Data Mining for Earth Science Data

Vipin Kumar

Army High Performance Computing Research Center
Department of Computer Science
University of Minnesota

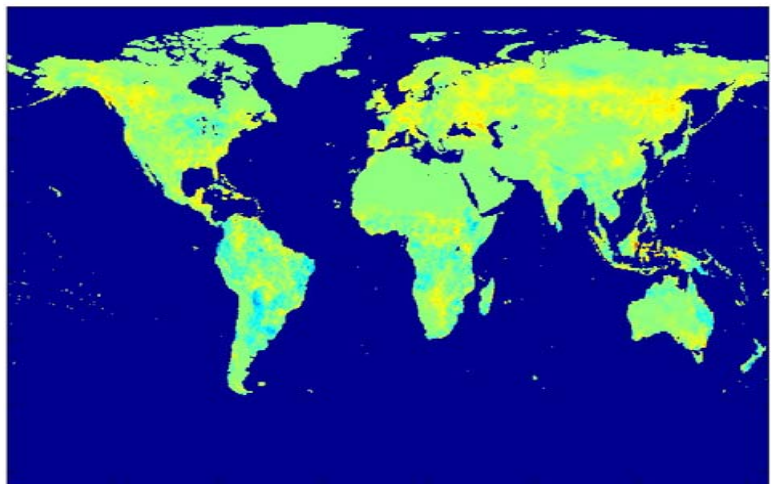
http://www.cs.umn.edu/~kumar

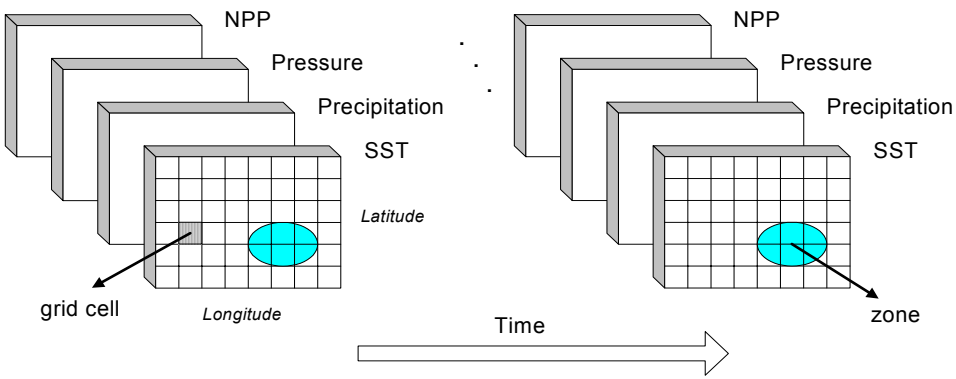
Collaborators:

G. Karypis, S. Shekhar, M. Steinbach, P.N. Tan (AHPCRC),
C. Potter, (NASA Ames Research Center),
S. Klooster (California State University, Monterey Bay).

This work was partially funded by NASA and Army High Performance Computing Center

Research Goals





Research Goals:

- modeling of ecological data
  - ➢ event modeling
  - ➢ zone modeling.
- finding spatio-temporal patterns
  - ➢ associations
  - ➢ predictive models.

A key interest is finding connections between the ocean and the land.

Sources of Earth Science Data

- Before 1950, very sparse, unreliable data.
- Since 1950, reliable global data.
  - Ocean temperature and pressure are based on data from ships.
  - Most land data, (solar, precipitation, temperature and pressure) comes from weather stations.
- Since 1981, data has been available from Earth orbiting satellites.
  - FPAR, a measure related to plant
- Since 1999 TERRA, the flagship of the NASA Earth Observing System, is providing much more detailed data.

Example Pattern: Teleconnections

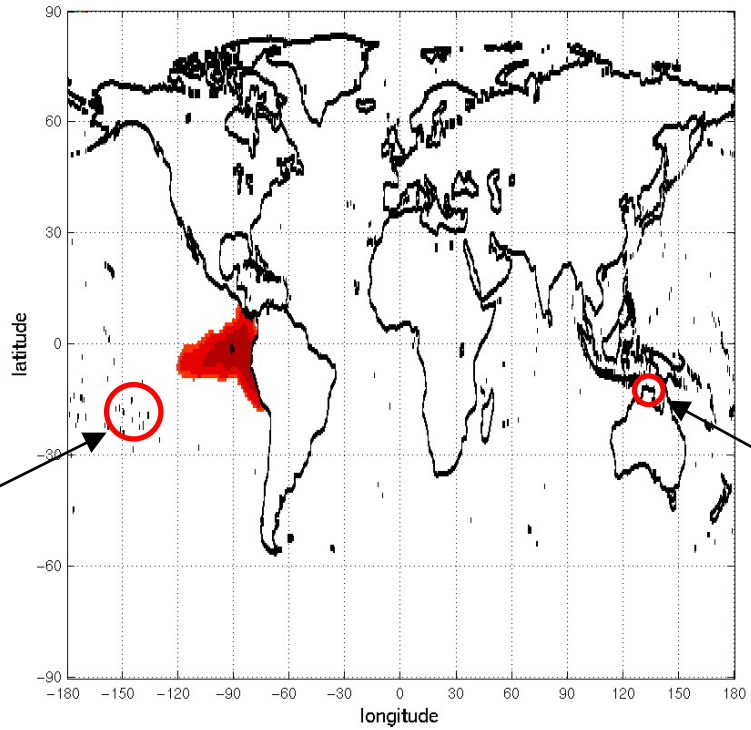
- Teleconnections are the simultaneous variation in climate and related processes over widely separated points on the Earth.

- For example, El Nino is the anomalous warming of the eastern tropical region of the Pacific, and has been linked to various climate phenomena.
  - Droughts in Australia and Southern Africa
  - Heavy rainfall along the western coast of South America
  - Milder winters in the Midwest

Relationship Between SOI and Sea Surface Temperature

SOI measures the pressure difference between Darwin and Tahiti.

The red region at the right is an area of the Pacific that warms when El Nino takes place.

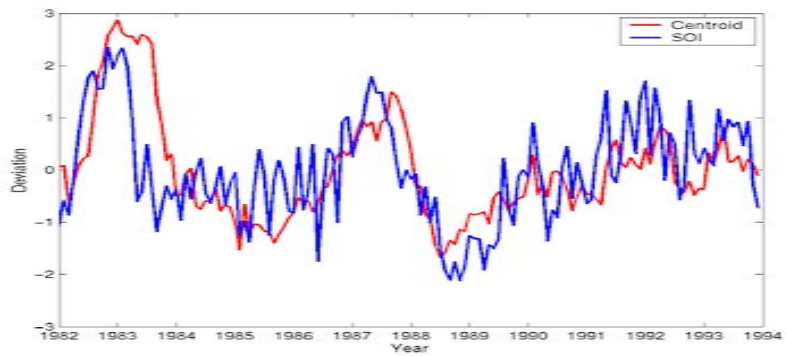


Tahiti

Darwin, Australia

Plot of time series for SOI (blue) and SST centroid of region shown above (red).

Correlation = 0.60



Net Primary Production (NPP)

- Net Primary Production (NPP) is the net assimilation of atmospheric carbon dioxide (CO_2) into organic matter by plants.
  - NPP is driven by solar radiation and can be constrained by precipitation and temperature.
- NPP is a key variable for understanding the global carbon cycle and ecological dynamics of the Earth.
- Keeping track of NPP is important because it includes the food source of humans and all other organisms.
  - Sudden changes in the NPP of a region can have a direct impact on the regional ecology.
- An ecosystem model for predicting NPP, CASA (the Carnegie Ames Stanford Approach) provides a detailed view of terrestrial productivity.

Why Statistics Is Not Sufficient

- Hypothesize-and-test paradigm is extremely labor-intensive.
  - Extremely large and growing families of interesting spatio-temporal hypotheses and patterns in ecological datasets.
- Classical statistics deals primarily with numeric data whereas ecological data contains many categorical attributes.
  - Types of vegetation, ecological events and geographical landmarks.
- Ecological datasets have selection bias in terms of being convenience or opportunity samples.
  - Not traditional statistical idealized random samples from independent, identical distributions.

Benefits of Data Mining

- Data mining provides earth scientist with tools that allow them to spend more time choosing and exploring interesting families of hypotheses.
  - However, statistics is needed to provide methods for determining the “statistical” significance of results.

- By applying the proposed data mining techniques, some of the steps of hypothesis generation and evaluation will be automated, facilitated and improved.

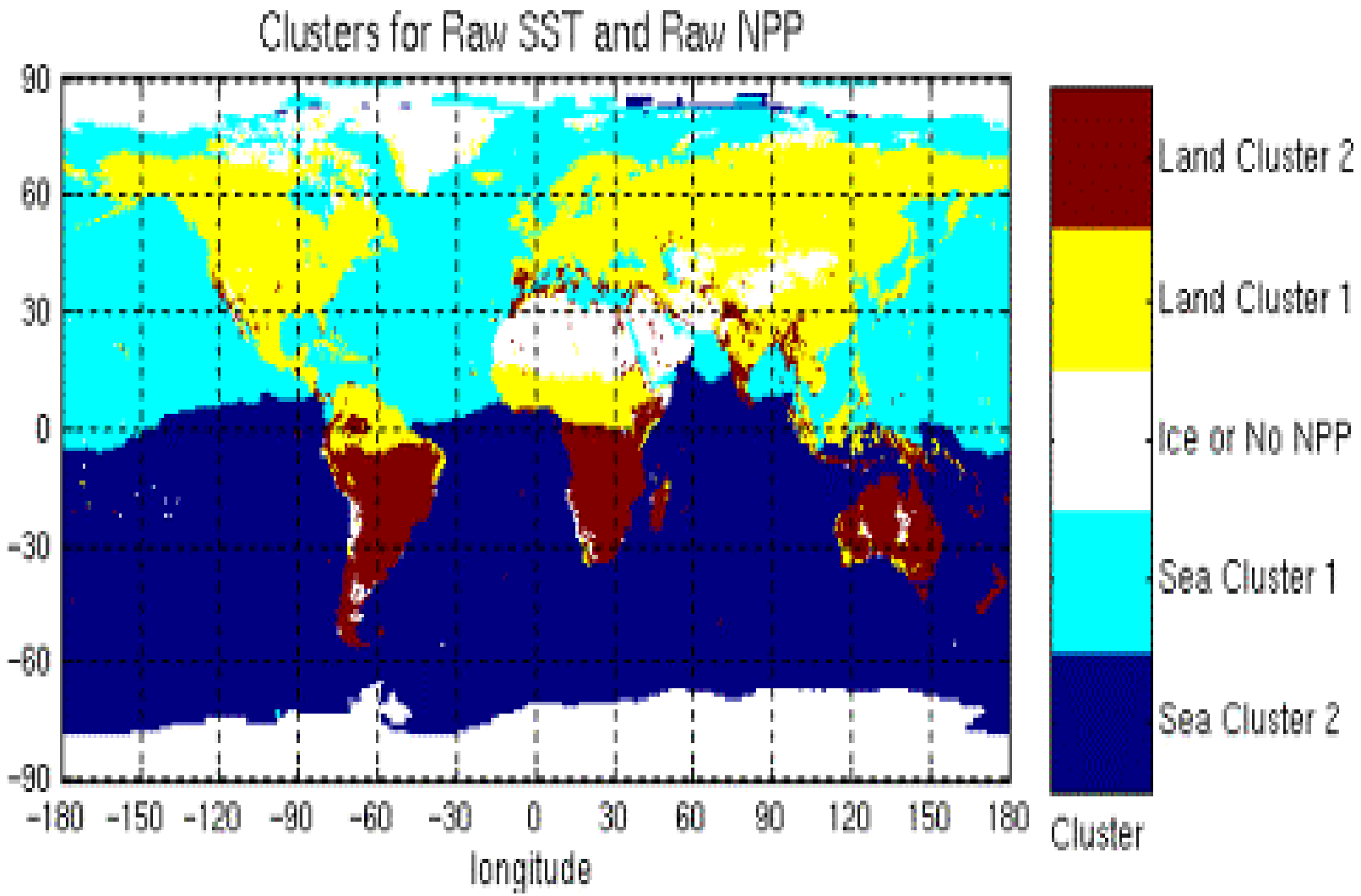
- Association rules provide a “new” framework for detecting relationships between events.

Clustering for Zone Formation

- Interested in relationships between regions, not “points.”
- For land, clustering based on NPP or other variables, e.g., precipitation, temperature.
- For ocean, clustering based on SST (Sea Surface Temperature).
- When “raw” NPP and SST are used, clustering can find seasonal patterns.
  - Anomalous regions have plant growth patterns which reversed from those typically observed in the hemisphere in which they reside, and are easy to spot.

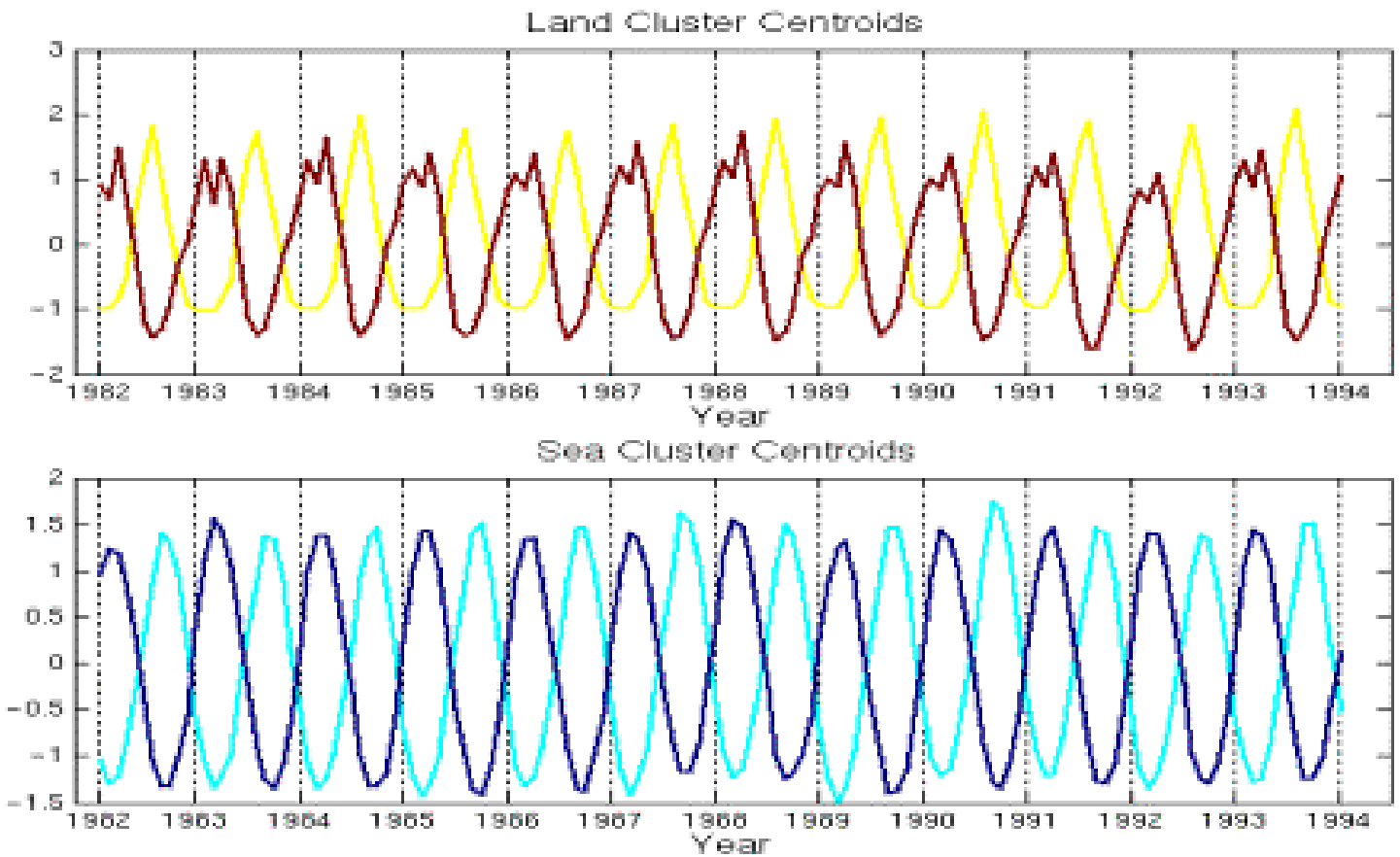
K-Means Clustering of Raw NPP and Raw SST

(Num clusters = 2)



K-Means Clustering of Raw NPP and Raw SST

(Num clusters = 2)



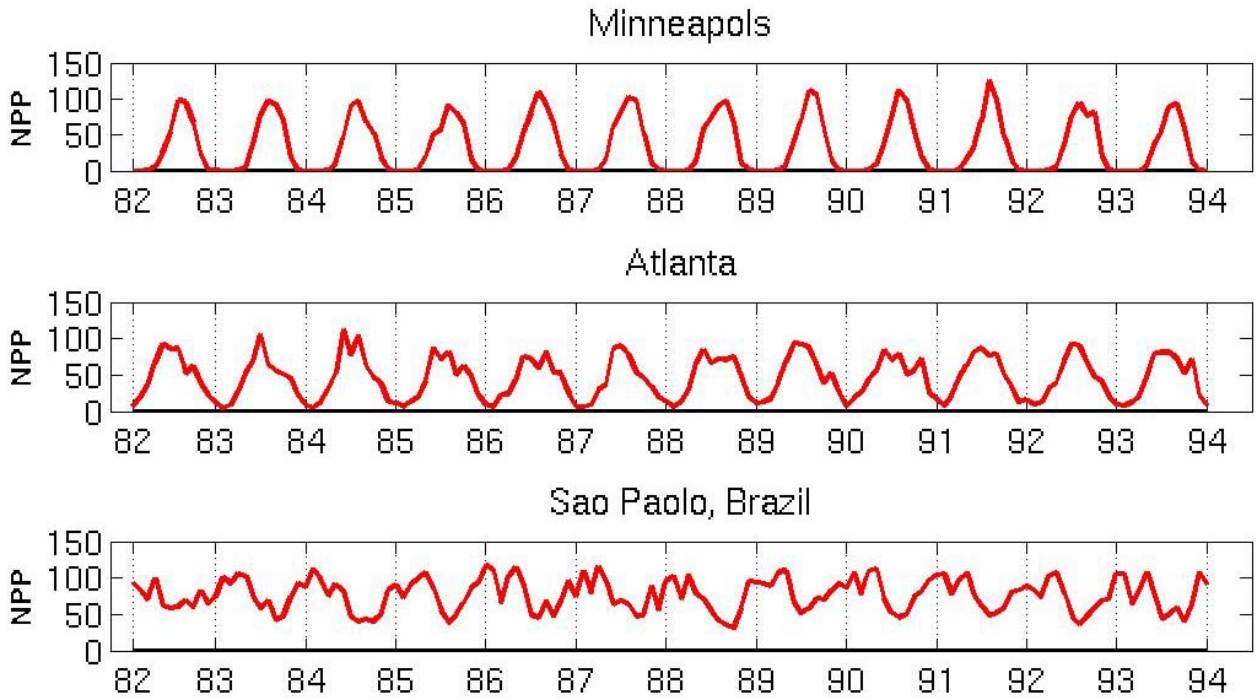
Land Cluster Cohesion: North = 0.78, South = 0.59

Ocean Cluster Cohesion: North = 0.77, South = 0.80

Preprocessing: Removing Seasonality

- Must remove seasonality to see events (anomalies) of interest.
  - 12 month moving average
    - Smoothes as well as removes seasonality
  - Discrete Fourier Transform
  - Monthly Z Score
    - Subtract of monthly mean and divide by monthly standard deviation
  - Singular Value Decomposition

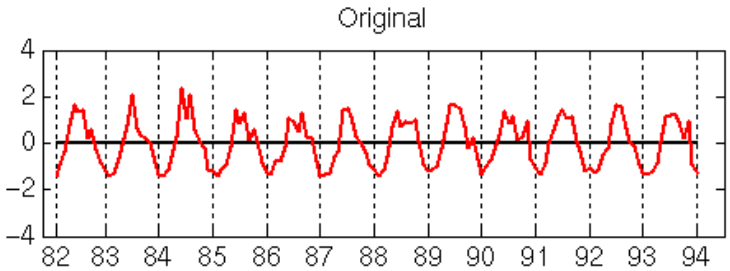
Sample NPP Time Series

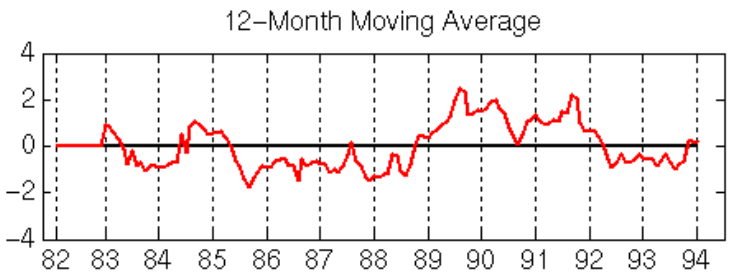


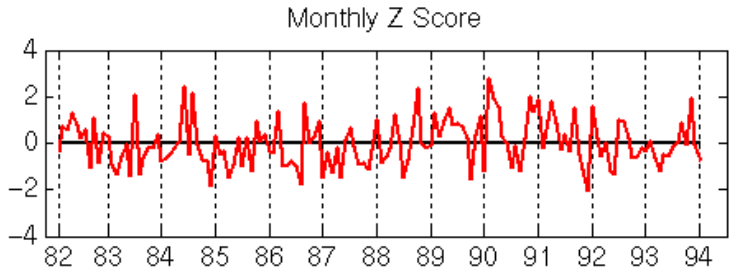
Correlations between time series

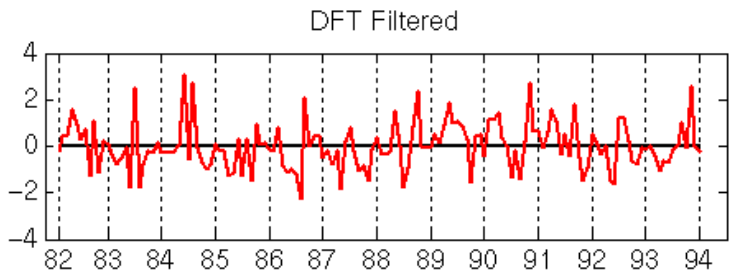
	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

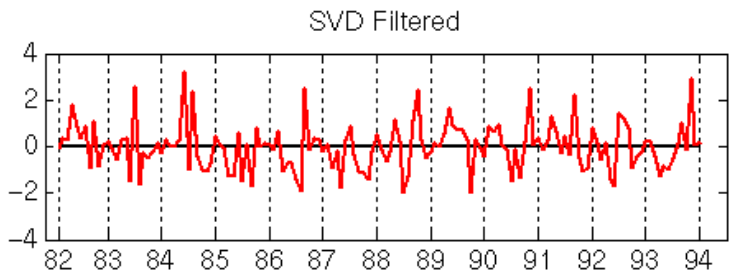
Removing Seasonality from Atlanta Time Series



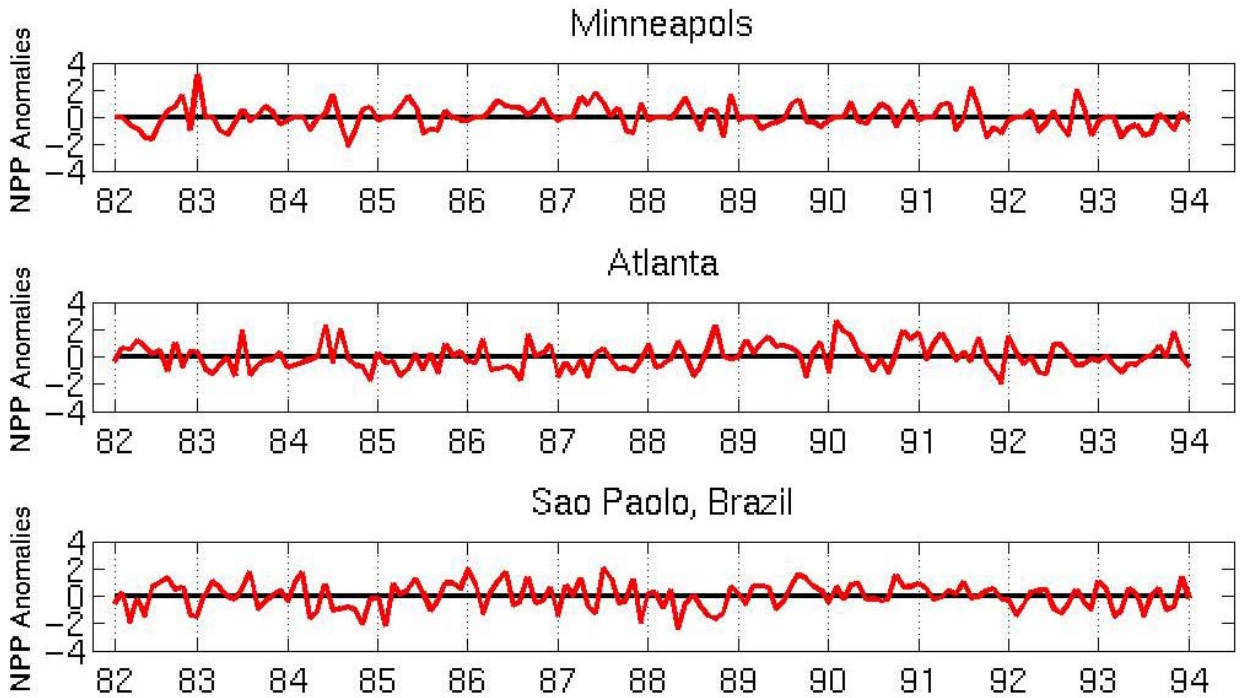








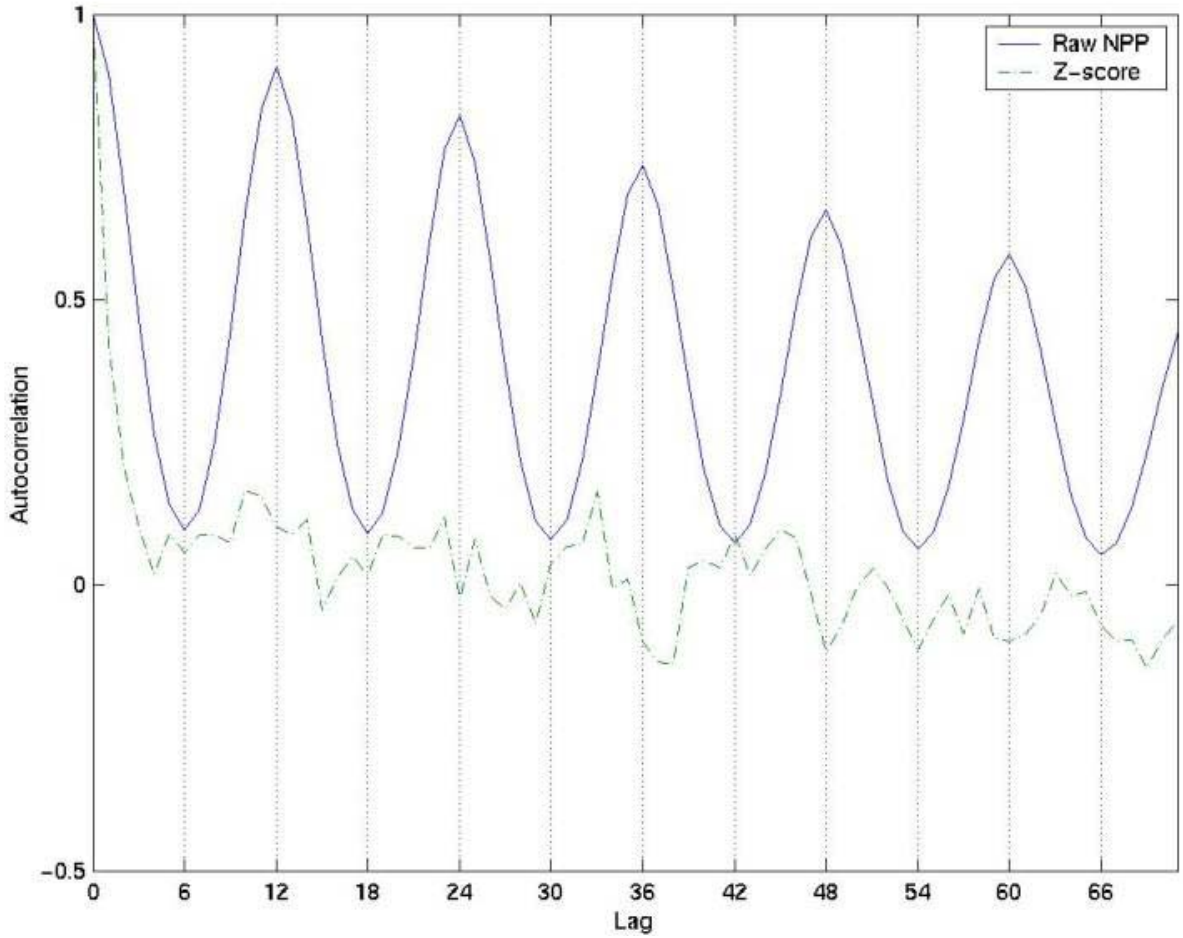
Seasonality Accounts for Much Correlation



Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000

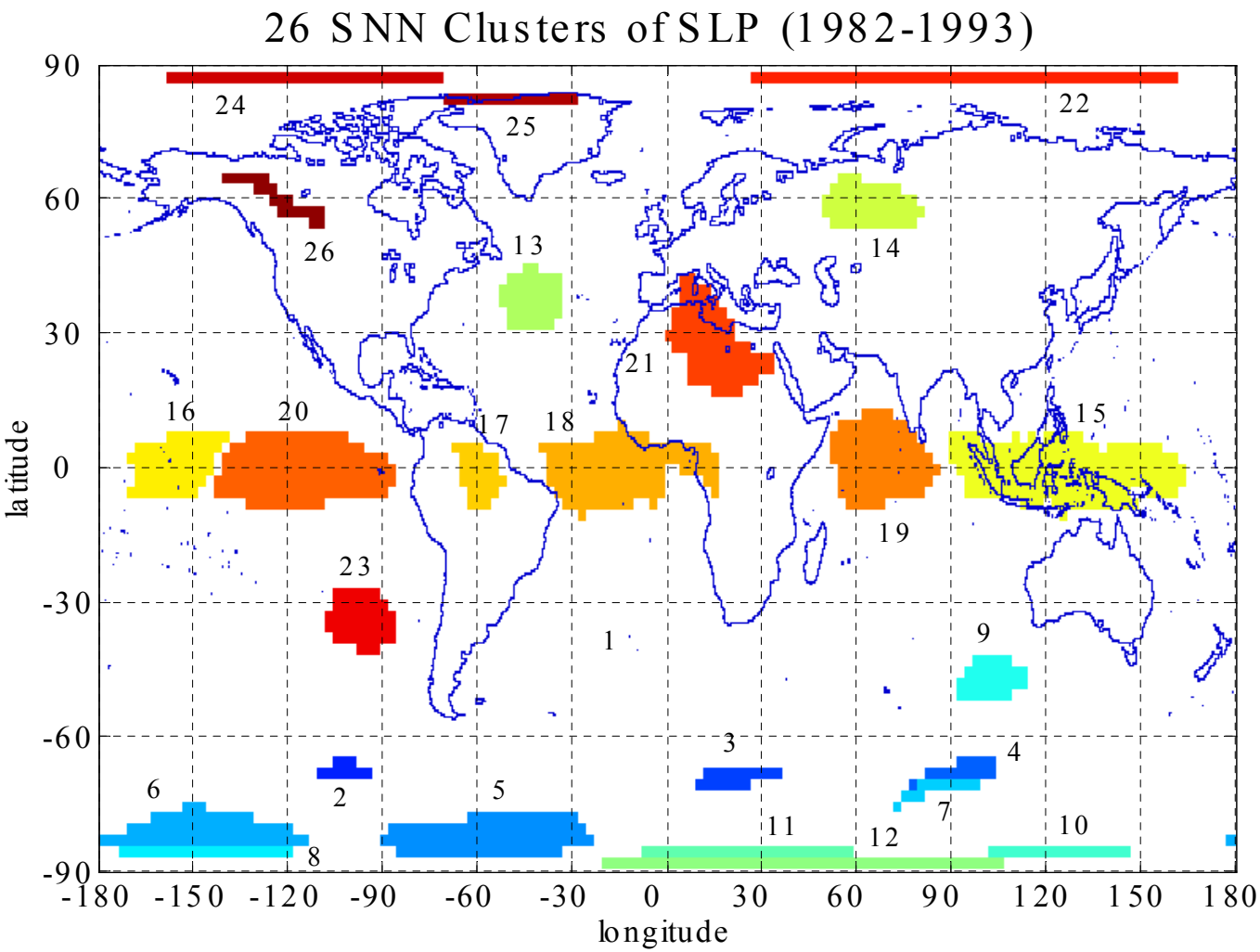
Removing Seasonality Removes Much of the Autocorrelation



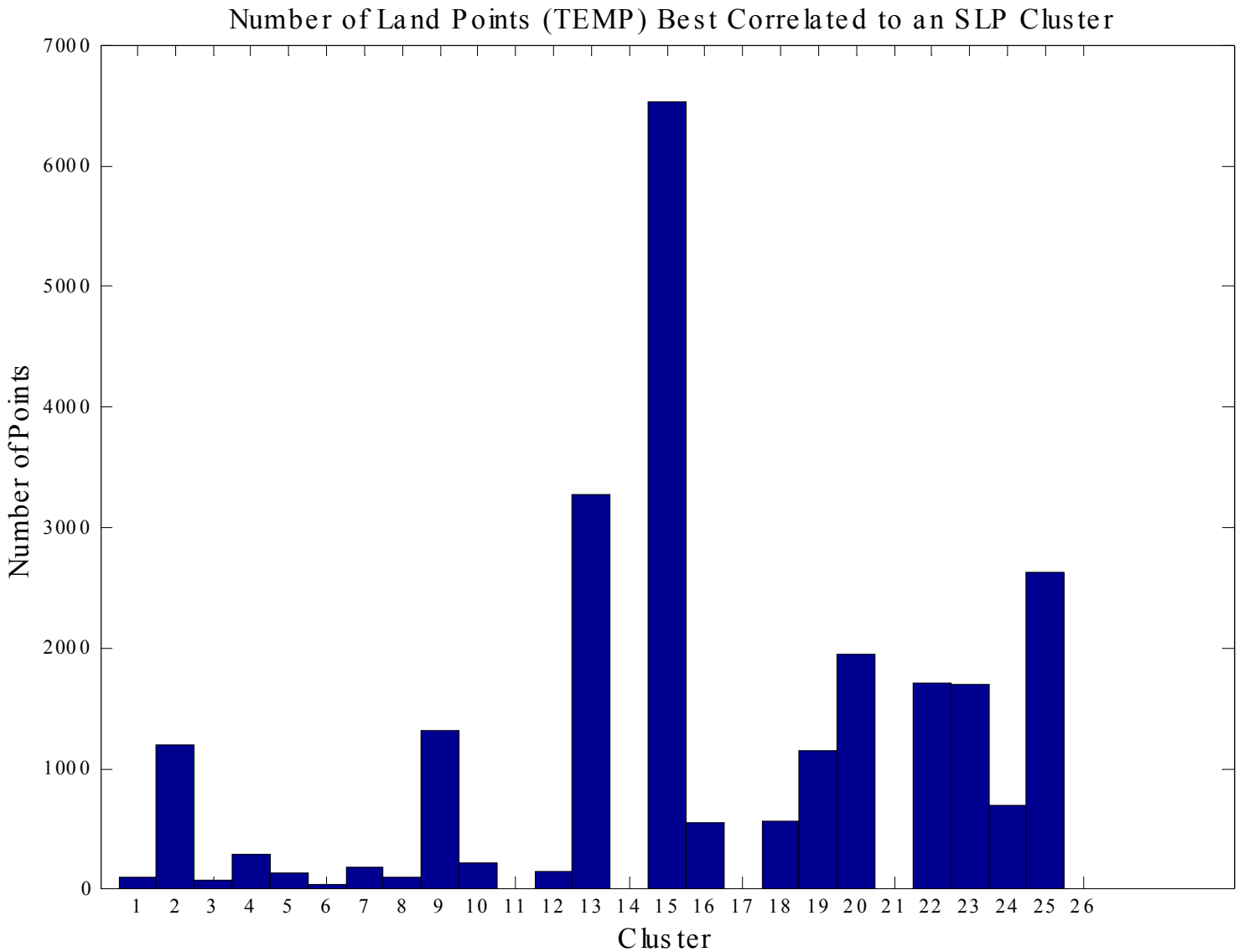
Discovery of Ocean Climate Indices

- **Use clustering to find areas of the oceans that have relatively homogeneous behavior.**
  - Cluster centroids are potential OCIs.
- **Evaluate the influence of potential OCIs on land points.**
- **Determine if the potential OCI matches a known OCI.**
- **For potential OCIs that are not well-known, conduct further analysis.**

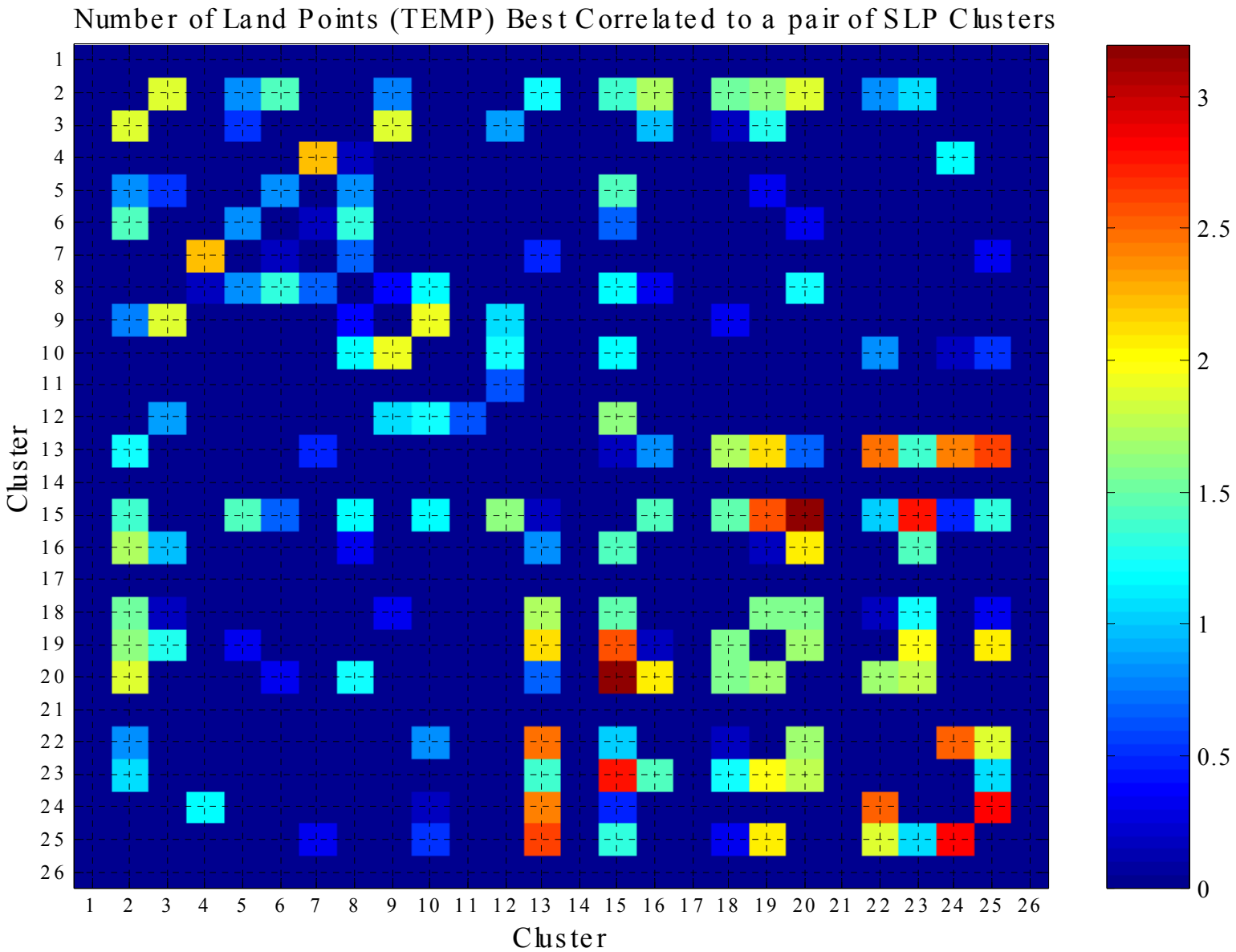
SLP Clusters



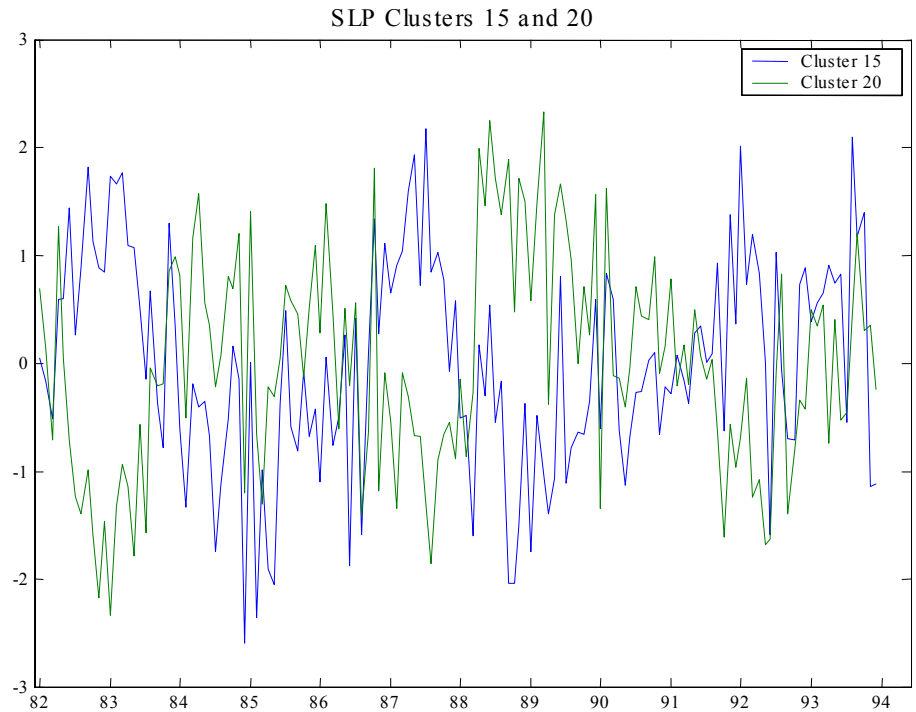
Number of Land Points Best Correlated to SLP Clusters



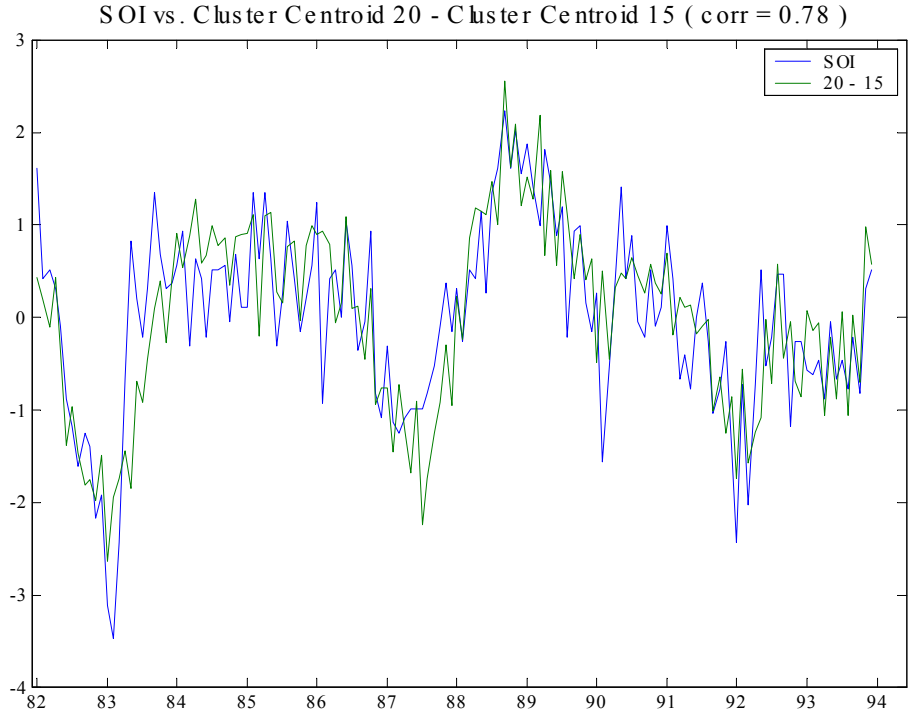
Number of Land Points Best Correlated to Pairs of SLP Clusters



Pairs of SLP Clusters that Correspond to SOI

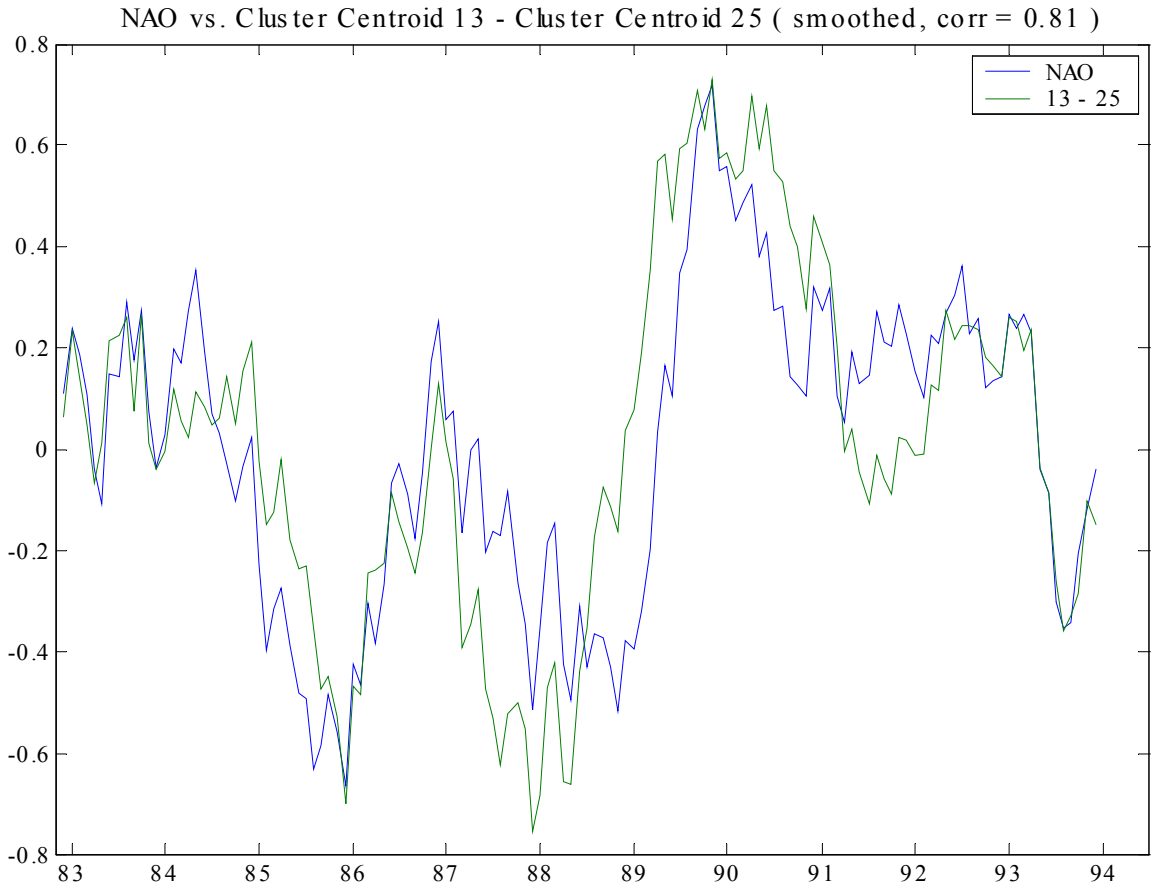


Centroids of SLP clusters 15 and 20 (near Darwin, Australia and Tahiti) 1982-1993.



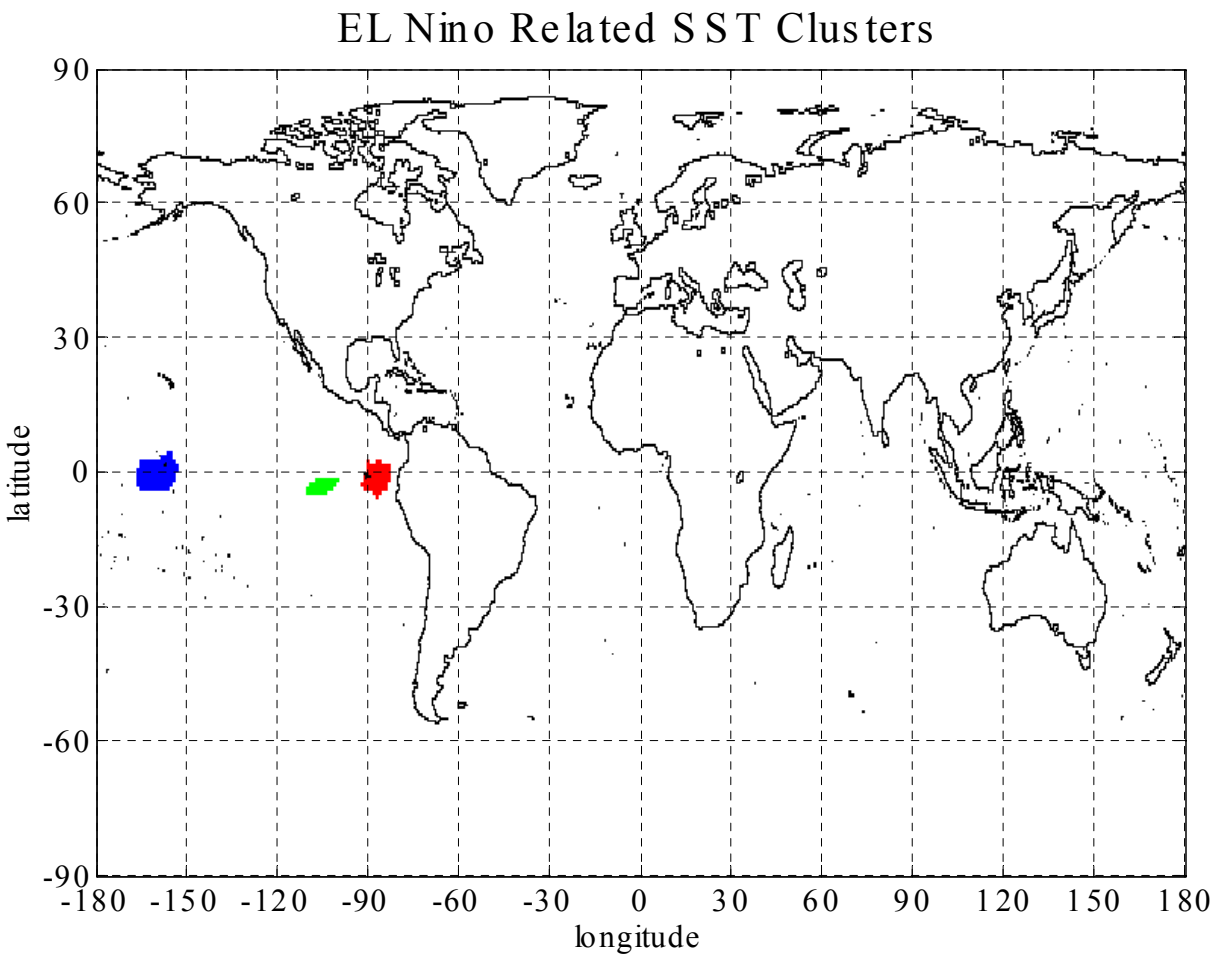
Centroid of cluster 20 – Centroid of cluster 15 versus SOI

Pairs of SLP Clusters that Correspond to NAO



Smoothed difference of SLP cluster centroids 13 and 25 versus North Atlantic Oscillation Index. (1982-1993)

SST Clusters that Correspond to El Nino Climate Indices

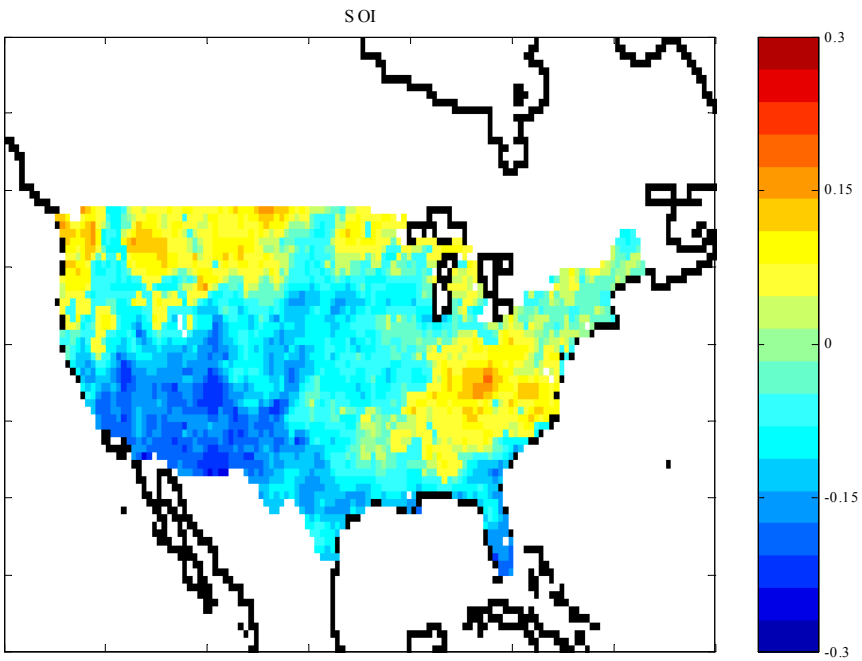


SNN clusters of SST that are highly correlated with El Nino indices.

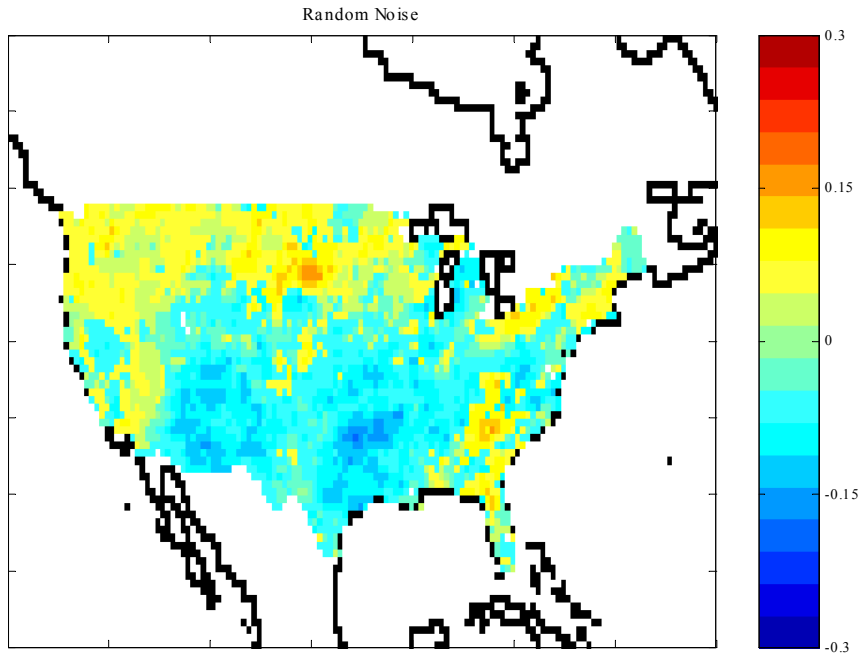
Niño Region	Range Longitude	Range Latitude
1+2	90°W-80°W	10°S-0°
3	150°W-90°W	5°S-5°N
3.4	170°W-120°W	5°S-5°N
4	160°E-150°W	5°S-5°N

El Nino Regions

Maps of Maximum Correlation (shifts 0-6 months)

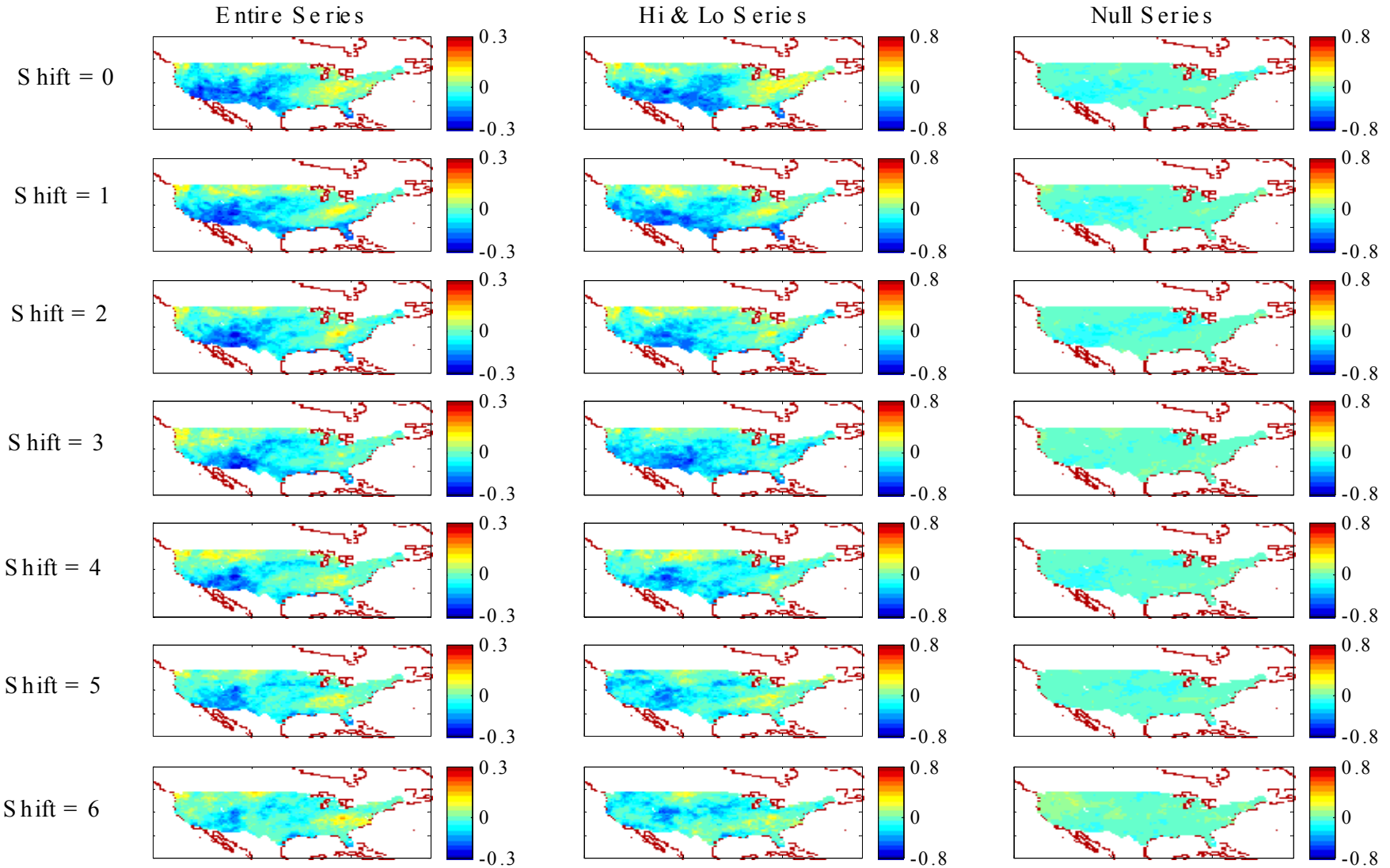


SOI

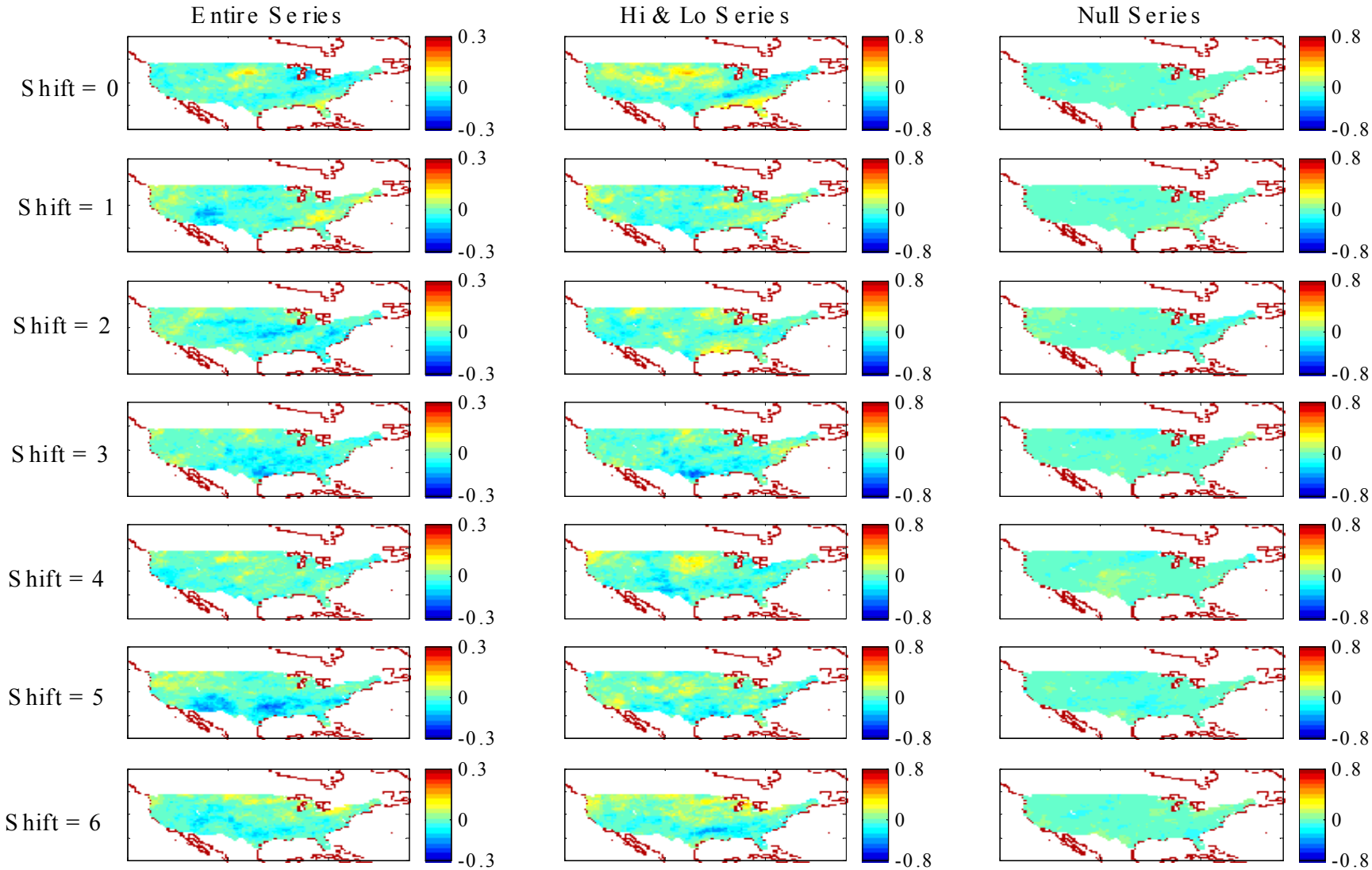


Random Noise

Ocean Climate Indices (SOI) have Persistent Correlation Patterns

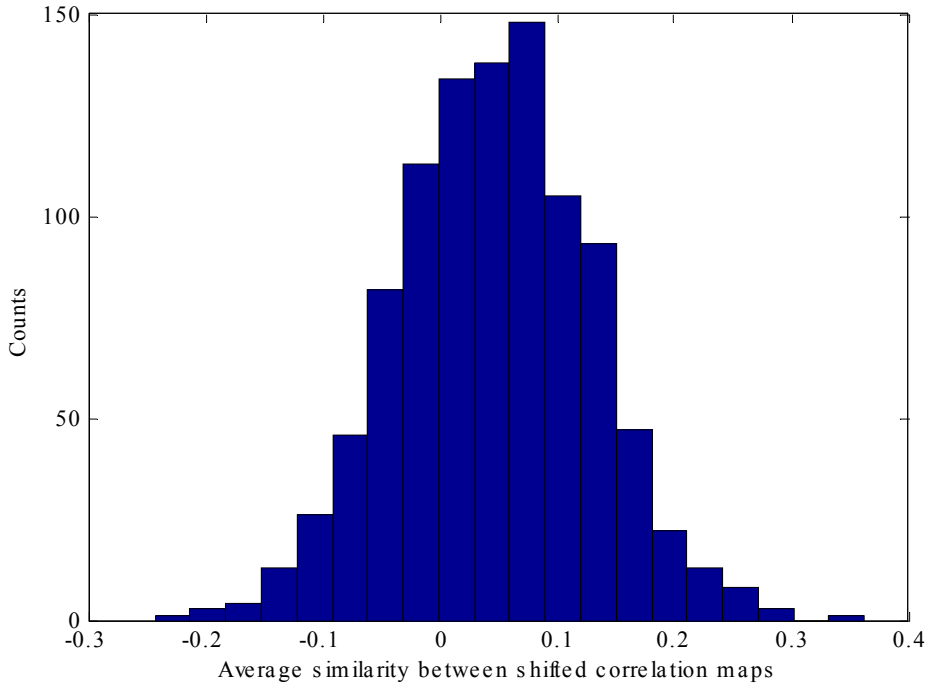


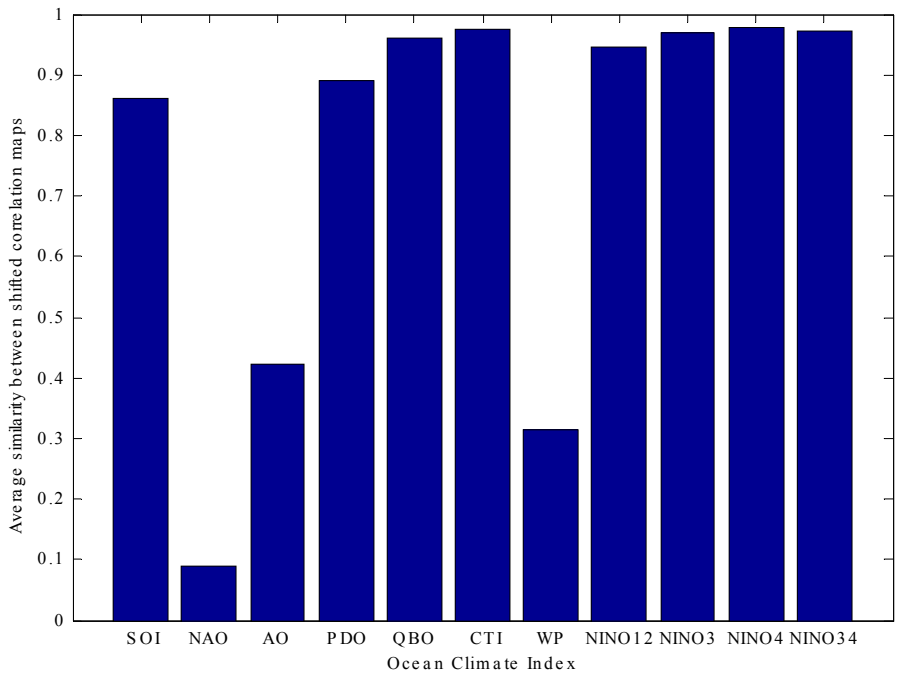
“Noise” Time Series do not have Persistent Correlation Patterns



Testing for Persistence via Average Similarity of Correlation Maps

- **Correlation Maps using Precipitation for the United States.**





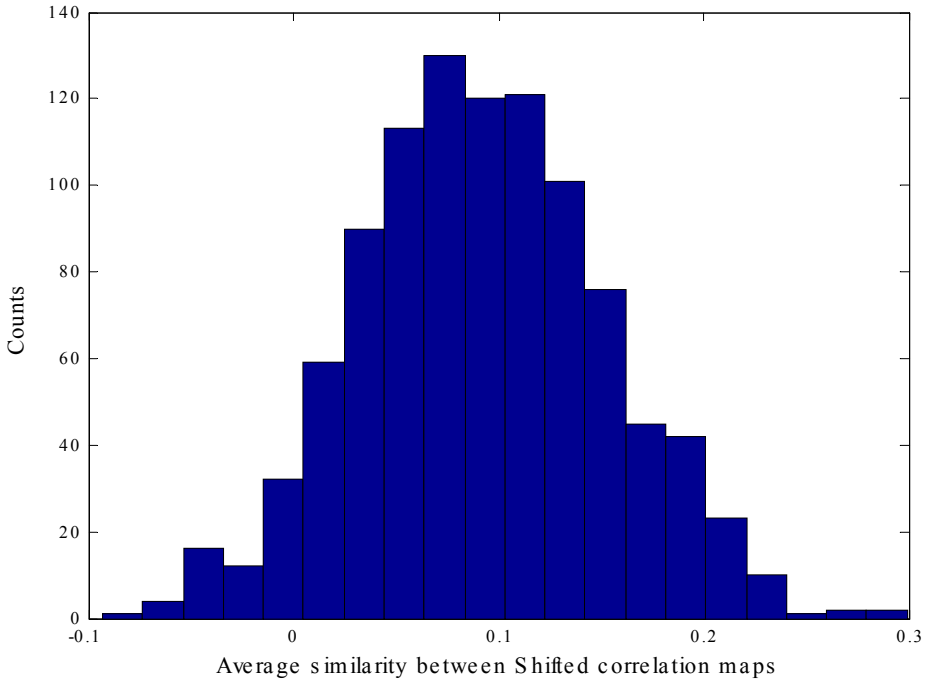
- Histogram of average similarity of shifted correlation maps for 1000 randomly generated time series.
- Average similarity for noise times series almost always between –0.2 and 0.3

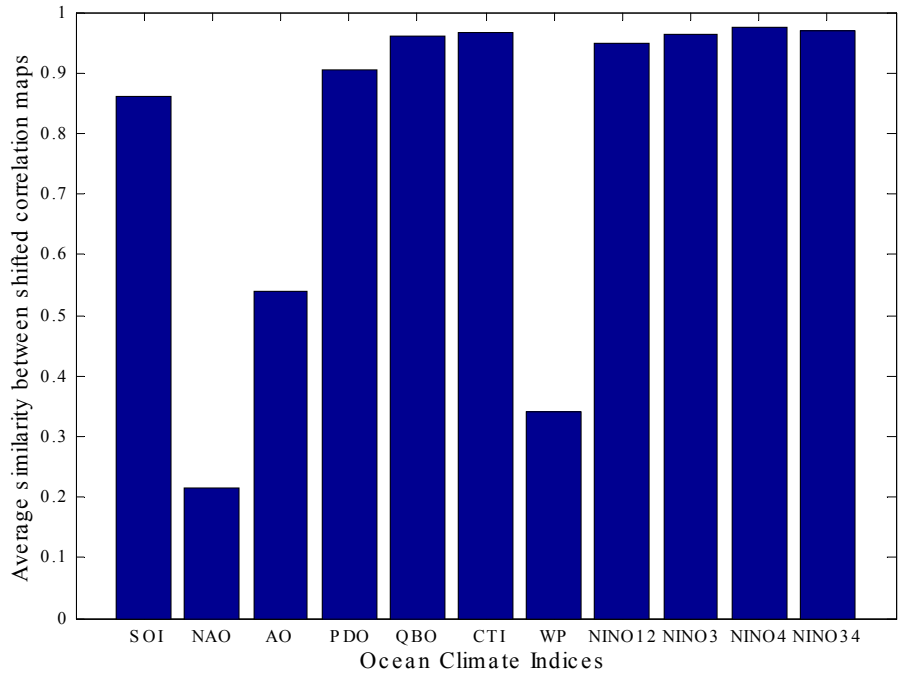
- Average similarity of shifted correlation maps for various OCIs.

$$Similarity = \frac{1}{p}\sum_{i=0}^{p-1} corr(M_i, M_{i+1})$$

Testing for Persistence via Average Similarity of Correlation Maps

- **Correlation Maps using Precipitation for the entire globe.**



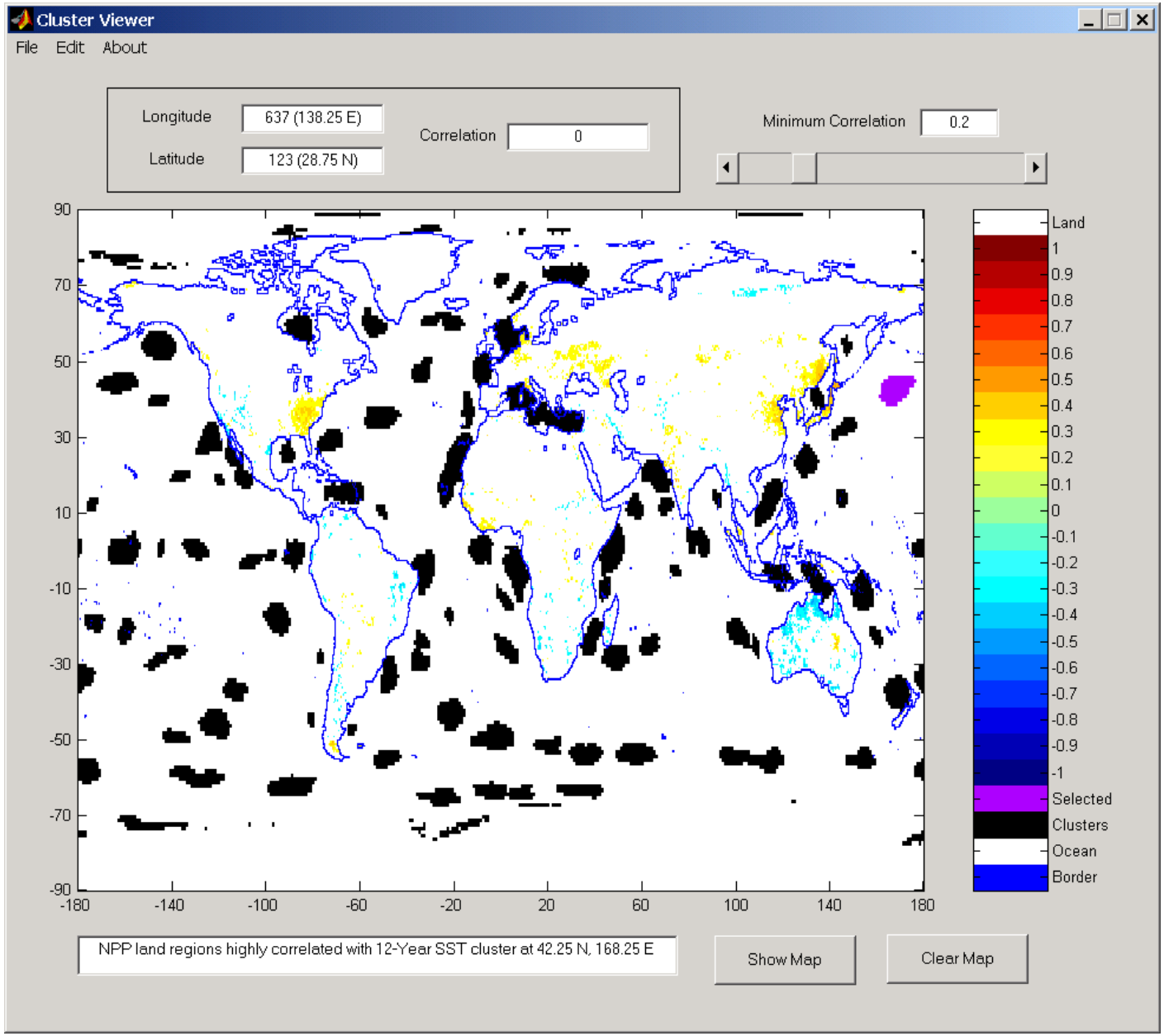


- Histogram of average similarity of shifted correlation maps for 1000 randomly generated time series.

- Average similarity of shifted correlation maps for various OCIs.

Cluster Viewer

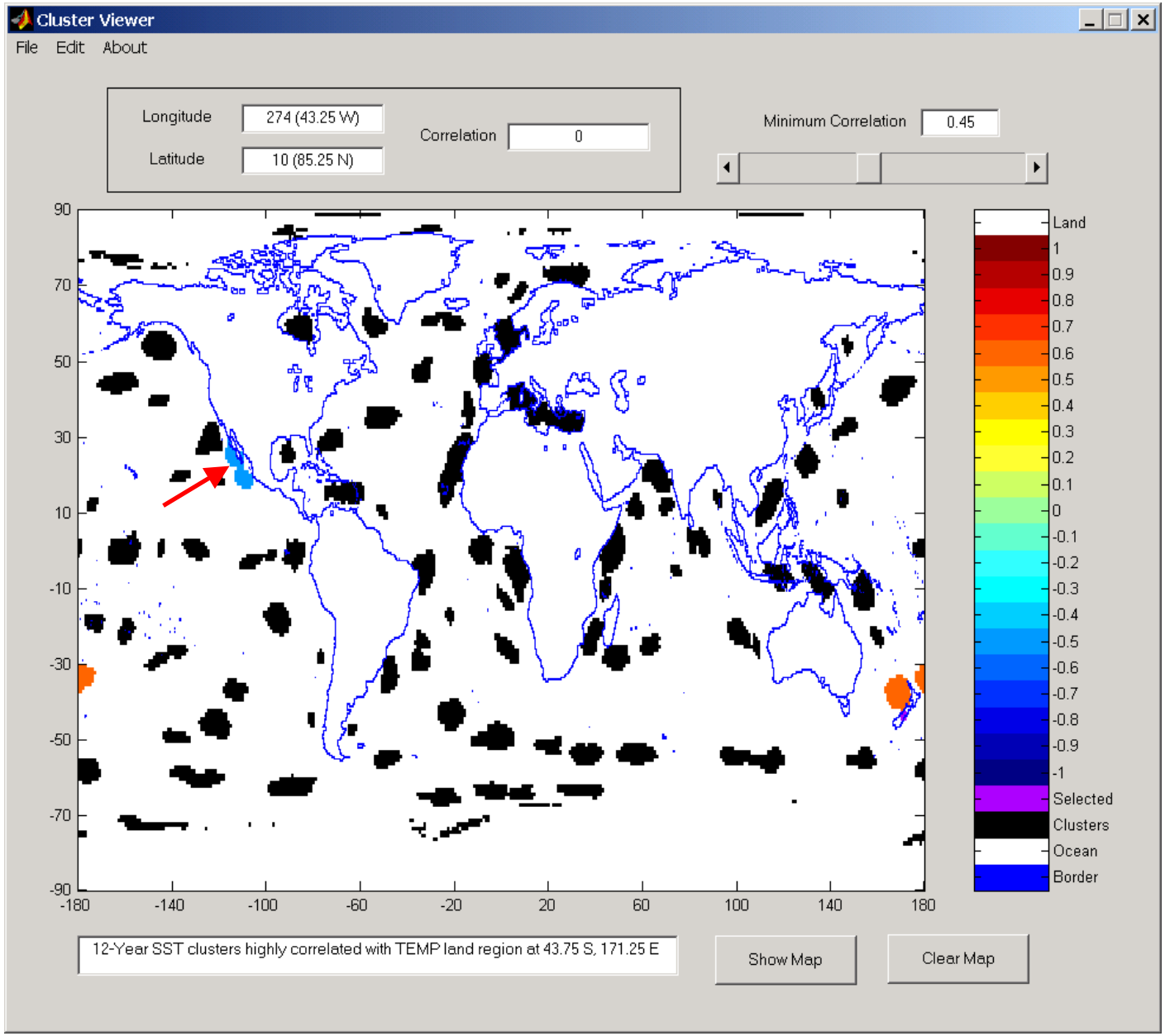
Cluster viewer showing land regions with positive or negative correlation > 0.2 with highlighted ocean cluster.



Cluster Viewer

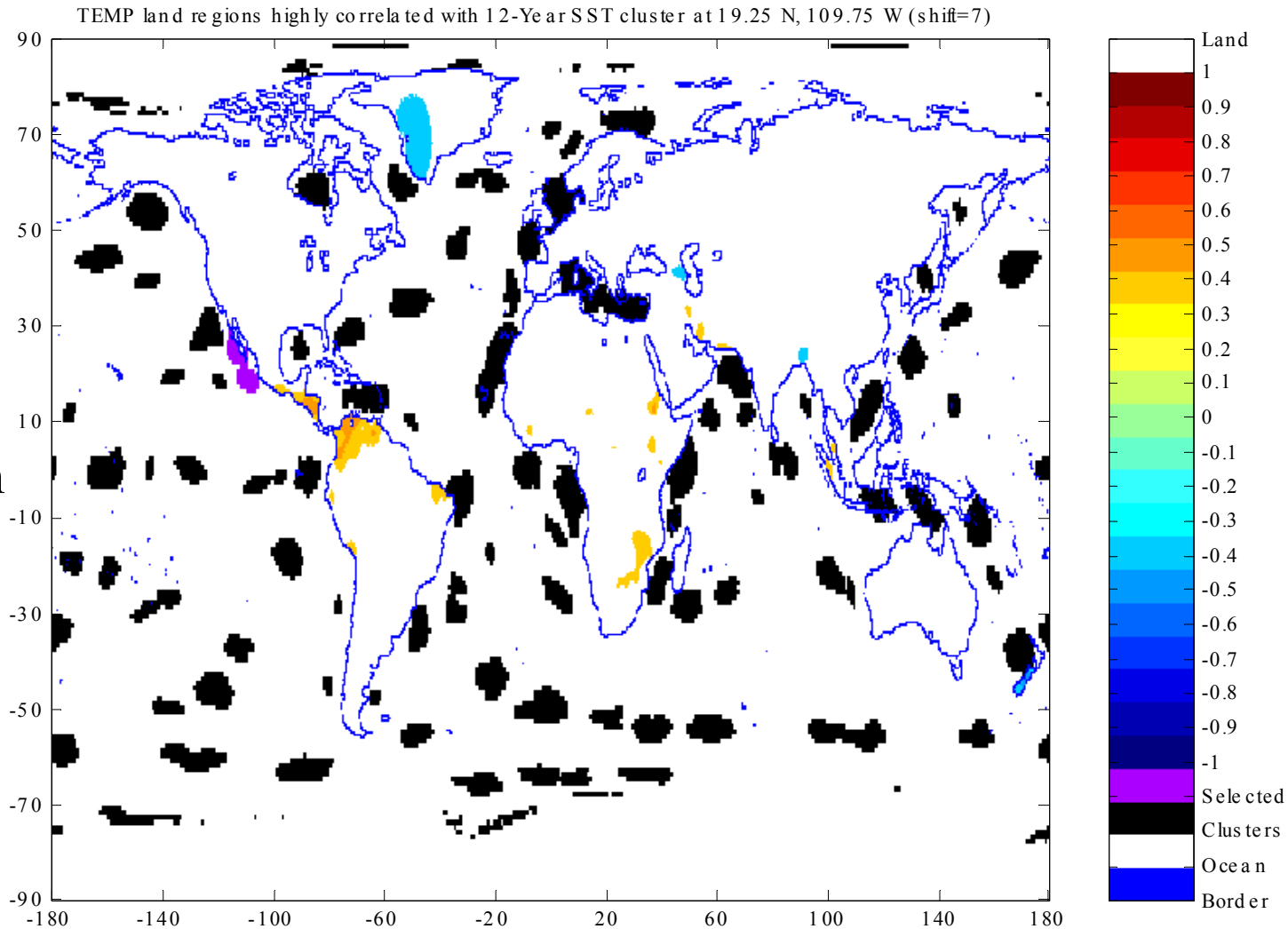
Cluster viewer showing clusters correlated (> 0.45) to a New Zealand land point)

Notice cluster off the coast of western Mexico, which is negatively correlated.



Cluster Viewer

Cluster viewer showing land points (Temp) correlated (> 0.34) to a cluster off the coast of western Mexico.



Statistical Issues

- Temporal Autocorrelation
  - Makes it difficult to calculate degrees of freedom and determine significance levels for tests, e.g., non-zero correlation.
  - Moving average is nice for smoothing and seeing the overall behavior, but introduces additional autocorrelation.
  - Removal of seasonality removes much of the autocorrelation (as long as not performed via the moving average).
- Measures of time series similarity
  - Detecting non-linear connections
  - Detecting connections that only exist at certain times.
    - Sometimes only extreme events have an effect.
  - Automatically detecting appropriate time lags.
  - Statistical tests for more sophisticated measures.

Statistical Issues …

- Detecting spurious connections.
  - We are performing many correlation calculations and there is a chance of spurious correlations.
    - Given that we have ~100,000 locations on the Earth for which we have time series, how many spuriously high correlations will we get when we calculate the correlation between these locations and a climate index?
  - Because of spatial autocorrelation, these correlations are not independent.
    - Again we have trouble calculating the degrees of freedom.

Mining Associations from Earth Science Data

- Earth Science data:
  - Data is continuous rather than discrete.
  - Data has spatial and temporal components.
  - Data can be multilevel
    - time and spatial granularities.
  - Observations are not i.i.d. due to spatial and temporal autocorrelations.
  - Data may contain noise, missing information and erroneous information
    - e.g., historical SST data between 1856-1941 is measured using wooden buckets.

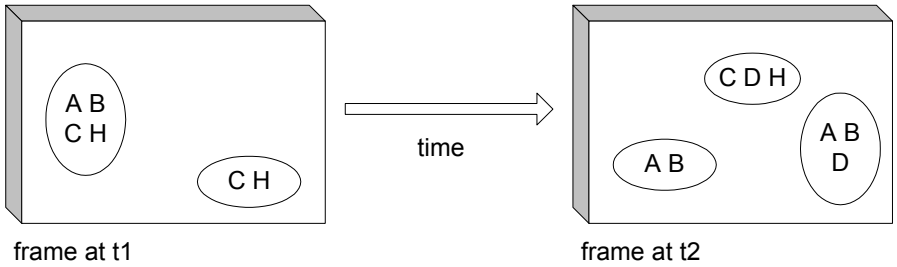
Issues in Mining Associations from Earth Science data

- How to define transactions?
  - What are the baskets?
  - What are the items?
- What are the patterns of interest?
  - Patterns due to anomalous events such as El-Nino and global warming.
  - Patterns that show teleconnections between land and ocean variables.
- How to modify existing association pattern discovery algorithms to accommodate spatio-temporal patterns.
- How to incorporate domain knowledge to filter out uninteresting patterns.

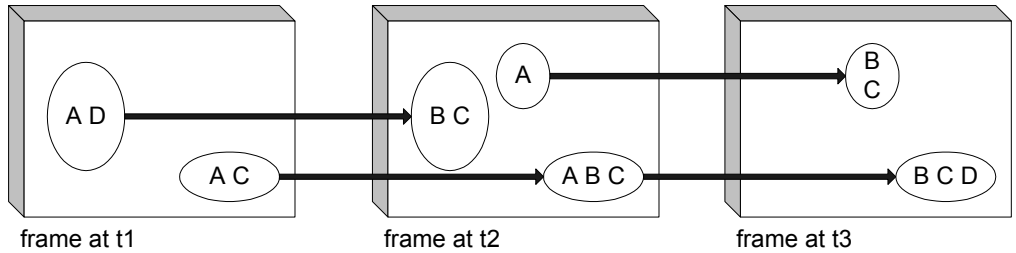
Types of Spatio-Temporal Association Patterns

Type of Pattern	Description
Intra-zone non-sequential	relationships among events in the same grid cell or zone, ignoring the temporal aspects of the data.
Intra-zone sequential	relationships among events happening in different grid cells or zones, ignoring temporal aspects of the data.
Inter-zone non-sequential	temporal relationships among events occurring within the same grid cell or zone.
Inter-zone sequential	temporal relationships among events occurring at different spatial locations.

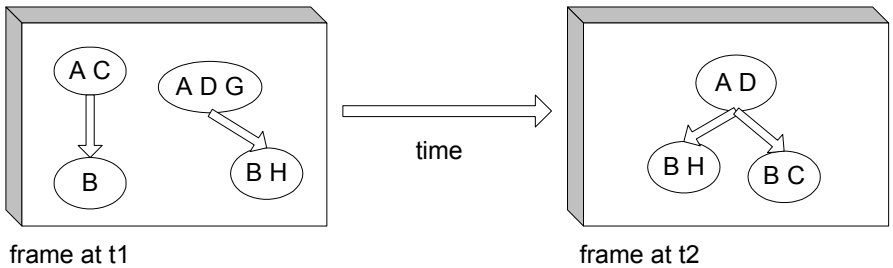
Types of Spatio-temporal Patterns



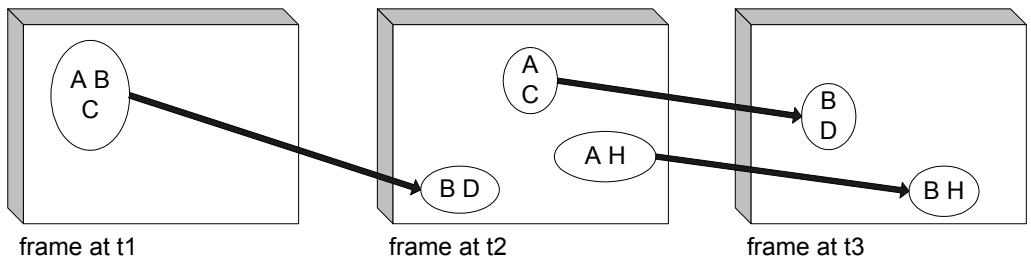
(a) Intra-zone non-sequential (e.g. {A,B}, {C,H})



(b) Intra-zone sequential (e.g. A ==> C)



(c) Inter-zone non-sequential (e.g. B to the south of A)

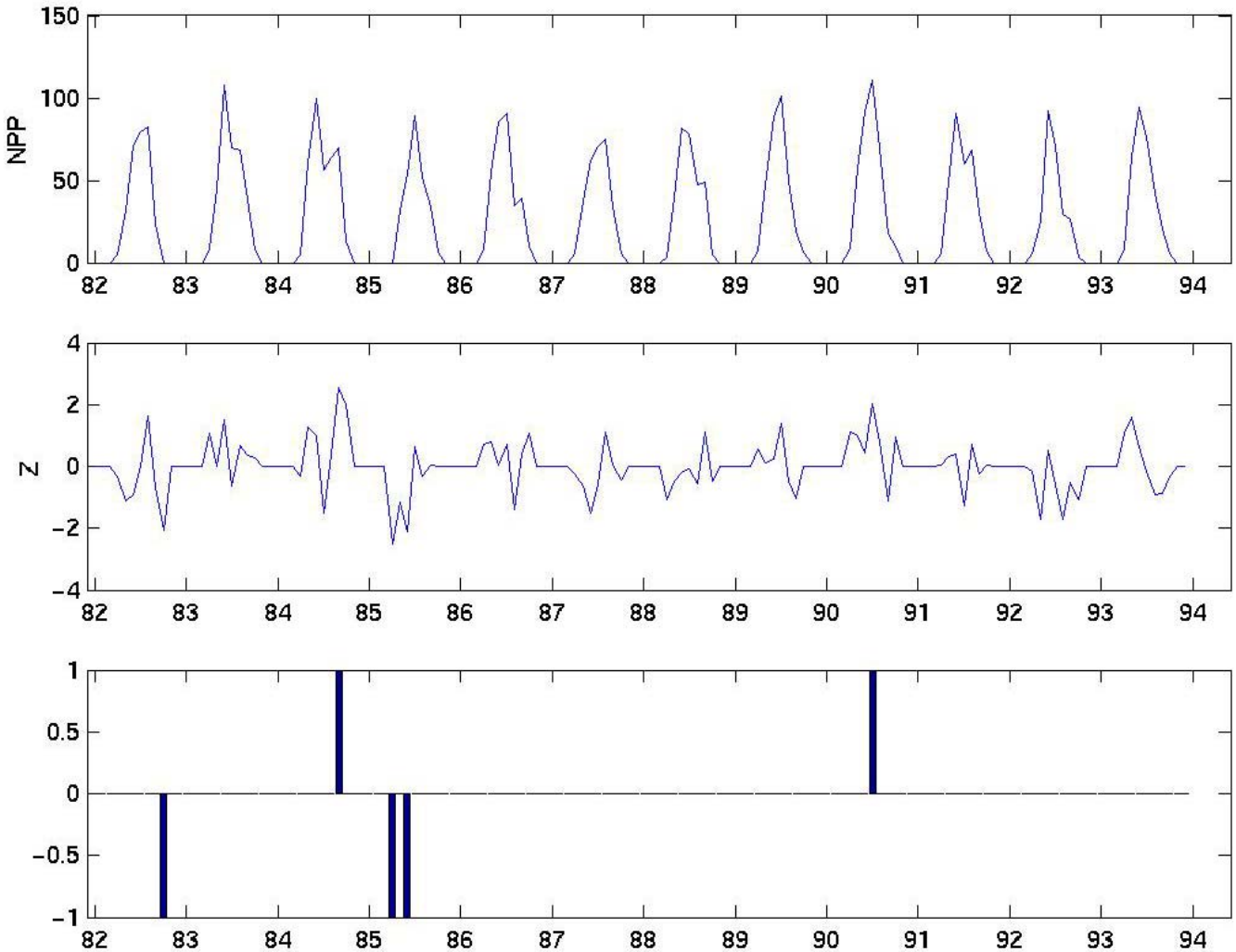


(d) Inter-zone sequential (e.g. B to the south of A in the future)

Feature Extraction

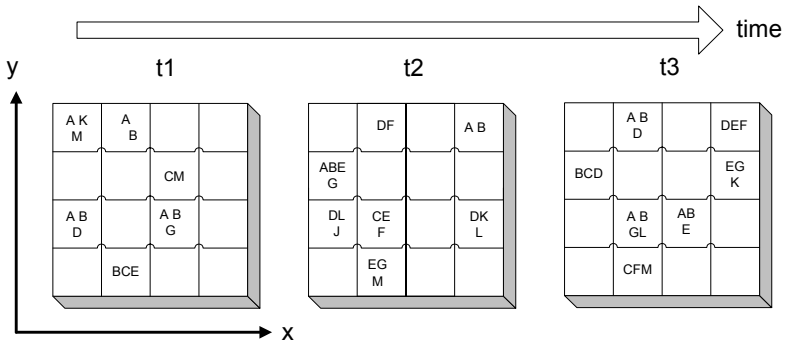
- Abstract events from time series.
- Events of interest include:
  - Temporal events:
    - Anomalous temporal events such as warmer winters and droughts.
    - Changes in periodic behavior such as longer growing seasons.
    - Trends such as increasing temperature (global warming).
  - Spatial events:
    - Large percentage of land areas in a certain region having below-average precipitation.
  - Spatio-temporal events:
    - Changes in circulation or trajectory of jet-streams.

Event Definition



Event Definition

- Convert the time series into sequence of events at each spatial location.



Grid Cell (x,y)	t1	t2	t3
(1,1)	∅	∅	∅
(1,2)	{A, B, D}	{D, L, J}	∅
(1,3)	∅	{A, B, E, G}	{B, C, D}
(1,4)	{A, K, M}	∅	∅
(2,1)	{B, C, E}	{E, G, M}	{C, F, M}
(2,2)	∅	{C, E, F}	{A, B, G, L}
(2,3)	∅	∅	∅
(2,4)	{A, B}	{D, F}	{A, B, D}
(3,1)	∅	∅	∅
(3,2)	{A, B, G}	∅	{A, B, E}
(3,3)	{C, M}	∅	∅
(3,4)	∅	∅	∅
(4,1)	∅	∅	∅
(4,2)	∅	{D, K, L}	∅
(4,3)	∅	∅	{E, G, K}
(4,4)	∅	{A, B}	{D, E, F}

Example of Intra-zone Non-sequential Associations

- Examples of intra-zone non-sequential association rules

1 PET-HI PREC-HI FPAR-HI TEMPAVE-HI ==> NPP-HI (support count = 99, confidence = 100%)

2 PET-HI TEMPAVE-LO ==> SOLAR-HI (support count = 167, confidence = 99.4%)

3 PET-HI PREC-HI FPAR-HI ==> NPP-HI (support count = 287, confidence = 98.6%)

4 NPP-LO PET-LO TEMPAVE-HI ==> SOLAR-LO (support count = 99, confidence = 98.0%)

5 PREC-HI FPAR-HI SOLAR-LO TEMPAVE-LO ==> PET-LO (support count = 154, confidence = 97.5%)

6 NPP-HI PREC-HI FPAR-HI SOLAR-LO TEMPAVE-LO ==> PET-LO (support count = 127, confidence = 97.0%)

7 NPP-LO PREC-HI SOLAR-LO TEMPAVE-LO ==> PET-LO (support count = 277 , confidence = 97.0%)

8 NPP-HI FPAR-HI SOLAR-LO TEMPAVE-LO ==> PET-LO (support count = 201, confidence = 96.6%)

9 PET-HI PREC-LO FPAR-LO TEMPAVE-HI ==> NPP-LO (support count = 126 , confidence = 95.5)

10 NPP-LO PREC-HI FPAR-LO SOLAR-LO TEMPAVE-LO ==> PET-LO (support count = 119, confidence =95.2%)

…..

147 FPAR-HI ==> NPP-HI (support count = 78108, confidence = 51.1%)

Finding Interesting Association Patterns

- Use domain knowledge to eliminate uninteresting patterns.
- A pattern is less interesting if it occurs at random locations.
- Approach:
  - Partition the land area into distinct groups (e.g., based on land-cover type).
  - For each pattern, find the regions for which the pattern can be applied.
  - If the pattern occurs mostly in a certain group of land areas, then it is potentially interesting.
  - If the pattern occurs frequently in all groups of land areas, then it is less interesting.

Example Using Land Cover Types

	Grassland	Barren land	Forests	Croplands	
Total grid points	n_1	n_2	n_3	n_4	$N = \Sigma n_i$
# grid points for which pattern can be applied	r_1	r_2	r_3	r_4	$R = \Sigma r_i$
Support count of pattern	s_1	s_2	s_3	s_4	$S = \Sigma s_i$

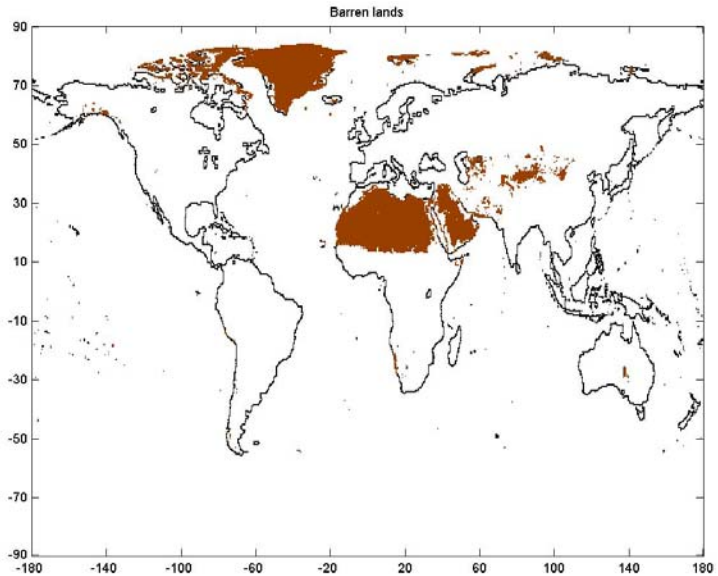
For each pattern p:

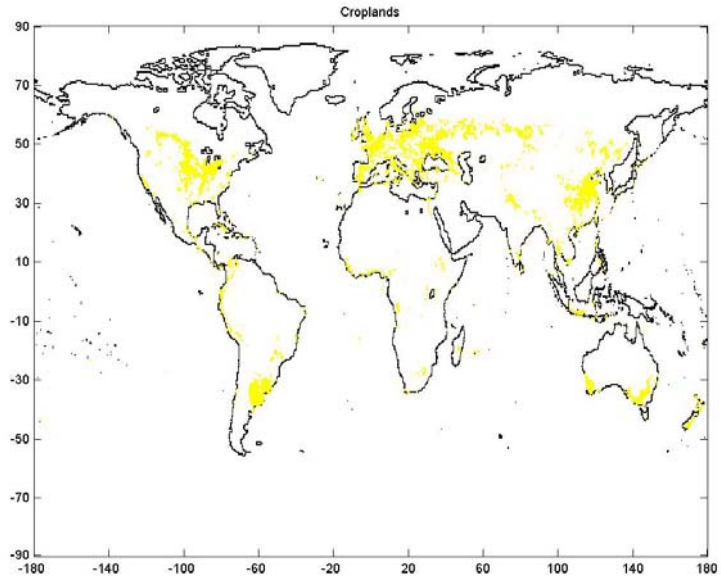
- Actual coverage for land cover type i = s_i /S
- Expected coverage for land cover type i = n_i /N
- Ratio of actual to expected coverage for land cover type i,

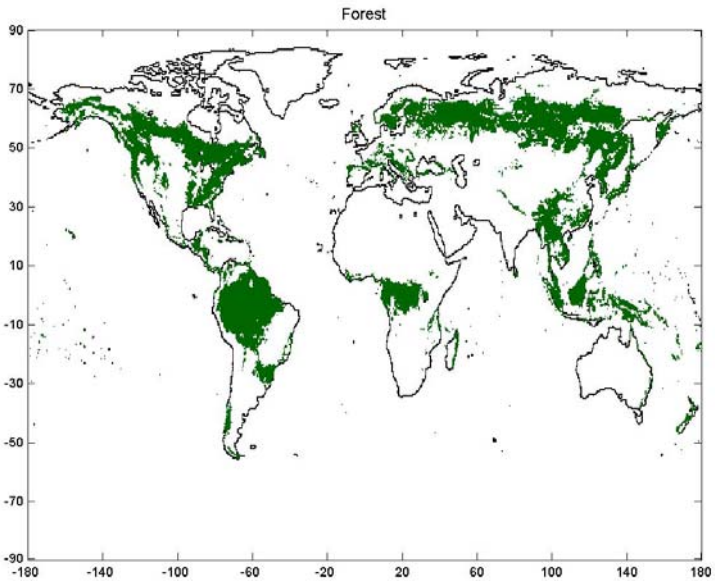
$$e_i = s_i N / n_i S$$

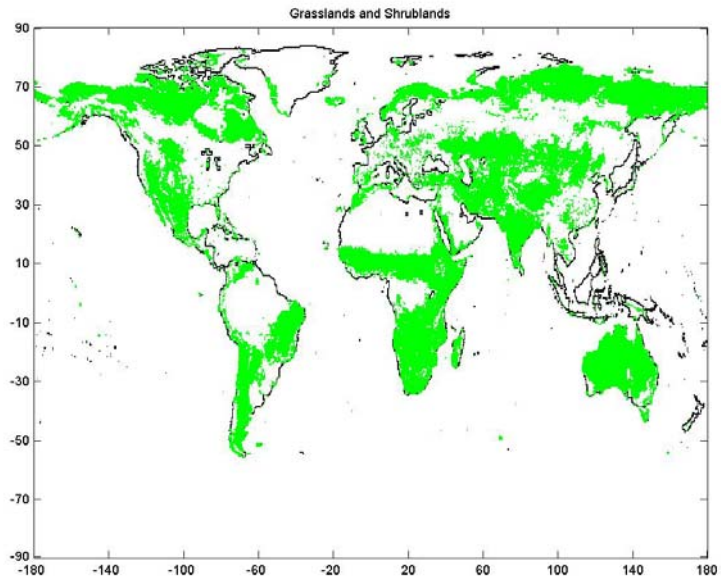
- Interest Measure $= 1 - \sum_{i=1}^{k} e_i^{\text{norm}} \log_k e_i^{\text{norm}}$
- If pattern occurs randomly, interest measure will be low.

Land Cover Types

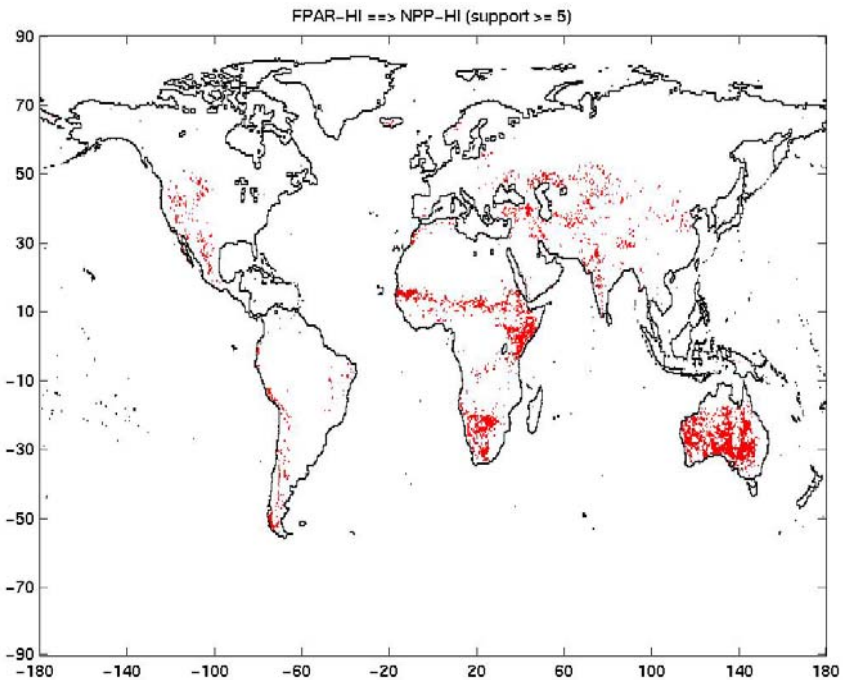




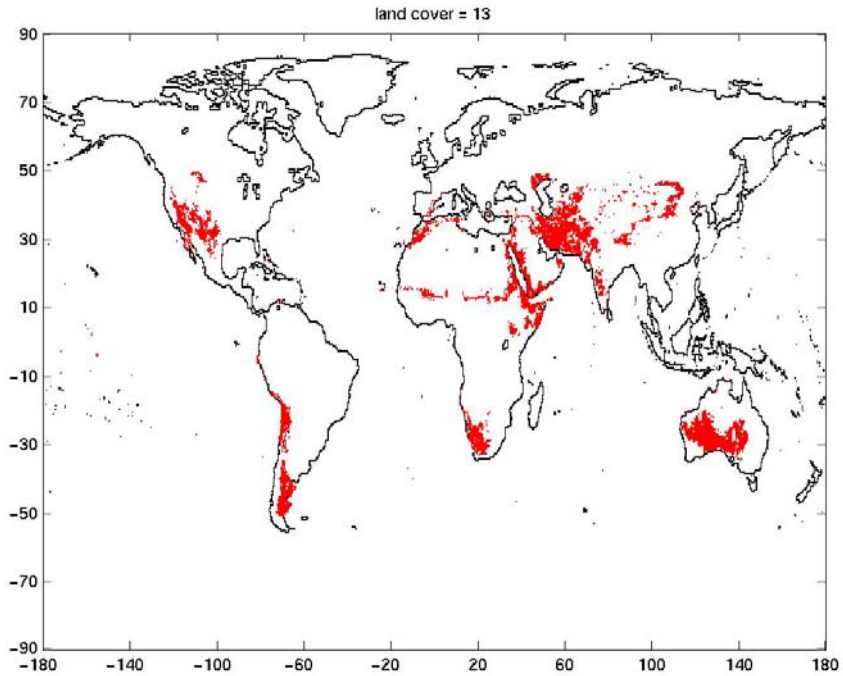




Intra-zone non-sequential Patterns



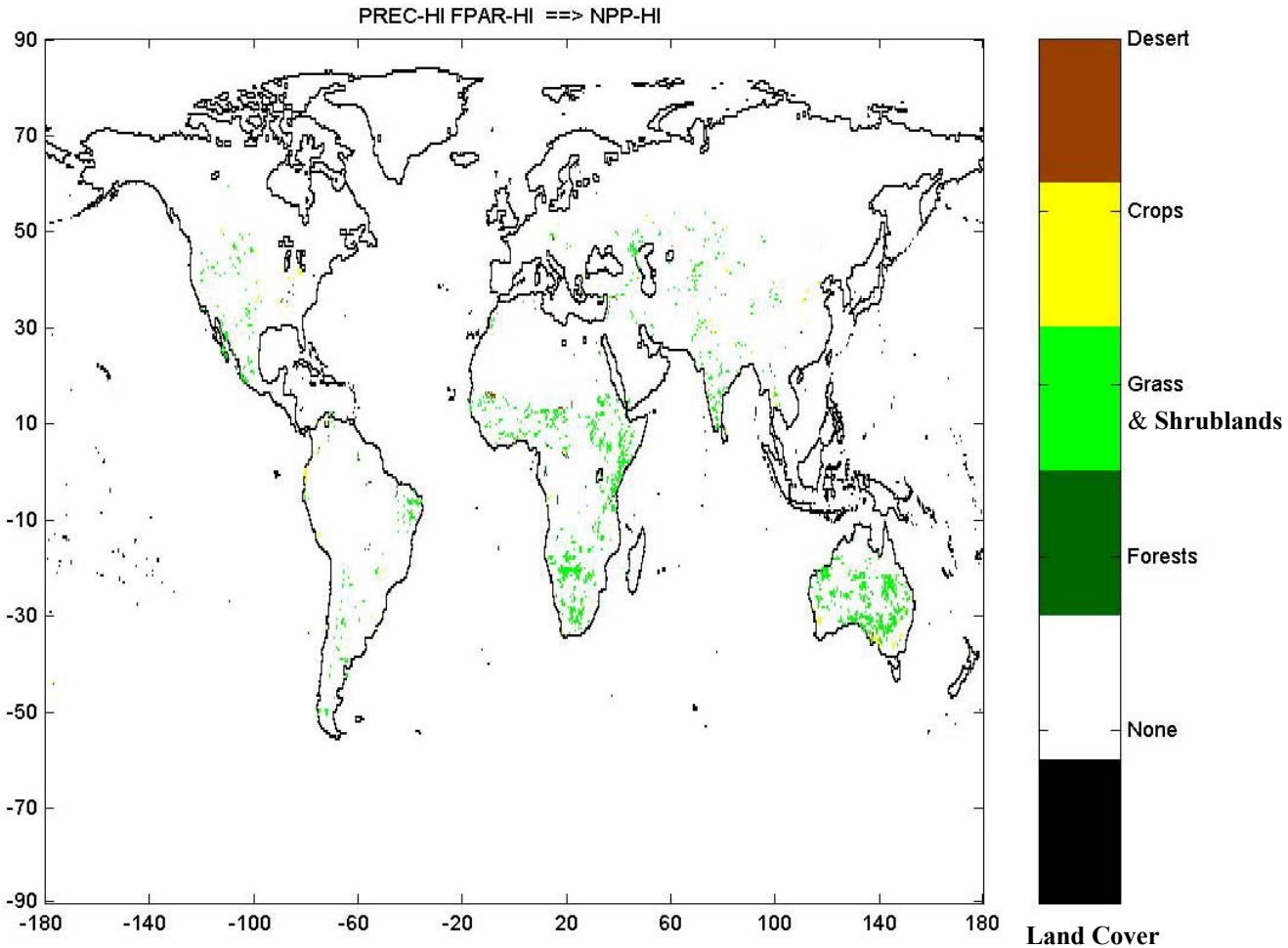
FPAR-Hi ➔ NPP-Hi (support ≥ 10)



Shrubland regions

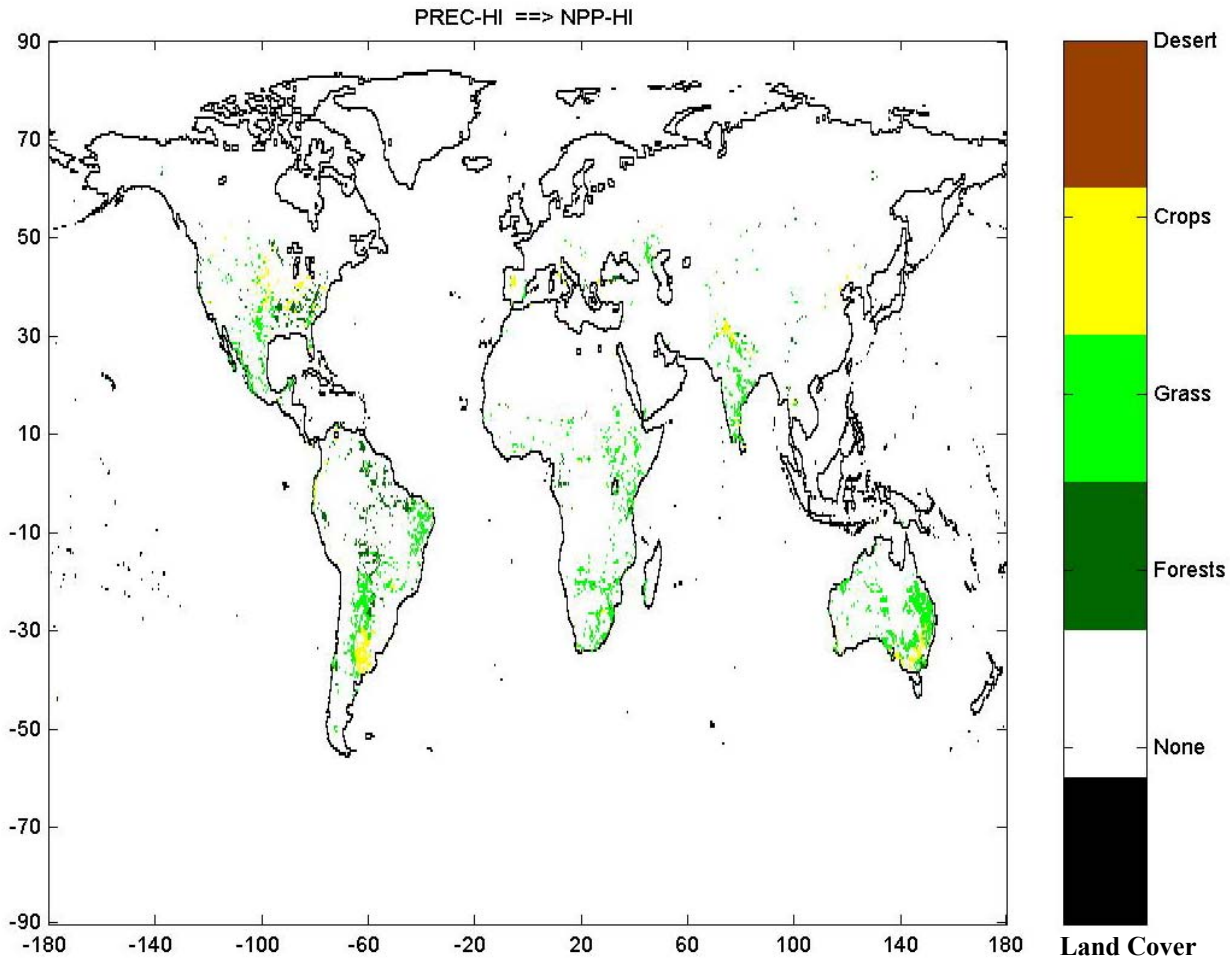
- Region corresponds to semi-arid grasslands, a type of vegetation, which is able to quickly take advantage of high precipitation than forests.
- Hypothesis: FPAR-Hi events could be related to unusual precipitation conditions.

Intra-zone non-sequential Patterns



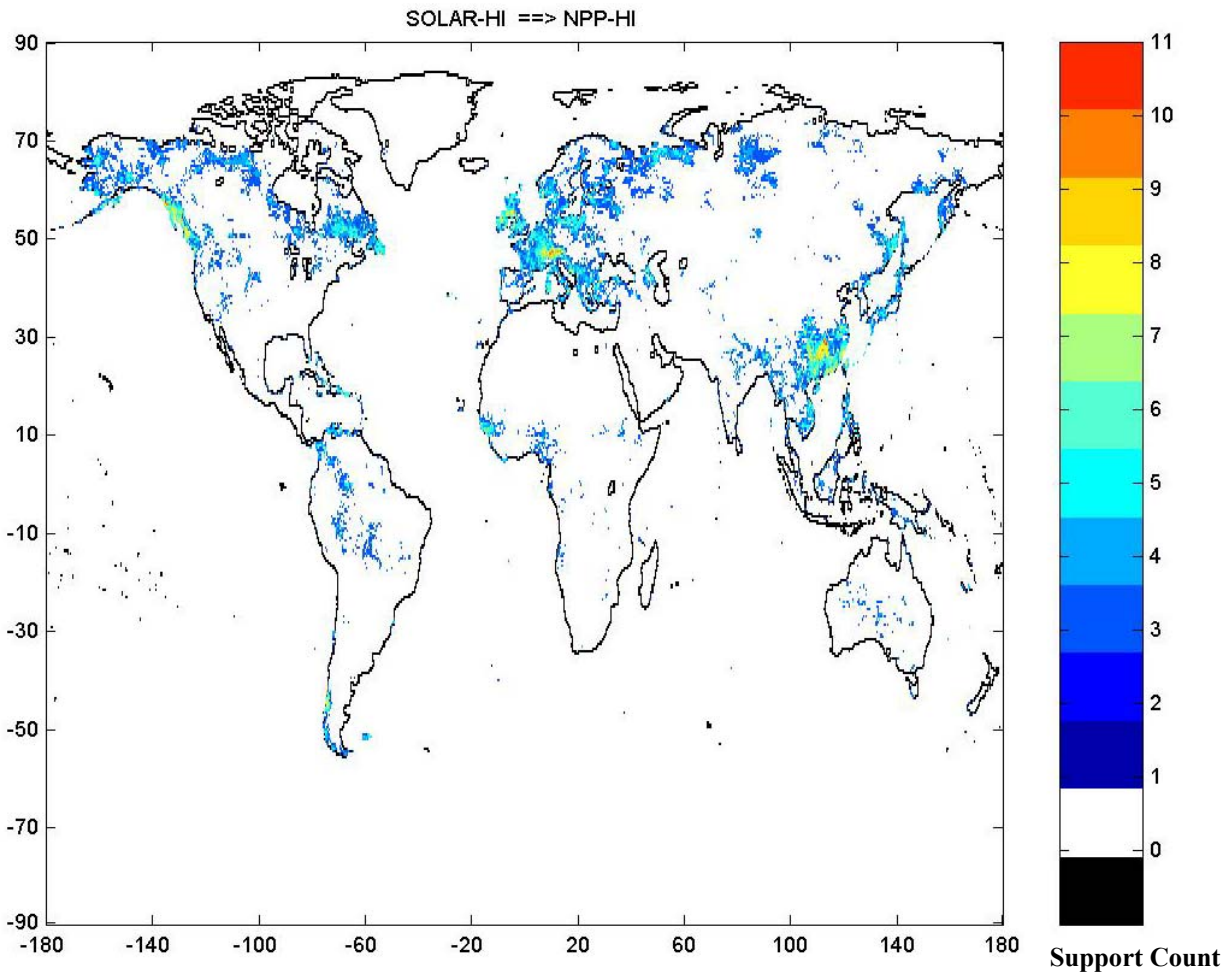
- Map agrees with hypothesis that Prec-Hi Fpar-Hi → NPP-Hi occurs mostly in shrubland and other type of grassland regions (support ≥ 3).

Intra-zone non-sequential Patterns



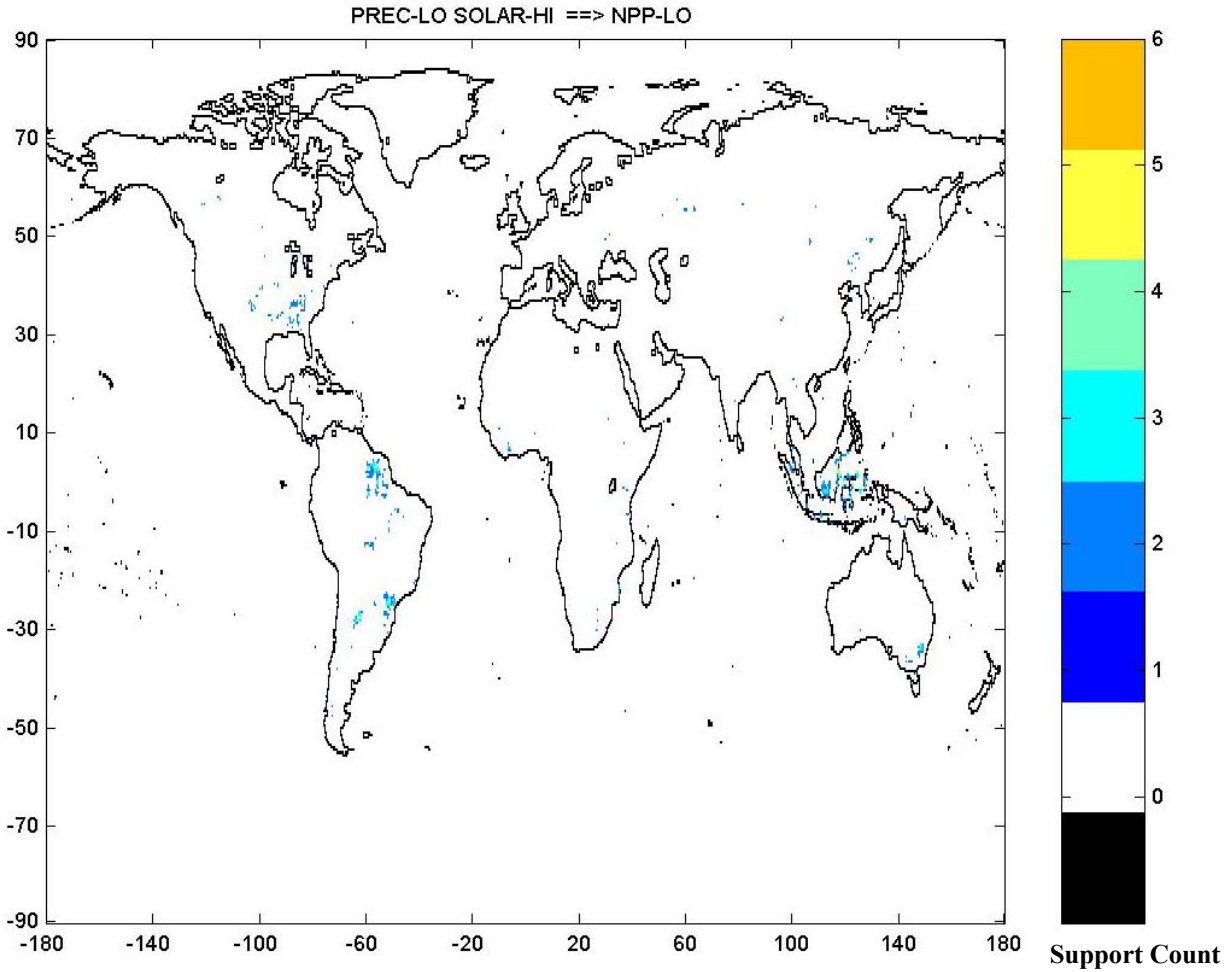
• Prec-Hi → NPP-Hi tends to occur in grassland and cropland regions (support ≥ 5).

Intra-zone non-sequential Patterns



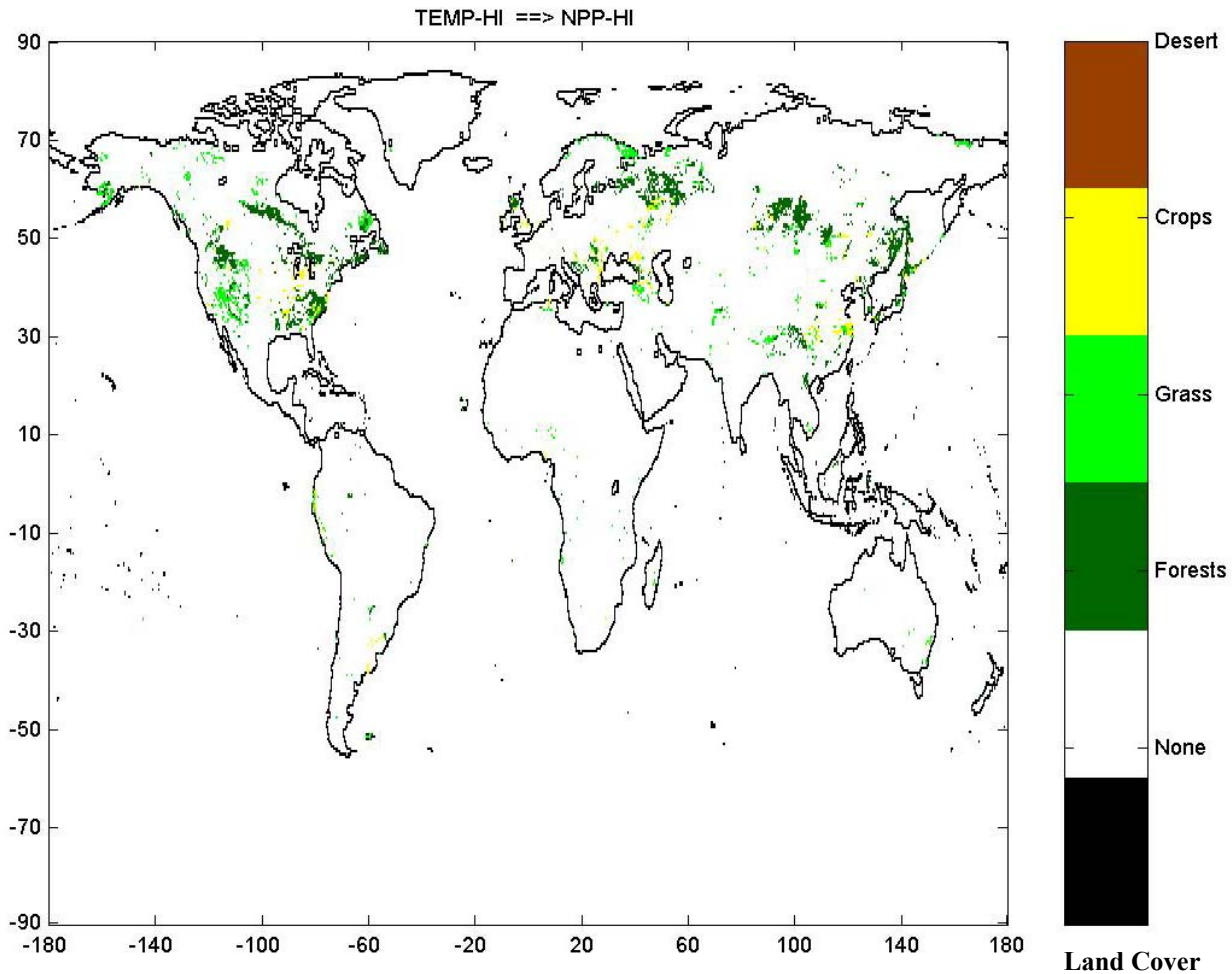
- Solar-Hi → NPP-Hi tends to occur in very cloudy (light limited) areas, like the Pacific NW and Canada/Alaska (support ≥ 3).

Intra-zone non-sequential Patterns



•Prec-Lo Solar-Hi → NPP-Lo tends to occur in drought-prone areas of tropical and sub-tropical zones, and areas of major forest fires (support ≥ 2).

Intra-zone non-sequential Patterns



• Temp-Hi → NPP-Hi tends to occur in the forest regions of the northern hemisphere (support ≥ 4).

Inter-zone and Sequential Associations

- Challenges:
  - Increased complexity due to co-occurrences of events derived from indices.
  - Support counting

Summary

- By using clustering we have made some progress towards automatically finding climate patterns that display interesting connections between the ocean and the land.
  - Possibility of discovering candidates for new climate indices.
- Association rules can uncover interesting patterns for Earth Scientists to investigate.
  - Challenges arise due to spatio-temporal nature of the data.
  - Need to incorporate domain knowledge to prune out uninteresting patterns.
- There are many statistical issues.
  - Key roles for statistics are providing some measure of confidence in the results and quantifying relationships.

Case Studies: Earth Science Data

- Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Chris Potter, Steven Klooster, Alicia Torregrosa, “Clustering Earth Science Data: Goals, Issues and Results”, Workshop on Mining Scientific Data, KDD 2001, San Francisco, CA, 2001.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Steven Klooster, Christopher Potter, Alicia Torregrosa, “Finding Spatio-Termporal Patterns in Earth Science Data: Goals, Issues and Results,” Temporal Data Mining Workshop, KDD 2001, San Francisco, CA, 2001.
- Vipin Kumar, Michael Steinbach, Pang-Ning Tan, Steven Klooster, Chris Potter, Alicia Torregrosa, “Mining Scientific Data: Discovery of Patterns in the Global Climate System”, Joint Statistical Meetings, Atlanta, GA, 2001.
- Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Chris Potter, Steven Klooster, “Data Mining for the Discovery of Ocean Climate Indices”, submitted to Workshop on Mining Scientific Data, 2002.