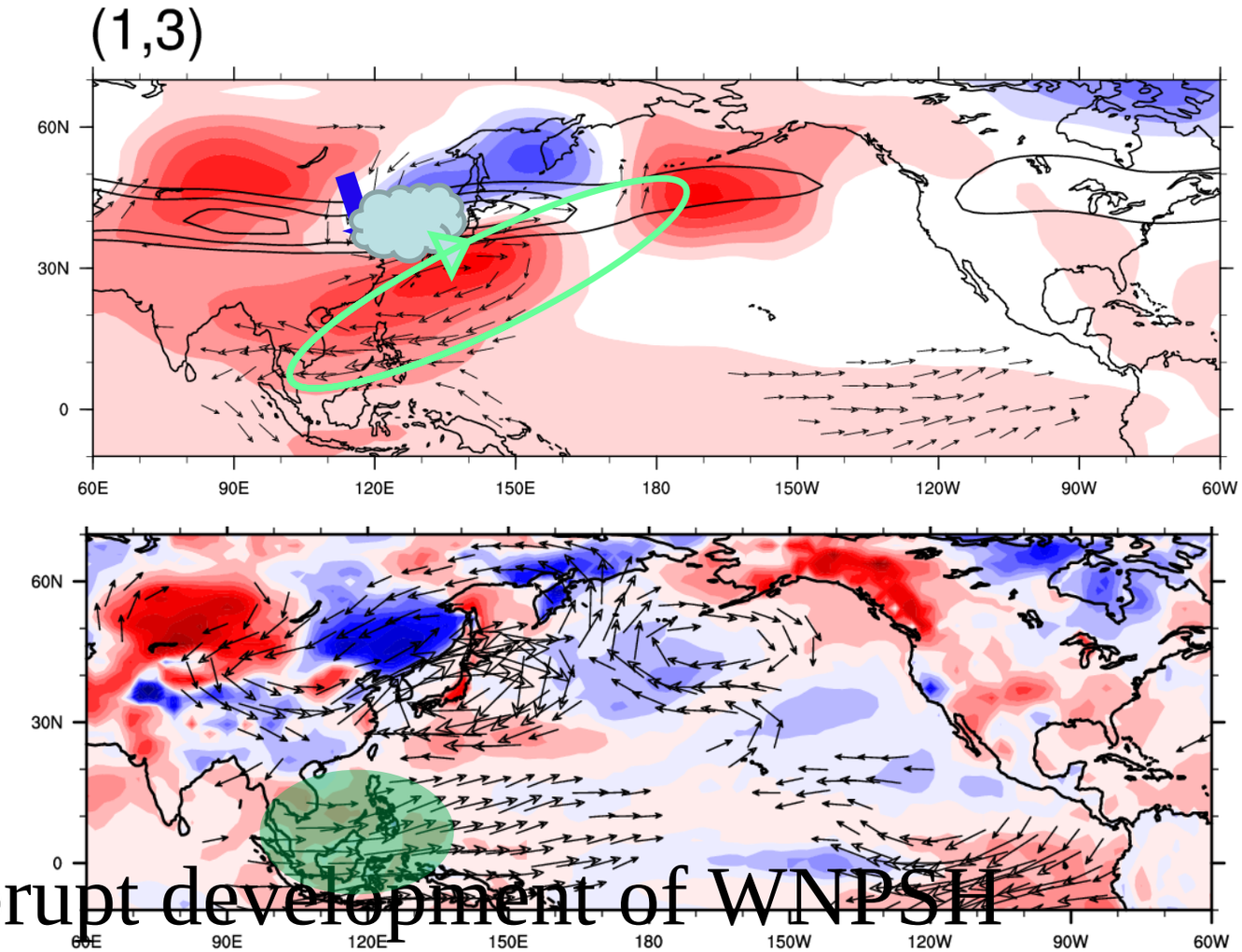
Changma Mode

post-Changma Mode



Large-scale circulation related to Changma mode





Contour : U200
Vector : Wind 850 a
Shading : Z500 a

Vector : U200 a
Shading : SKT a

Abrupt development of WNPSH

- Advection of warm and moist air by the low-level winds is essential for generating convective instability and sustaining the convective activity (Ha et al., 2005)

Software

- http://www.cis.hut.fi/research/som_lvq_pak.shtml - C programs
- http://www.cis.hut.fi/somtoolbox/ - Matlab
- https://github.com/sajinh/SOMPAK_4R - Ruby (maintained by me)

Thank You!


2010/12/26 15:15


2011/04/24 12:49


2010/11/11 10:11


2011/07/24 12:51