Application of Clustering to Earth Science Data: Progress and Challenges

Michael Steinbach
Shyam Boriah
Vipin Kumar
University of Minnesota

Pang-Ning Tan
Michigan State University

Christopher Potter
NASA Ames Research Center

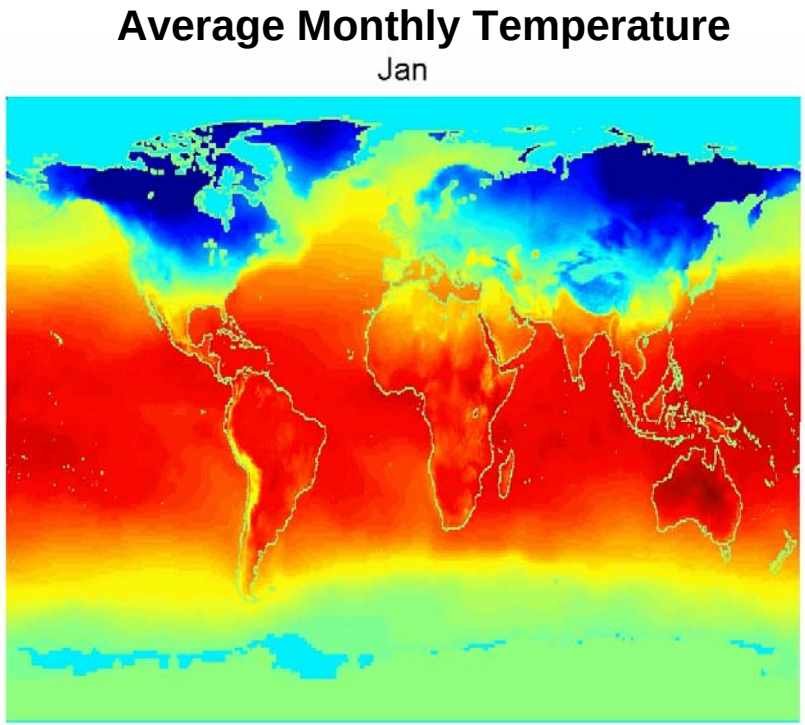
Steven Klooster
California State University, Monterey Bay

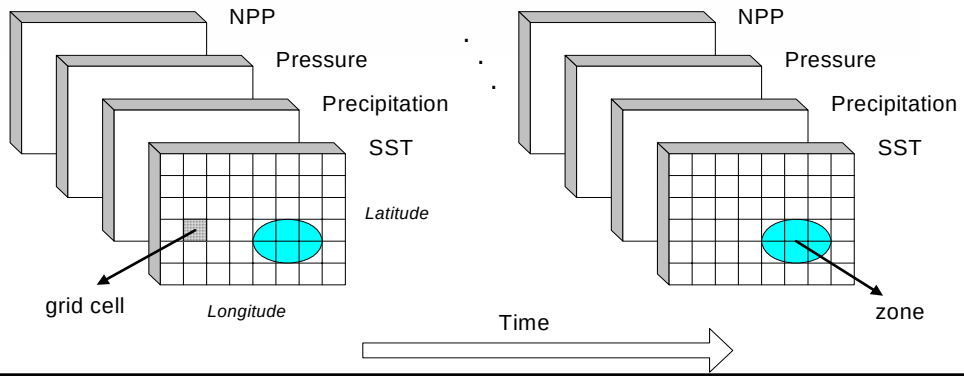
NASA funded project: *Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining*
Additional support from Army High Performance Computing Research Center

Overview

- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results
- Challenges
- Conclusion

Research Goal





Research Goal:

- Find global climate patterns of interest to Earth Scientists

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

Satellite Data:

- Global snapshots of values for a number of variables on land surfaces or water
- Span a range of 10 to 50 years
- Gridded

The El Niño Climate Phenomenon

- El Niño is the anomalous warming of the eastern tropical region of the Pacific (off the coast of Peru)

Normal Year: Trade winds push warm ocean water west, cool water rises in its place

El Niño Year: Trade winds ease, switch direction, warmest water moves east.

http://www.usatoday.com/weather/tg/wetnino/wetnino.htm

The El Niño Climate Phenomenon

Overview

- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results
- Challenges
- Conclusion

Climate Indices: Connecting the Ocean/Atmosphere and the Land

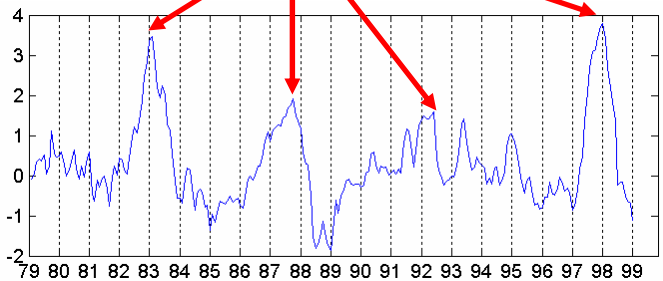
- A climate index is a time series of temperature or pressure
  - Similar to business or economic indices
  - Based on Sea Surface Temperature (SST) or Sea Level Pressure (SLP)

Dow Jones Index (from Yahoo!)

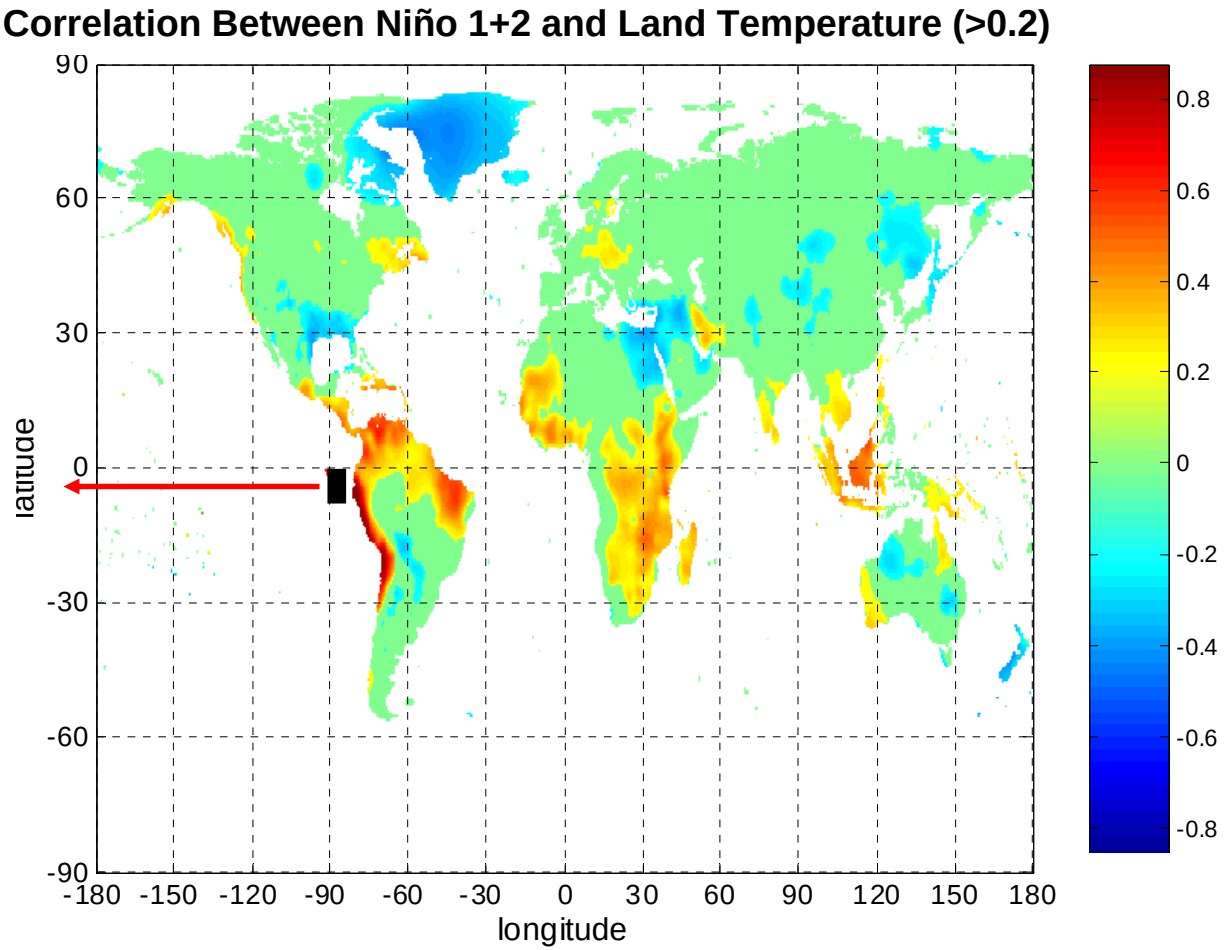
- Climate indices are important because
  - They distill climate variability at a regional or global scale into a single time series
  - They are well-accepted by Earth scientists
  - They are related to well-known climate phenomena such as El Niño

A Temperature Based Climate Index: NINO1+2

El Niño Events

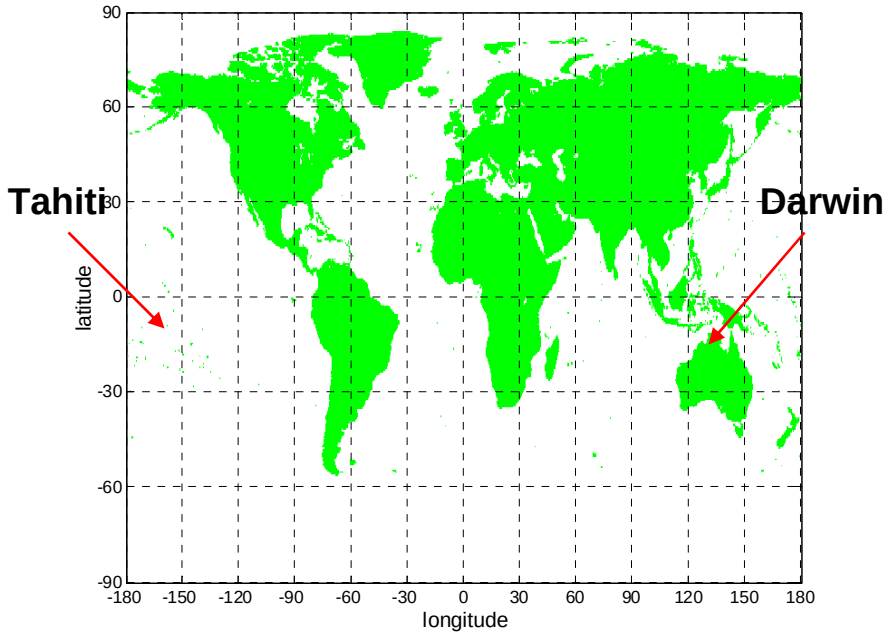


Niño 1+2 Index



A Pressure Based Climate Index: SOI

- The Southern Oscillation Index (SOI) is also associated with El Niño.
- Defined as the normalized pressure differences between Tahiti and Darwin Australia.
- Both temperature and pressure based indices capture the same El Niño climate phenomenon.



List of Well Known Climate Indices

Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP poleward of 20° N
PDO	**Pacific Decadel Oscillation:** Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	**Quasi-Biennial Oscillation Index:** Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	**Cold Tongue Index:** Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	**Western Pacific:** Represents a low-frequency temporal function of the 'zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

Overview

- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results
- Challenges
- Conclusion

Discovering Climate Indices: Traditional Approaches

- Earth scientists have discovered currently known climate indices
  - Observation
    - The El Niño phenomenon was first noticed by Peruvian fishermen centuries ago
    - They observed that in some years the warm southward current, which appeared around Christmas, would persist for an unusually long time, with a disastrous impact on fishing
  - Eigenvalue techniques such as Singular Value Decomposition (SVD)

Overview

- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results
- Challenges
- Conclusion

Discovering Climate Indices via Data Mining

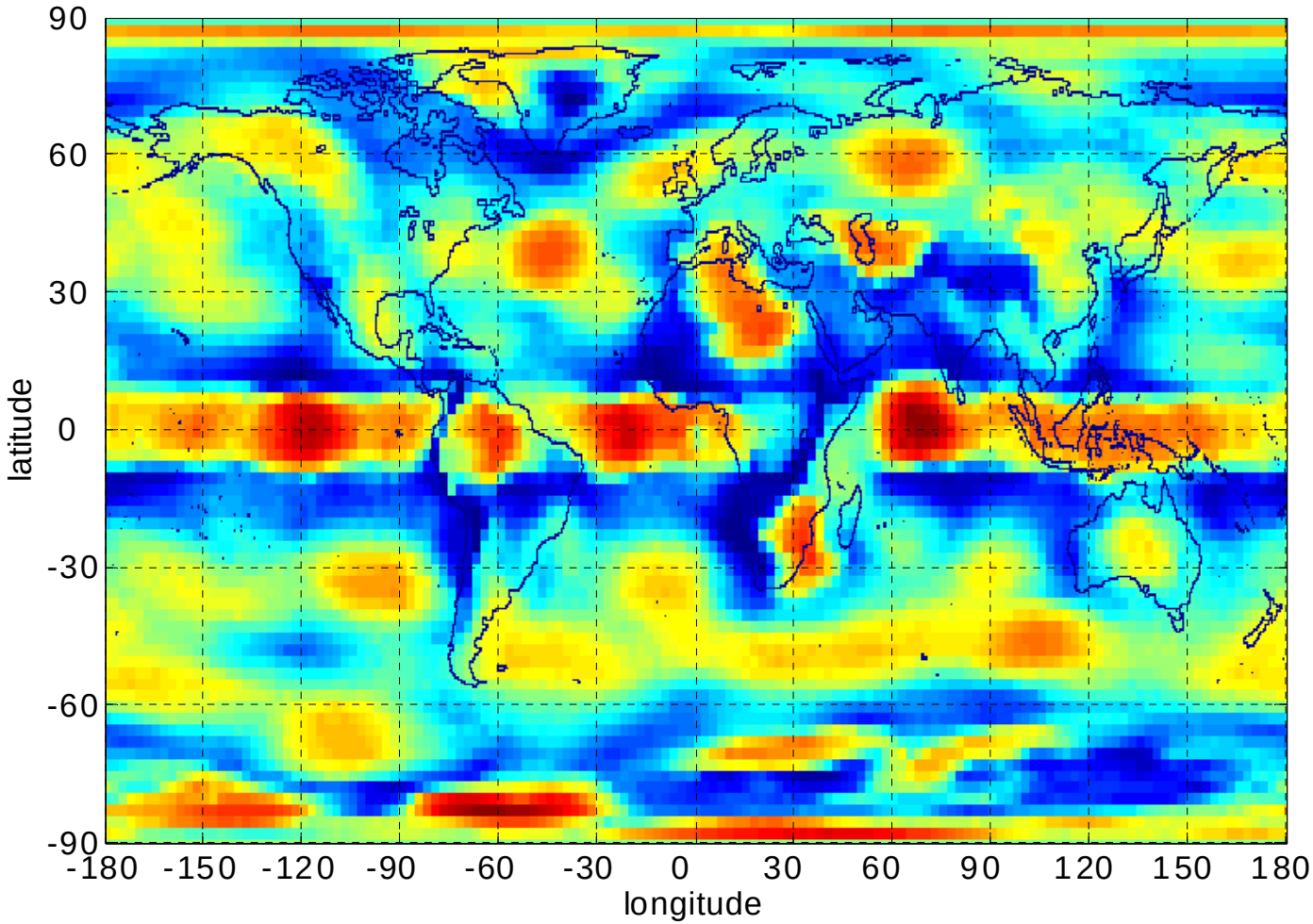
- Clustering provides an alternative approach for finding candidate indices
  - Clusters represent ocean regions with relatively homogeneous behavior
  - The centroids of these clusters are time series that summarize the behavior of these ocean areas, and thus, represent potential climate indices

- Need to evaluate the “influence” of potential indices on land points

Shared Nearest Neighbor (SNN) Clustering

- Density based clustering approach
  - Determine the density of each point (time series)
    - Density is high if most of your neighbors have you as a neighbor
  - Perform the clustering using the density
    - Identify and eliminate noise and outliers, which are points with low density
    - Identify core points, which are time series with high density
    - Build clusters around the core points

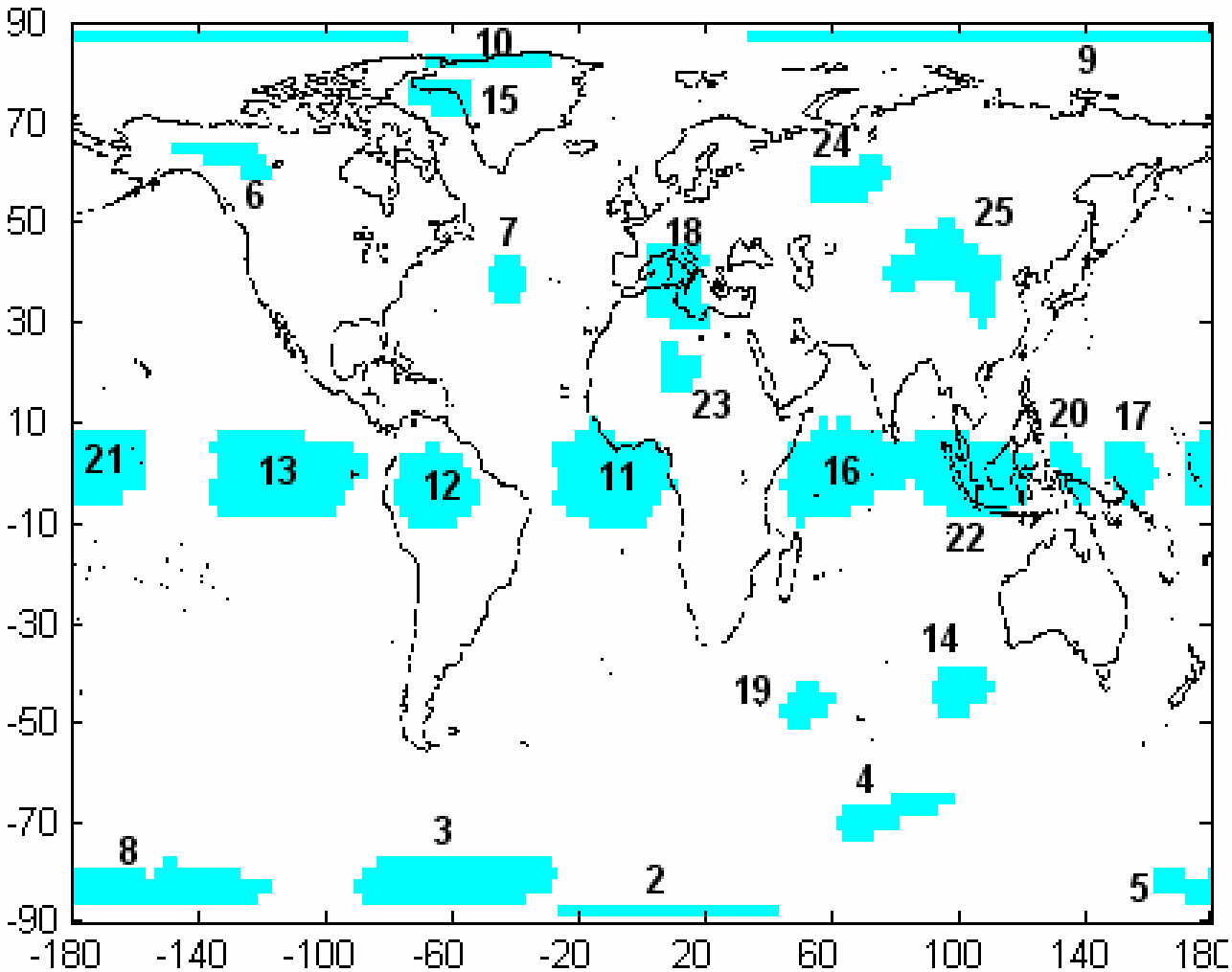
SNN Density of SLP Time Series



Redder areas are high density, i.e., high homogeneity.

SLP Clusters

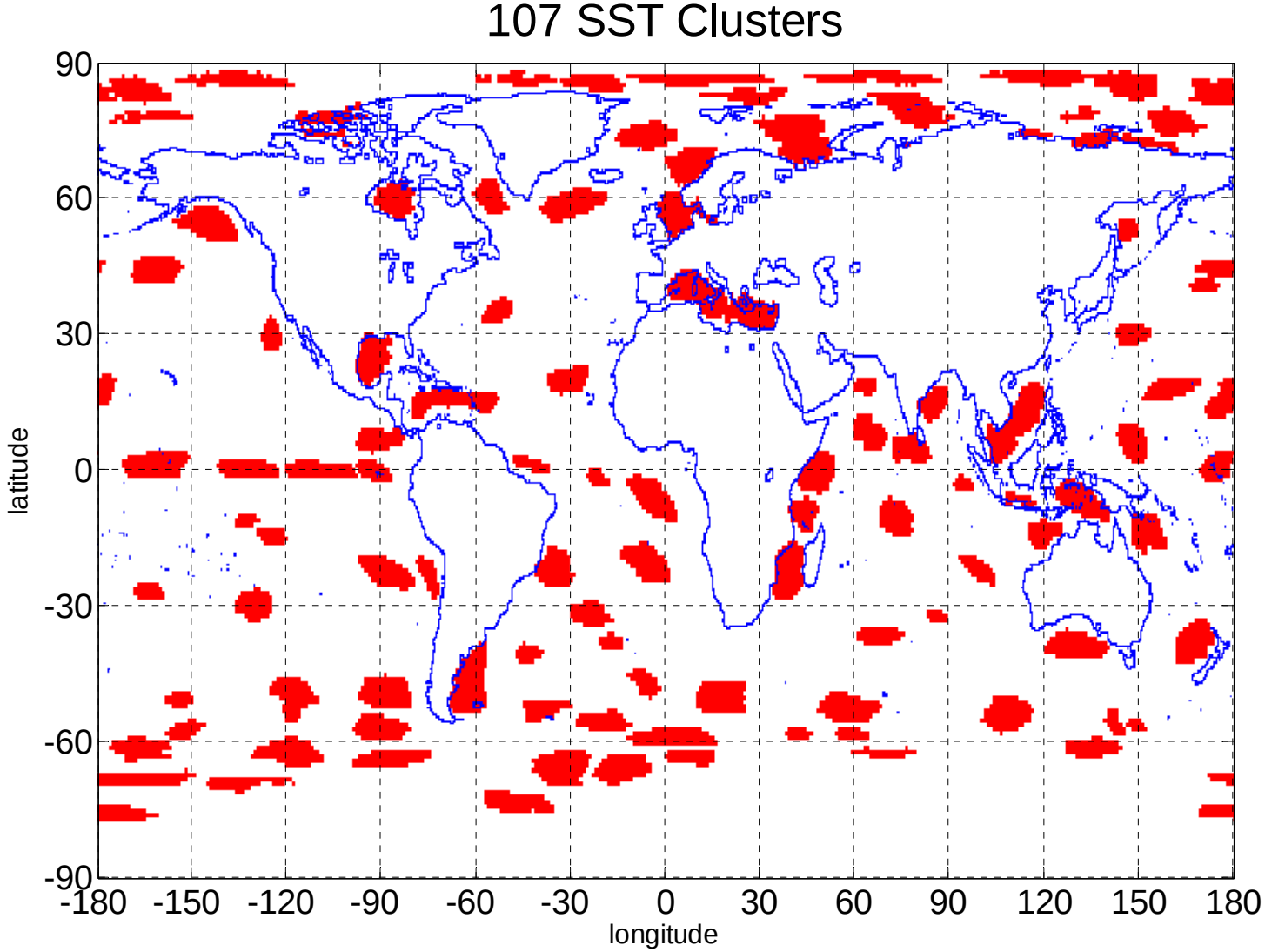
25 SLP Clusters



Overview

- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results: SST-based clusters
- Challenges
- Conclusion

SST Clusters



Correlation of Known Indices with SST Cluster Centroids and SVD Components

Known Indices	Best Matching SST Cluster Centroid	Best Matching SVD Component
SOI	0.7006	0.5427
NAO	0.2973	0.1774
AO	0.2383	0.2301
PDO	0.5172	0.4684
QBO	0.2675	0.3187
CTI	0.9147	0.6316
WP	0.2590	0.1904
NINO1+2	0.9225	0.5419
NINO3	0.9462	0.6449
NINO3.4	0.9196	0.6844
NINO4	0.9165	0.6894

SST based cluster centroids have better correlation to known indices than SVD based indices in all but one case.

Red indicates higher magnitude of correlation.

Area-weighted Correlation of Known Indices with SST Cluster Centroids and SVD Components

Known Indices	Area Weighted Correlation for		
	Index	Best Centroid	Best SVD Component
SOI	0.1550	0.1768	0.1240
NAO	0.1328	0.1387	0.0929
AO	0.1682	0.1912	0.0929
PDO	0.1378	0.1377	0.0891
QBO	0.0671	0.1377	0.0850
CTI	0.1702	0.1708	0.1240
WP	0.1117	0.1714	0.1240
NINO1+2	0.1558	0.1608	0.2091
NINO3	0.1774	0.1708	0.2091
NINO 3.4	0.1800	0.1714	0.2091
NINO 4	0.1696	0.1768	0.2091

SST based cluster centroids have higher area-weighted correlation than SVD based indices and known indices in most cases.

Red indicates higher correlation.

Overview

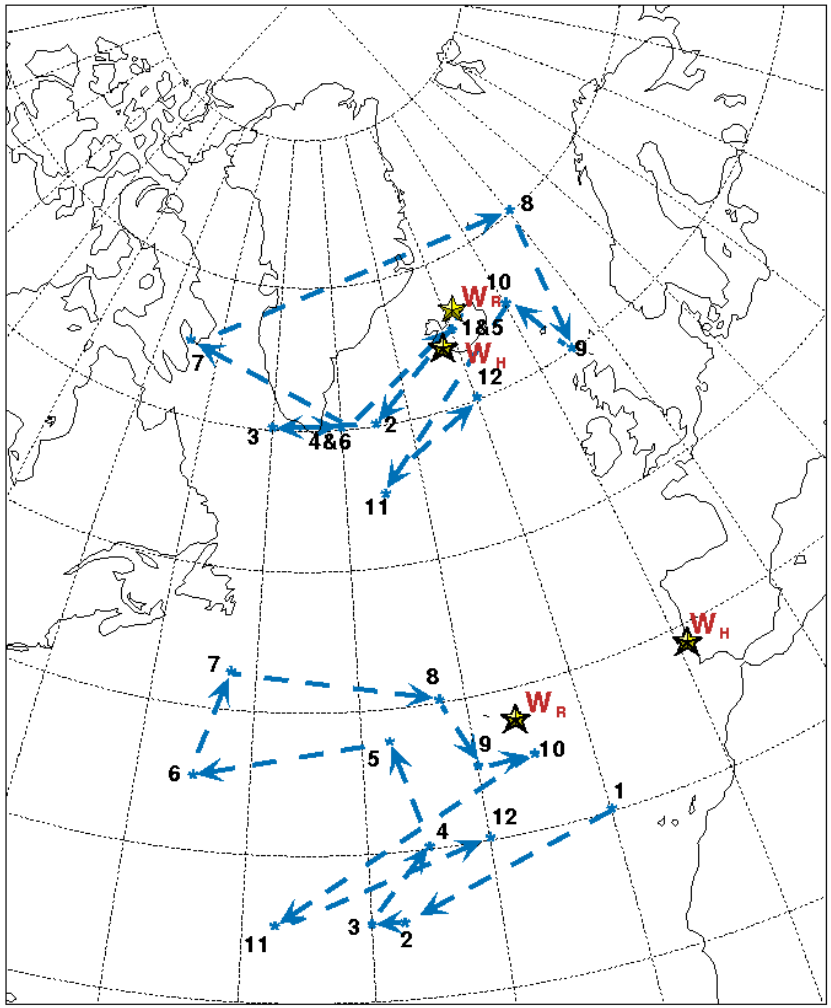
- Background
- Climate indices
- Techniques for Discovering Climate Indices
  - Traditional approaches
  - Clustering
- Results
- Challenges
- Conclusion

“Dynamic” Clusters

Motivation:

- In Earth Science, most phenomena of interest evolve in space and time
- Most well-known indices based on data collected at fixed land stations
- However, underlying phenomenon may not occur at exact location of the land station. e.g. NAO

An example – NAO



Source: Portis et al, Seasonality of the NAO, AGU Chapman Conference, 2000.

- NAO refers to swings in the atmospheric sea level pressure difference between the Arctic and the subtropical Atlantic
- Computed as the normalized difference between SLP at a pair of land stations in these two regions of the North Atlantic Ocean
- Exact location of centers of the phenomena do not necessarily correspond to fixed land stations

Dynamic Clusters

- Need for novel clustering approaches that can find “dynamic” or “moving” clusters
- Climate phenomenon can be captured much more accurately by having a notion of a dynamic cluster
- This would represent a dynamic phenomenon based on satellite data for the entire region
- Such clusters might move in space and/or expand or contract in extent over time
- Successful development of dynamic clustering techniques will provide better insight as to how climate indices and their impact on land climate change over time

Previous work in this area

- Two common approaches:

1. View data as collection of snapshots taken at different time periods[1]
   - Spatial clustering is performed for each snapshot independently
   - Correspondence between clusters in different time periods established by measuring fraction of objects the clusters share in common
2. Spatial clustering initially applied to data in first snapshot[2]
   - Clusters updated incrementally taking into account changes due to moving objects that leave or join the cluster as time progresses

- Limitations: Clustering performed along one of the dimensions (either spatial or temporal, not both) – loses information about contiguity of objects along other dimension

1. P. Laknis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In Proc. of the 9th International Symposium on Spatial and Temporal Databases (SSTD 2005), Angra dos Reis, Brazil, 2005.
2. Y. Li, J. Han, and J. Yang. Clustering moving objects. In Proc. of the 2004 ACM SIGKDD Int'l Conf on Know ledge Discovery and Data Mining, 2004.

Possible approaches

- Directly optimize a spatio-temporal objective function
- Partition space-time dimensions into spatio-temporal cells, and aggregate cells that have similar values while preserving spatial and temporal proximities
- Challenges:
  - Feature extraction: granularity of spatio-temporal cells may be too fine. Need to determine (spatial and temporal) neighborhood size
  - Spatio-temporal clustering: Formulate objective function.
  - Spatial constraints: MBR?
  - Visualization

Conclusion

- Briefly described current progress in applying clustering to find climate indices
- One of the remaining challenges is to model dynamic clusters
- Future work:
  - Investigate several approaches to modeling dynamic clusters
  - Take into account domain specific factors such as seasonality, land cover and geographical boundaries
  - Other challenges include those applicable to most clustering algorithms, i.e. handling outliers, parameter initialization and scalability

Questions?

Project web page:

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