

# Discovery of Patterns from Global Earth Science Data Sets



Vipin Kumar

University of Minnesota

kumar@cs.umn.edu  
www.cs.umn.edu/~kumar



Collaborators and Group Members:

**Chris Potter**  
NASA Ames

**Steve Klooster**  
California State University

**Sudipto Banerjee, Shyam Boriah**  
**Joe Knight, Michael Steinbach**  
University of Minnesota

**Pang-Ning Tan**  
Michigan State University



Research funded by ISET-NOAA, NSF, NASA and Cisco

# Discovery of Patterns from Global Earth Science Data Sets

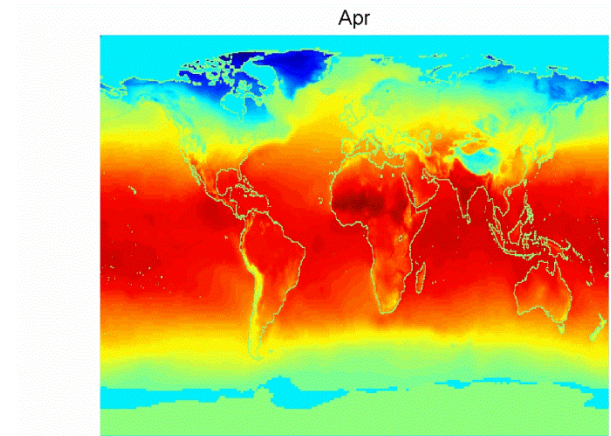
**Science Goal:** Understand global scale patterns in biosphere processes

## Earth Science Questions:

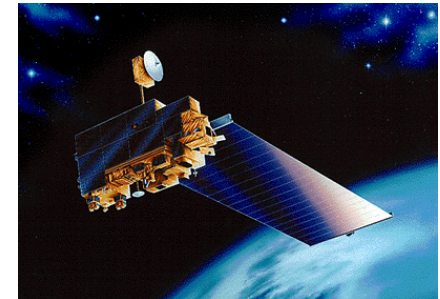
- What is the scale and location of natural and human-induced changes?
- How are ocean, atmosphere and land processes coupled?

## Data sources:

- Weather observation stations
- High-resolution EOS satellites
  - 1982-2000 AVHRR at  $1^\circ \times 1^\circ$  resolution ( $\sim 115\text{km} \times 115\text{km}$ )
  - 2000-present MODIS at  $250\text{m} \times 250\text{m}$  resolution
- Model-based data from forecast and other models
  - Sea level pressure 1979-present at  $2.5^\circ \times 2.5^\circ$
  - Sea surface temperature 1979-present  $1^\circ \times 1^\circ$
- Data sets created by data fusion



Monthly Average Temperature

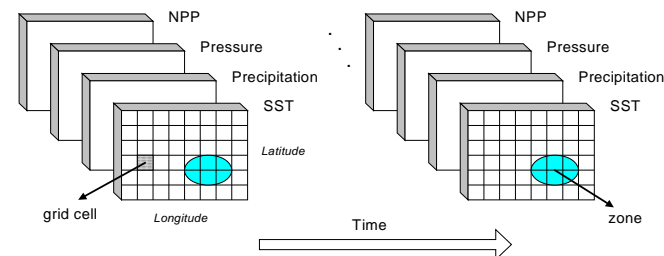


Earth Observing System

# Data Mining Challenges

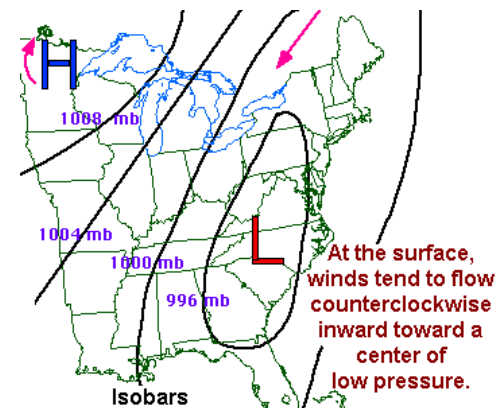
- **Spatio-temporal nature of data**

- Traditional data mining techniques do not take advantage of spatial and temporal autocorrelation.



- **Scalability**

- Size of Earth Science data sets can be very large, especially for data such as high-resolution vegetation
- Grid cells can range from a resolution of  $2.5^\circ \times 2.5^\circ$  (10K locations for the globe) to  $250\text{m} \times 250\text{m}$  (15M locations for just California; about 10 billion for the globe)



- **High-dimensionality**

- Long time series are common in Earth Science

- **Noise and missing values**

# Interdisciplinary Collaboration

Team Members include Earth Scientists, Remote Sensing scientists, Computer Scientists (data mining, HPC)

Results of Collaboration:

- Over a dozen publications in Earth Science Journals
- Several publications in computer science and Earth Science conferences
- Two NASA press releases, *Mechanical Engineering* cover article

NASA News

National Aeronautics & Space Administration  
Ames Research Center  
Moffett Field, California 94034-1000



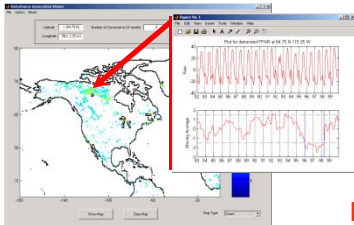
## NASA DATA MINING REVEALS A NEW HISTORY OF NATURAL DISASTERS

NASA is using satellite data to paint a detailed global picture of the interplay among natural disasters, human activities and the rise of carbon dioxide in the Earth's atmosphere during the past 20 years....

[http://www.nasa.gov/centers/ames/news/releases/2003/03\\_51AR.html](http://www.nasa.gov/centers/ames/news/releases/2003/03_51AR.html)



## Detection of Ecosystem Disturbances:



This interactive module displays the locations on the earth surface where significant disturbance events have been detected.

**Disturbance Viewer**



ON THE COVER: *Data Mining*

## mining what others miss

Highlighting the subtleties in  $10^{12}$  bytes of data, technology tries to clear up its own mess.

Simulating natural phenomena, mapping the human genome, and discovering ways to improve product quality all have one thing in common: They generate tremendous amounts

# Focus of the talk: Two Earth Science Questions

---

- Land Cover Change Detection
  - What is the scale and location of land cover changes and their impact on the carbon cycle?
- Climate Indices: Connecting the Ocean/Atmosphere and the Land
  - How are ocean, atmosphere and land processes coupled?

# Land Cover Change Detection

**Goal:** Determine **where**, **when** and why land cover changes occur

- E.g. Deforestation, Urbanization, Agricultural intensification, crop rotation

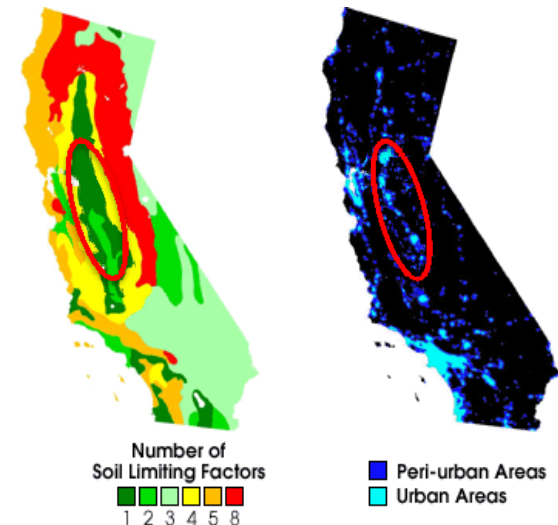
## Motivation:

- Land cover change has impacts on a wide range of issues from Local climate to Diversity/abundance of terrestrial species to Commodity prices
- Conversion of natural land cover can have undesirable environmental consequences, such as on the carbon cycle



**Deforestation** changes local weather. Cloudiness and rainfall can be greater over cleared land (image right) than over intact forest (left).

**Urbanization** E.g., housing development in prime agricultural land.





# Forest Cover Change

- Changes in forests account for over **20%** of the greenhouse gas emissions
    - 2<sup>nd</sup> only to fossil fuel emissions
  - Terrestrial carbon can provide up to **25%** of the climate change solution
  - Ability to monitor changes in global forest cover over space and time is critical for enabling inclusion of forests in carbon trading
- ⇒ The need for a scalable technological solution to assess the state of forest ecosystems and how they are changing has become increasingly urgent.



**Deforestation** moves large amounts of carbon into the atmosphere in the form of CO<sub>2</sub>.

# Data

Rich amounts of data from remotely sensed images are available for detecting changes in land cover.

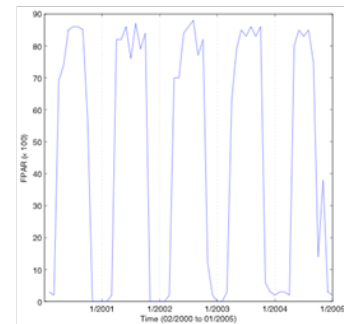


Global EVI in Summer, 2000.



Global EVI in Winter, 2001.

- MODIS algorithms have been used to generate the EVI at 250-meter spatial resolution from Feb 2000 to the present
- Enhanced Vegetation Index (EVI) represents the "greenness" signal (area-averaged canopy photosynthetic capacity), with improved sensitivity in high biomass cover



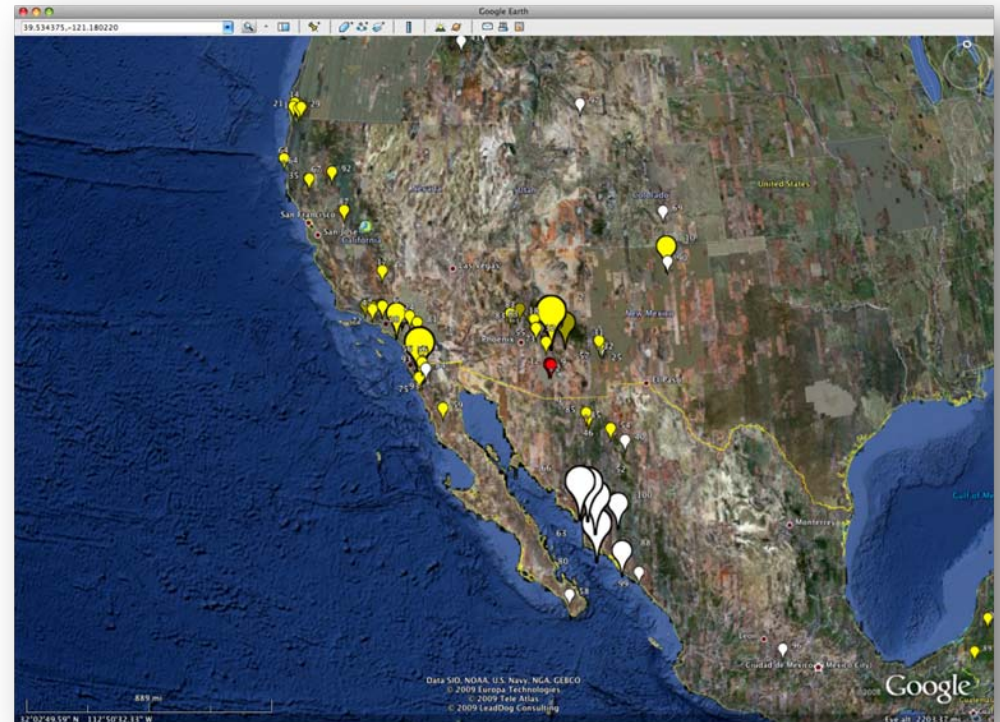
NASA's Terra satellite platform launched in 1999 has the Moderate Resolution Imaging Spectroradiometer (MODIS)

**Image Source:** NASA/Goddard Space Flight Center Scientific Visualization Studio



# Goals

- Advance state of the art in land cover change detection using a time series approach
- **Develop new algorithms:**
  - novel time series change detection
  - land cover change characterization
- **Facilitate regional and global analysis** of major changes in land cover
- Form the basis for a system that will help quantify the **carbon impact** of these changes
- Provide **ubiquitous web-based access** to changes occurring across the globe, creating public awareness



# Previous work: Land cover change detection

- Primarily based on examining differences between two or more satellite images acquired on different dates.
- Focus has been on relatively small areas.
- Detected only changes of specific types of interest.

## Limitations:

- Unable to detect changes outside the image acquisition window.
- Difficult to identify when the change has occurred.
- Parameters such as rate of change, extent, speed, and pattern of growth cannot be derived.
- Quantitative assessment of carbon impact cannot be derived
- Inherently unsuitable for global analysis.



# Previous work: Time Series Change Detection

Time series change detection problem has been addressed in a variety of fields under different names:

E.g. statistical process control, curve segmentation (computer graphics & vision), segmented regression

- Statistics
- Signal processing
- Control theory
- Industrial process control
- Computer graphics & vision (curve segmentation)
- Network Intrusion Detection
- Fraud Detection (telecommunications, etc.)
- Health Care (Statistical Surveillance)
- Industrial Processes (process control and quality control)
- **Land Cover Change**

- Parameter Change  
CUSUM-type approaches, Page [1957], Chernoff and Zacks [1964], Picard [1985]
- Segmentation  
Linear Model: Himberg et al. [2001], Keogh et al. [2001], Hawkins and Merriam [1973]  
Polynomial Model: Guralnik and Srivastava [1999]  
Wavelet Model: Sharifzadeh et al. [2005]
- Predictive  
Ge and Smyth [2000], Roy, Jin, Lewis and Justice [2005]
- Subspace Approach  
Moskvina and Zhigljavsky [2003]
- Anomaly Detection  
Chan and Mahoney [2005], Yamanishi and Takeuchi [2002], Ide and Kashima [2004], Chandola, Banerjee and Kumar [2008]

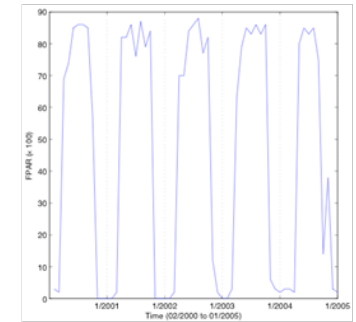
# Limitations of Existing Time Series Change Detection Techniques

---

- Many techniques do not scale to massive datasets
  - Designed for **single** long time series
  - Earth Science data sets can be very large
    - ◆ 2.5° x 2.5°: 10K locations for the globe
    - ◆ 250m x 250m: 10 billion for the globe
- Seasonality of Earth Science data and/or intra-season variability is not accounted for.
- Spatial and temporal autocorrelation are not exploited.
- Results may be difficult to interpret, e.g. from a predictive model.

# A novel change detection algorithm: Preliminary Results

- Detects land cover change using a time series approach
- High accuracy
- Low computational requirements
- Intuitive, interpretable



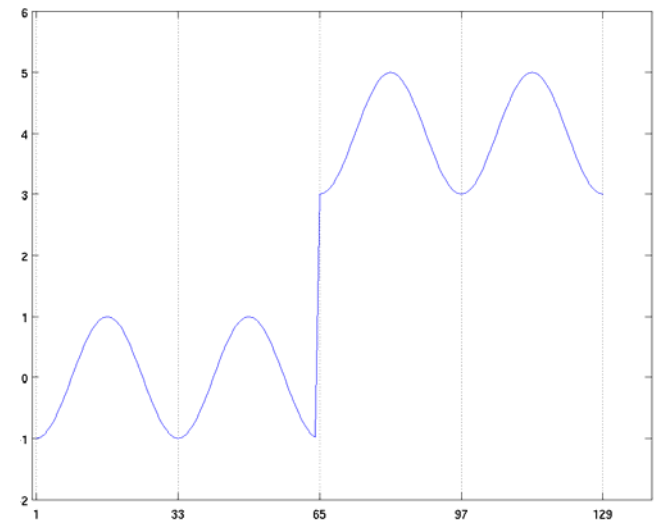
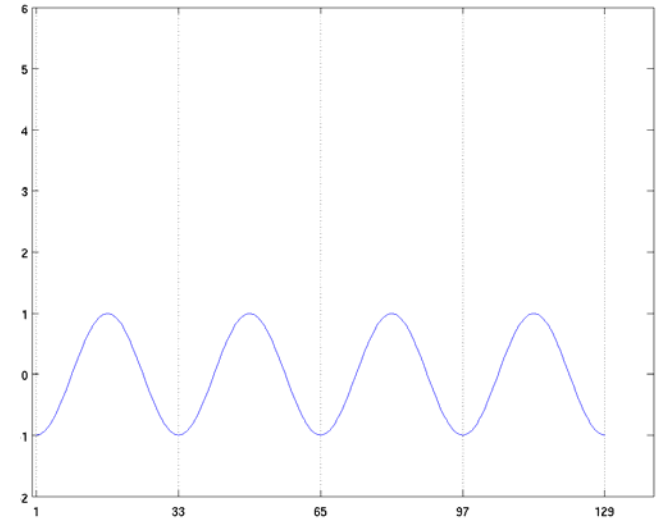
- Application to 250m x 250m EVI data from California
  - Detects a number of interesting land cover changes including logging, forest fires and conversion from desert to farmland [Boriah 2008].
- Application on a global scale using 4km x 4km FPAR data over land areas classified as forest
  - Detected major forest fires worldwide.

S. Boriah, V. Kumar, M. Steinbach, et al., *Land cover change detection: a case study*, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, Las Vegas, Nevada, USA, 2008.

# Algorithm: Recursive Merging

- Exploit the major mode of behavior (yearly cycle) in order to detect changes.
- The time series for each location is processed as follows:
  1. The two most similar seasons are merged, and the distance/similarity is stored.
  2. Step 1 is applied recursively until one season is left.
  3. The **change score** for this location is based on whether any of the observed distances are extreme (e.g. ratio of maximum distance/minimum distance).

The algorithm produces a ranked list of pixels that are most likely to have changed, when applied to a data set of multiple time series.



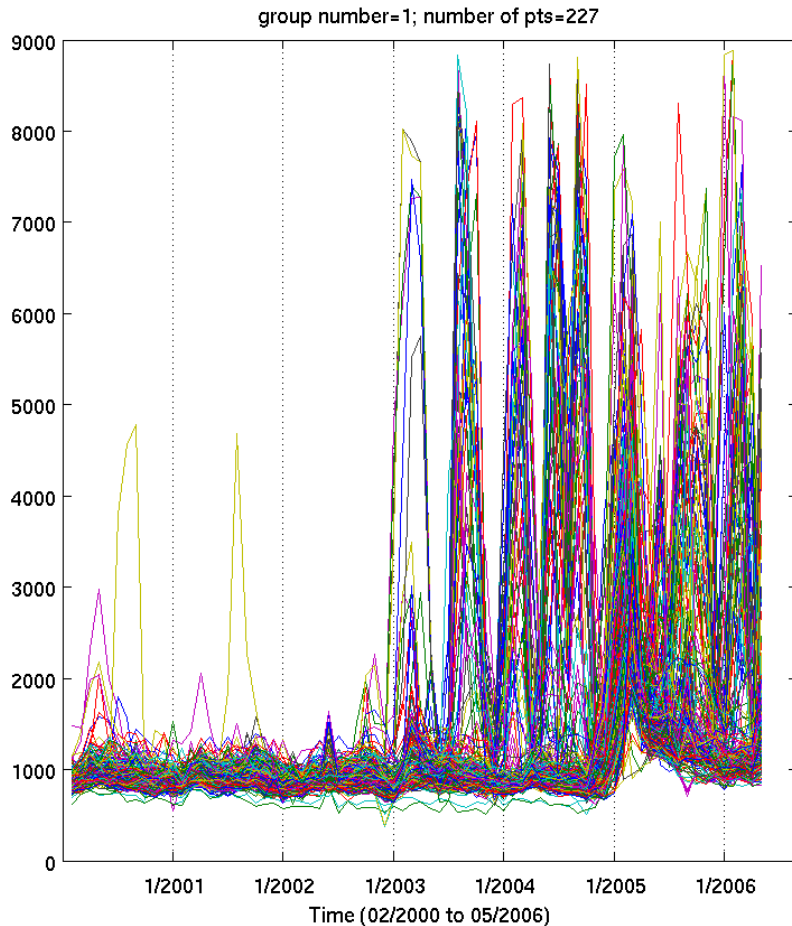


# Study Focus: Entire California

---

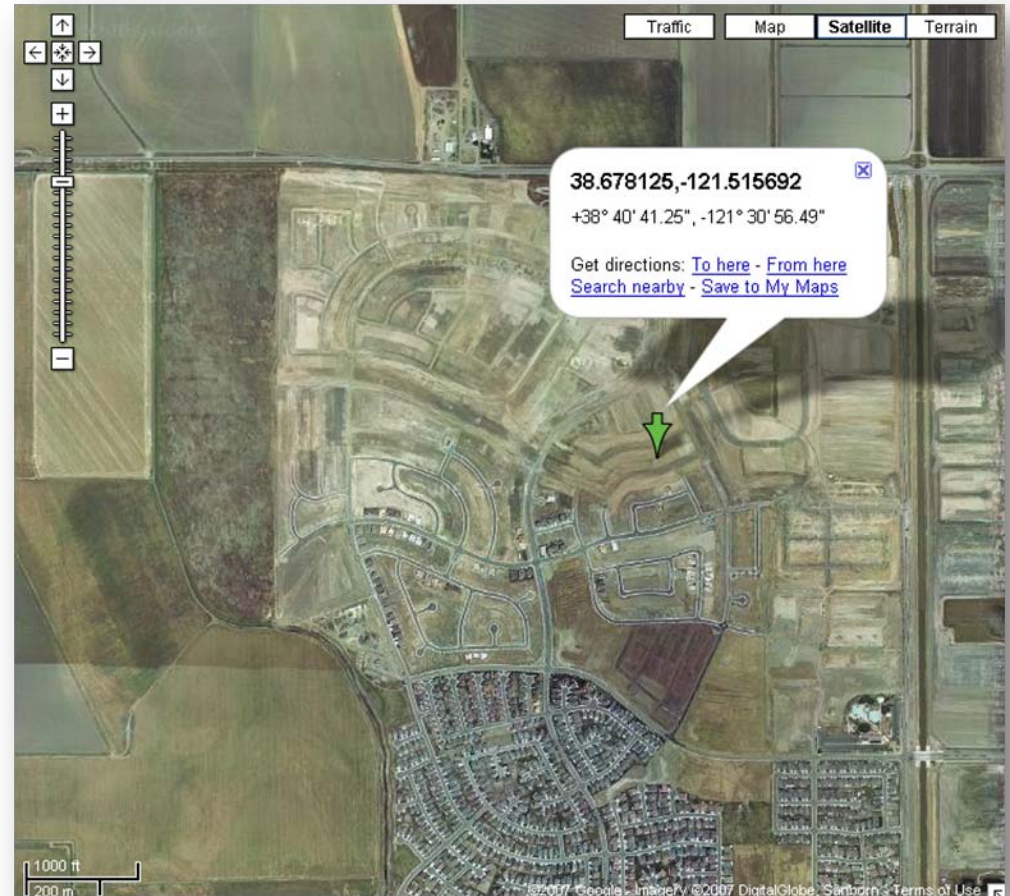
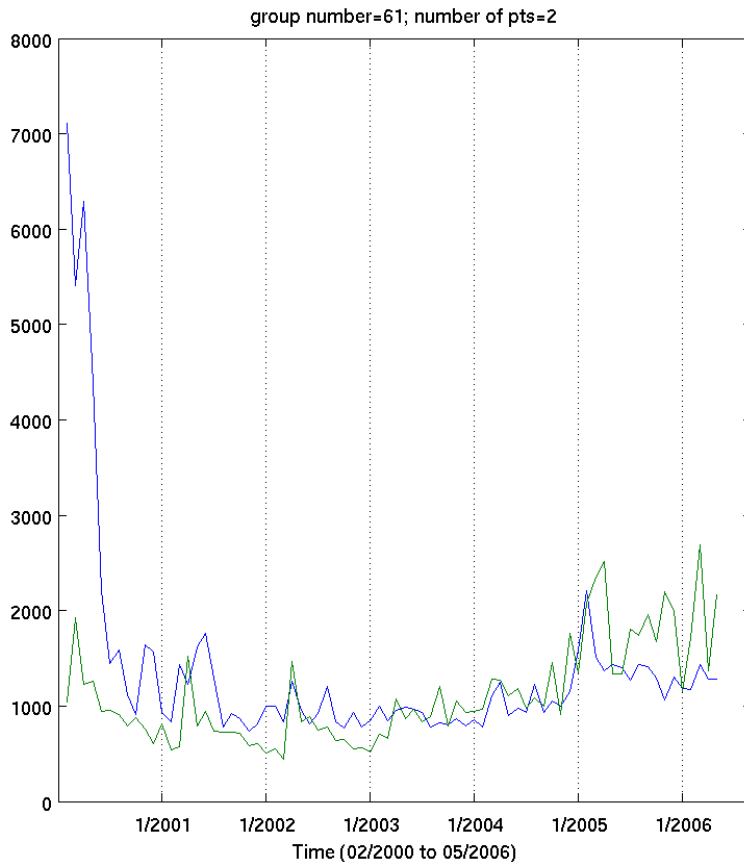
- Data has 5,165,205 locations
- After applying our algorithm, 2,833 locations with change points are detected at a high threshold
- The larger data has more types of changes:
  - Desert to farmland
  - Farmland to subdivision
  - Desert to golf course
  - Logging in tropical forest
  - Forest fires
  - ..
  - ..

# Example: Conversion to farmland



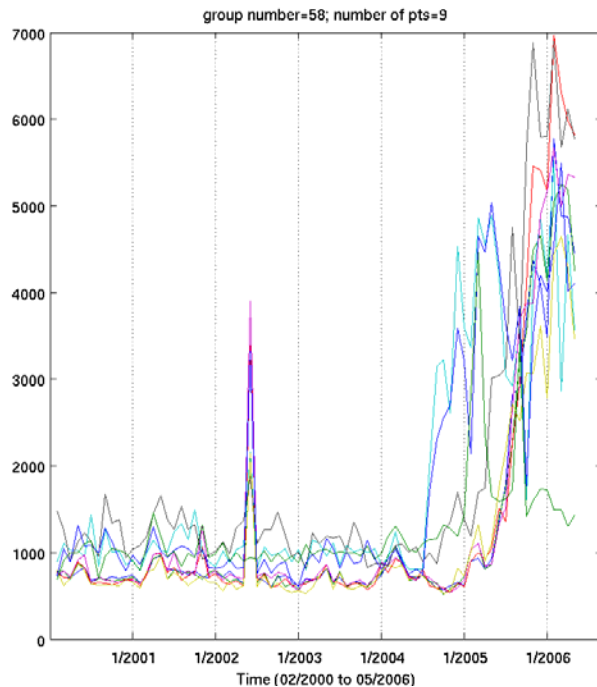
Group of pixels that were all detected by our algorithm, spatially located close to each other

# Example: Farmland to subdivision

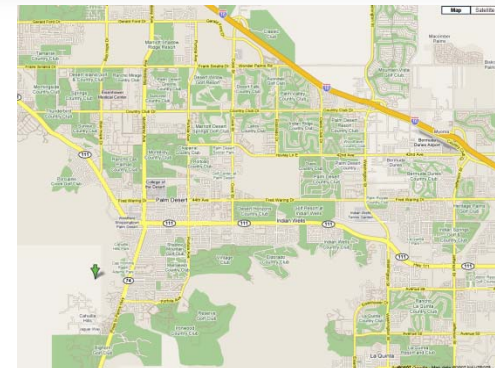


Location in Sacramento where farm land has been cleared and a subdivision is being built.

# Bunch of Golf Courses in SE California Desert



- New golf course being built in Palm Desert, CA
- This town has over 100 golf courses, putting intense pressure on the water supply

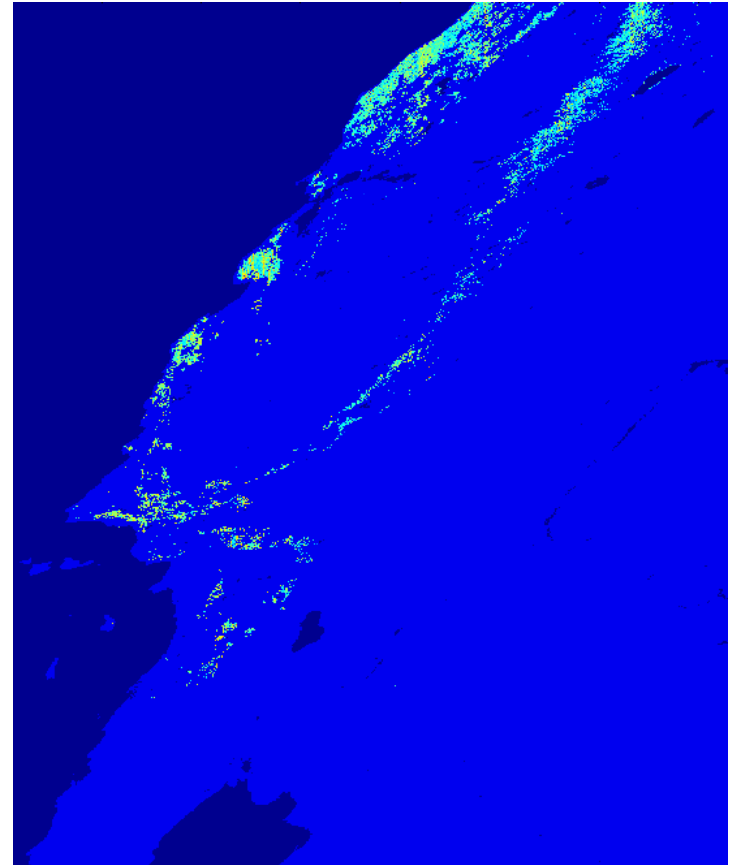




# Study Focus: Forests in California

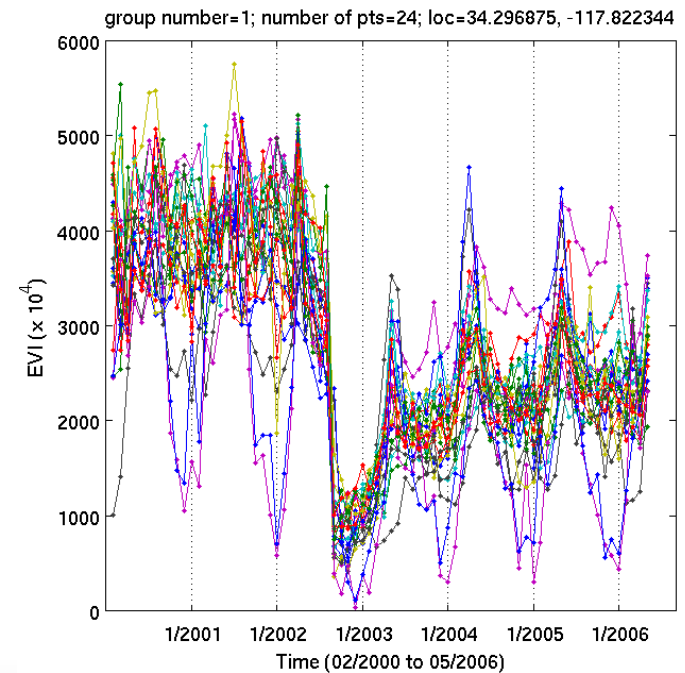
---

- 380,285 locations
- Covers the following land cover types:
  - Evergreen Needleleaf Forest
  - Evergreen Broadleaf Forest
  - Deciduous Needleleaf Forest
  - Deciduous Broadleaf Forest
  - Mixed Forests
- Majority of change in forests in CA due to forest fires
- Forest fires are easily verifiable using database maintained by the state dept of forestry and fire prevention.

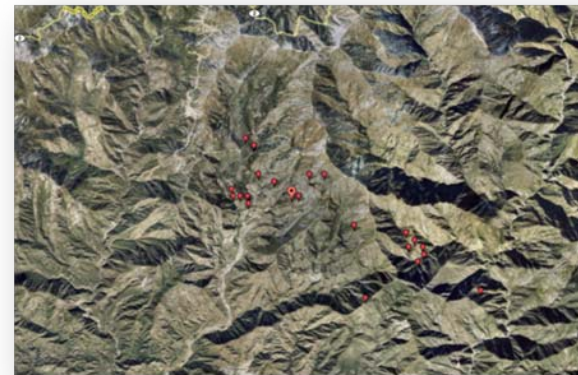


# Fires Detected by our Algorithm

Year	Forest Fire
September 2002	Curve
June 2002	Troy
June 2002	Wolf
July 2002	Pines
May 2004	Cachuma
Mid-2003	Spanish
Late 2003	Grand Prix
October 2004	Rumsey
Mid 2001	Poe
< 2000	Kirk Complex
September 2001	Darby
Mid-2004	Geysers



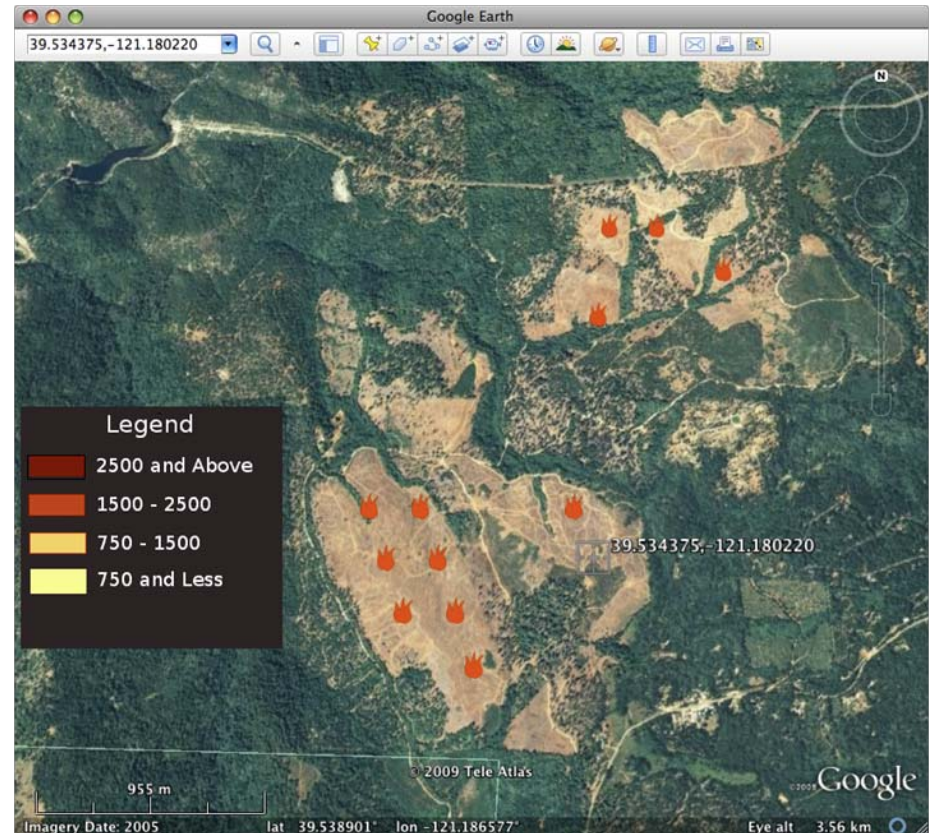
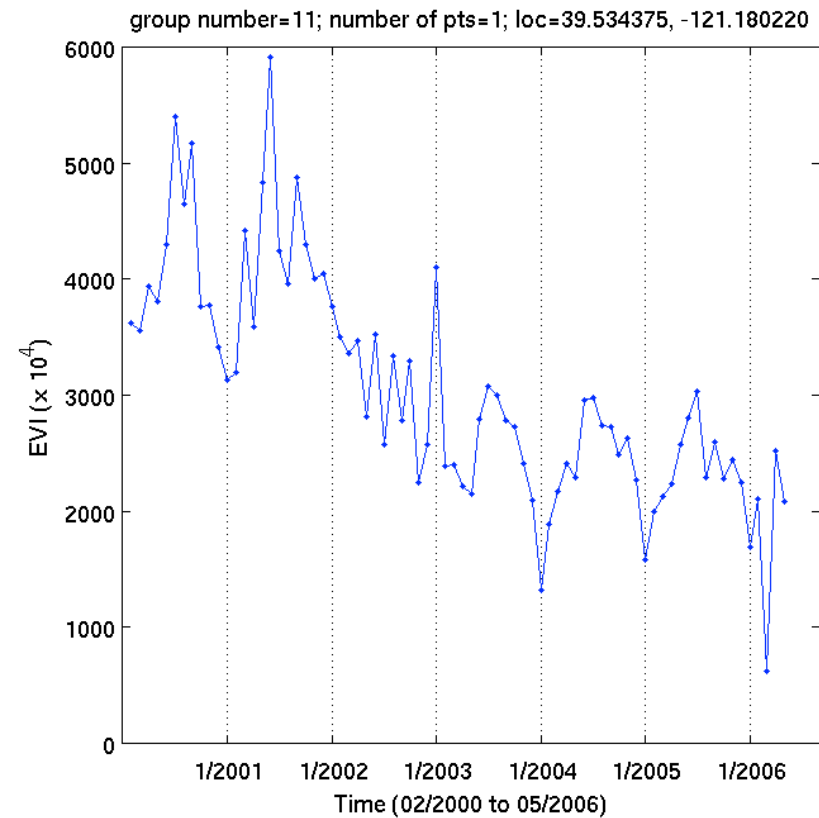
An informal evaluation **of the top 3800 points (1% of all locations)** showed an overwhelming majority appeared to correspond to forest fires.



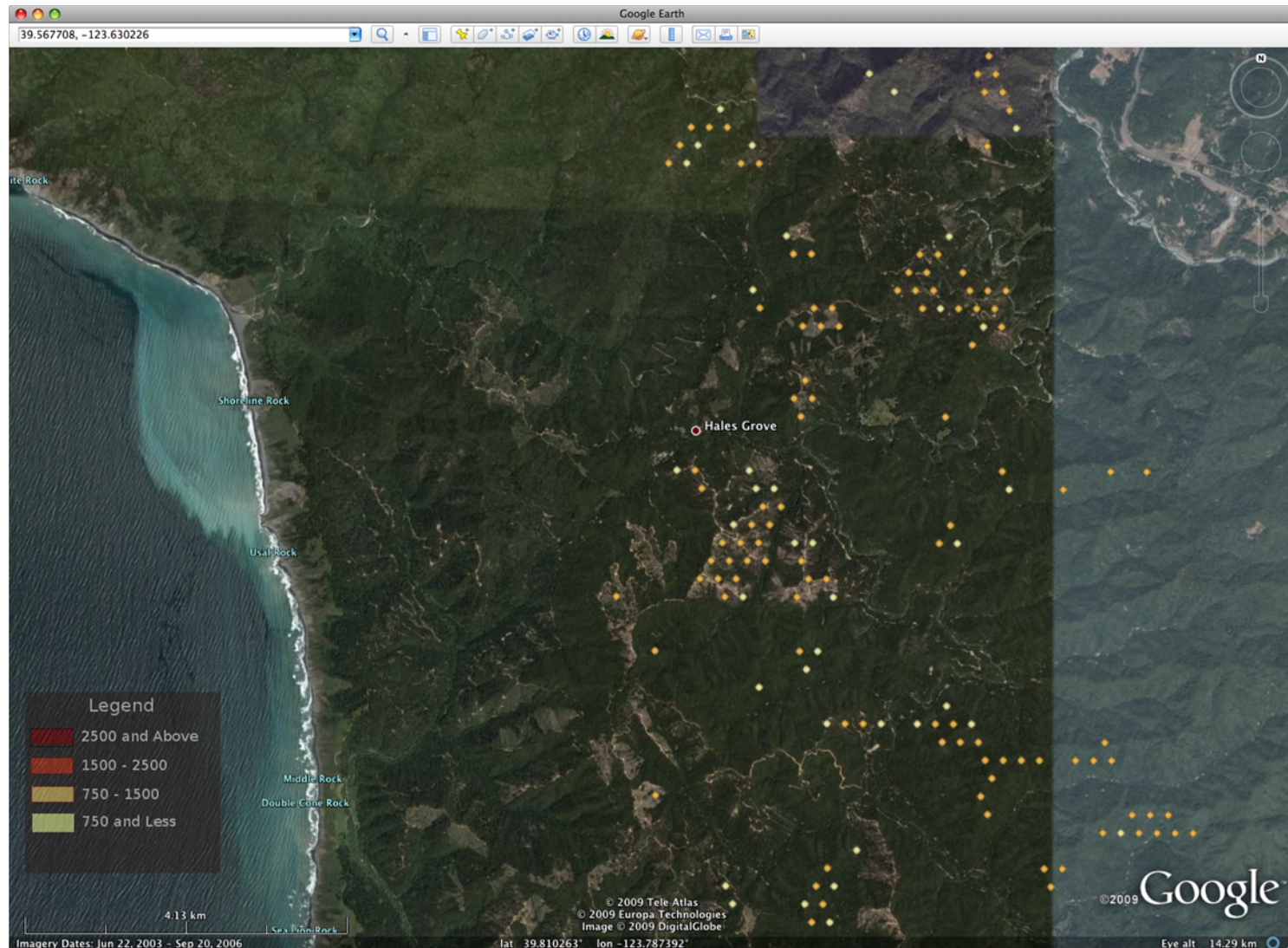
Curve Fire,  
September  
2002, San  
Gabriel  
Canyon, 20K  
acres.



# Logging in Northern CA



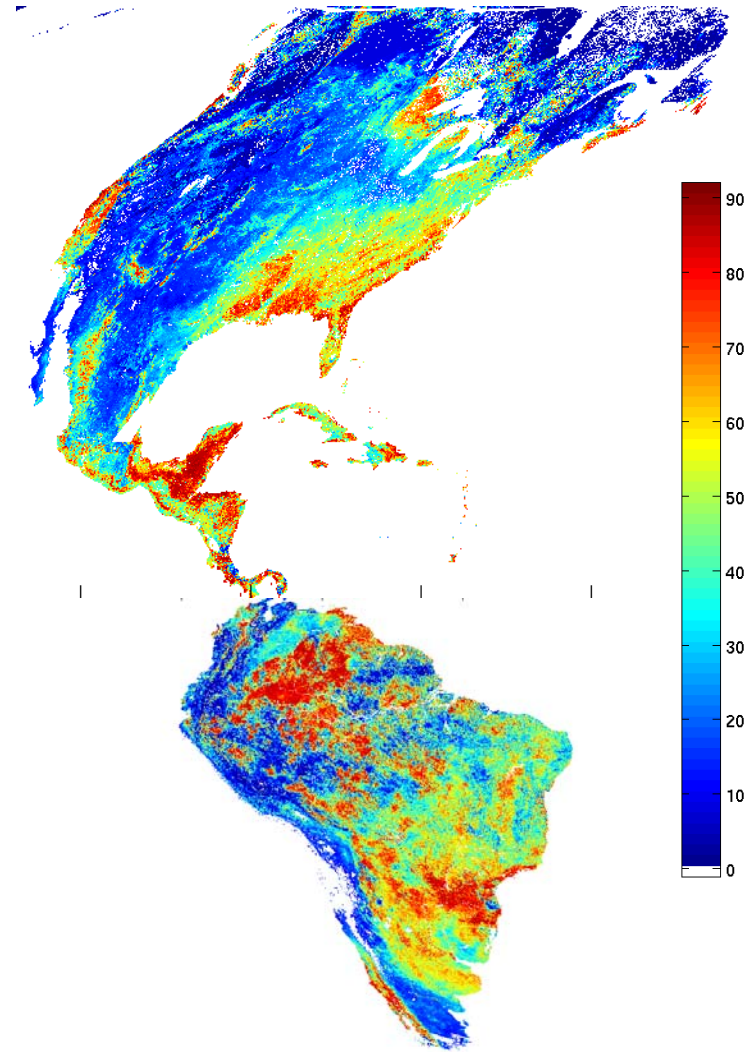
# Logging in Mendocino County, CA



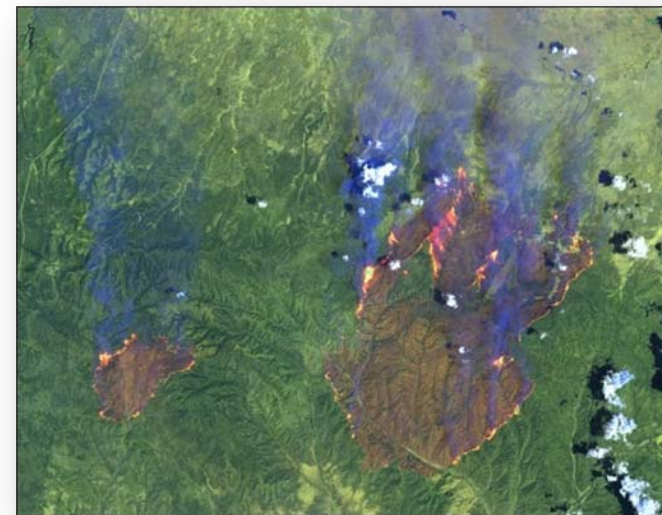
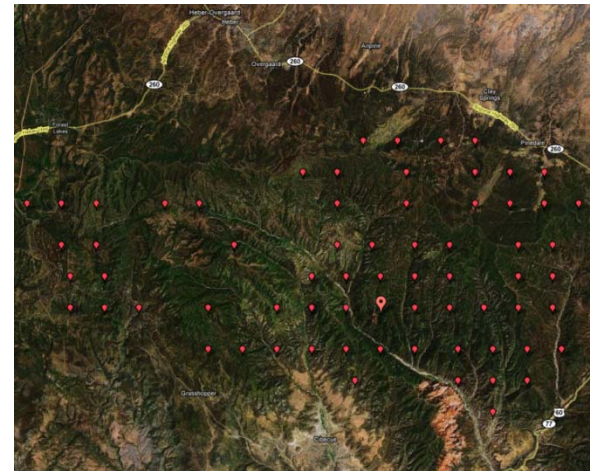
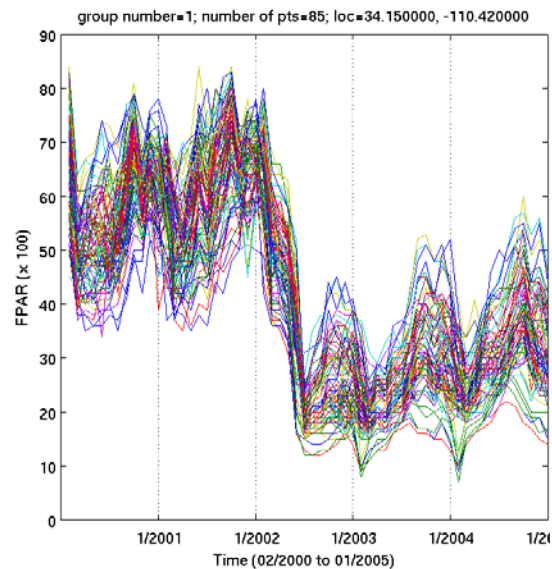


# Study Focus: Global FPAR Data

- FPAR is available globally at 4km resolution from February 2000—December 2006.
- Similar to EVI but not as sensitive in high biomass cover areas (higher values of index)



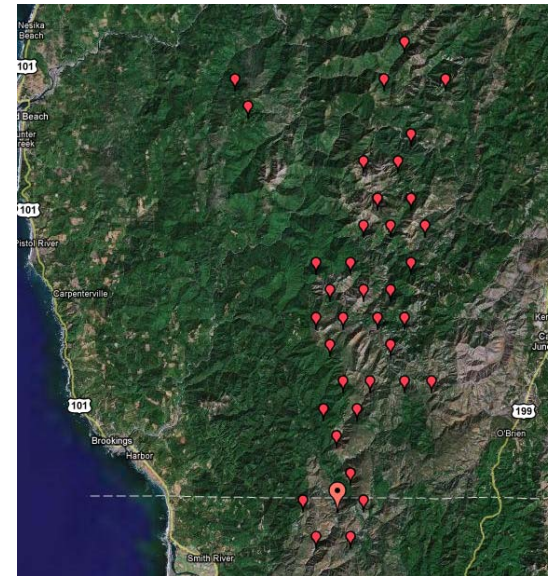
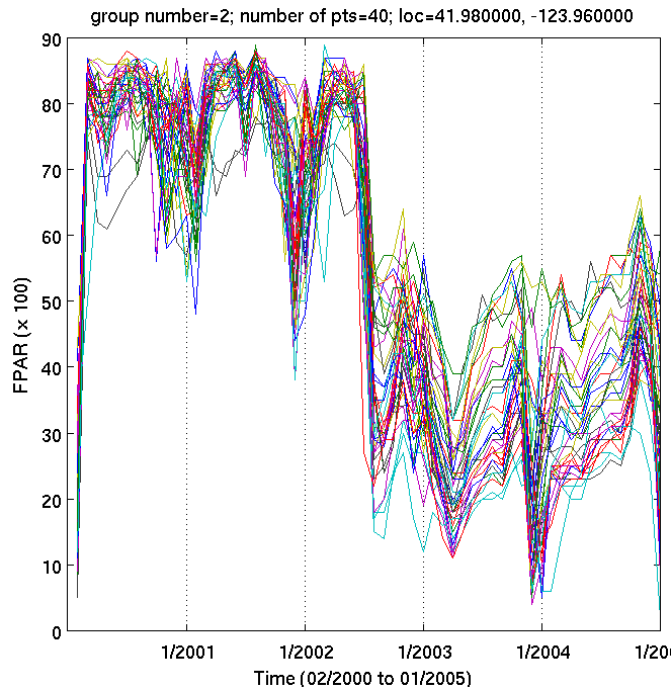
# FPAR Results: North America



Courtesy: NASA

- Large change in forested area near Phoenix
- Corresponds to Rodeo fire in June 2002

# FPAR Results: North America



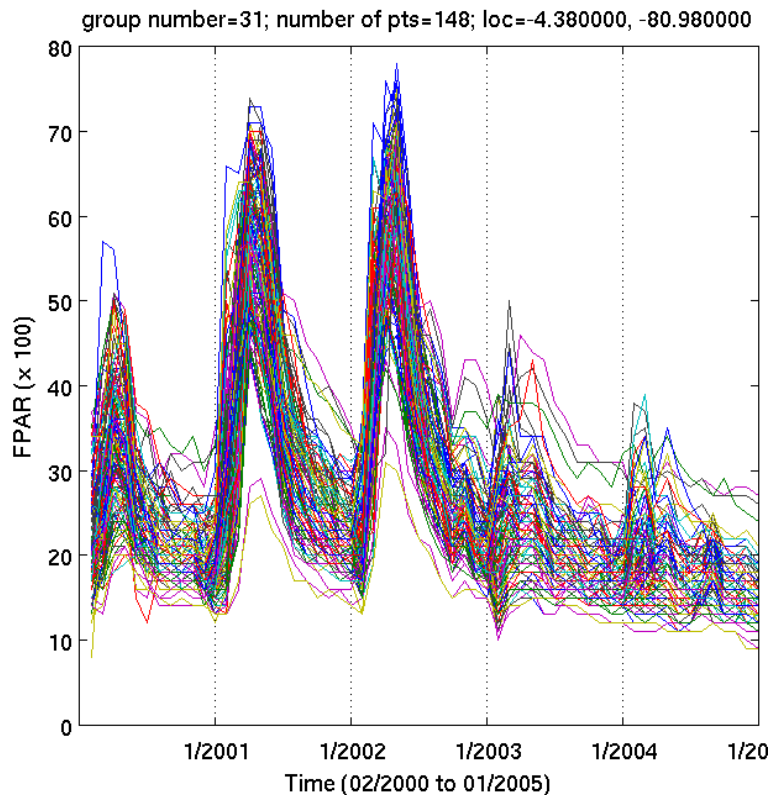
- Large region along California/Oregon border
- Corresponds to a Biscuit Fire in July 2002 (500,000 acres burned).



Courtesy: NASA



# FPAR Results: South America

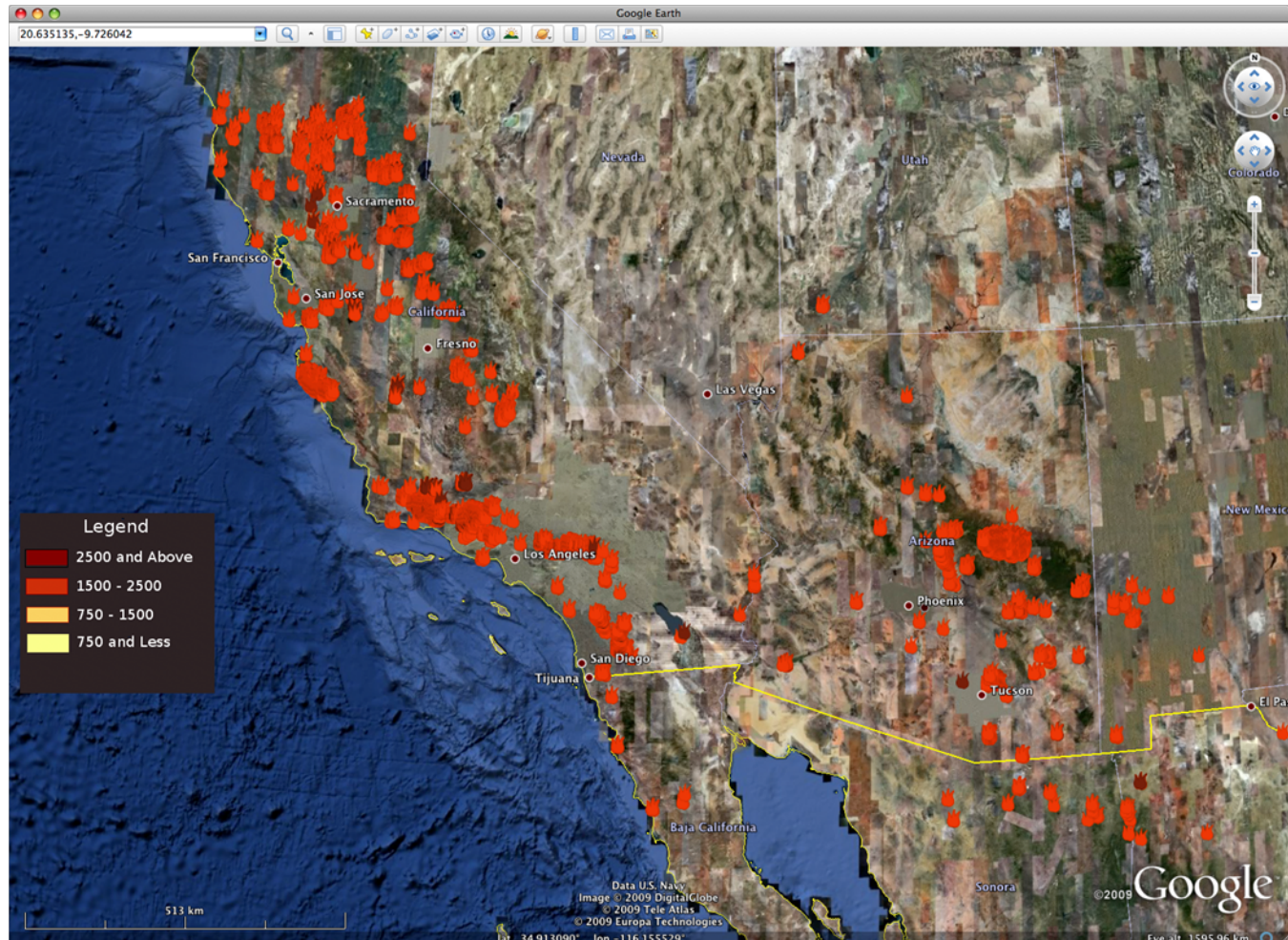


Large event spanning hundreds of miles - possibly due to El-Nino related drought.



# History of fires in CA/NE/AZ 2000-2008

20,000 pixels detected over a 10-year period.



# Full Animation





# Applications being studied

- **Global forest tracking**

- carbon pool mapping
- deforestation monitoring

- **Agricultural data**

- Tracking conversion of soybean fields and fallow land to corn for biofuels
- Estimating annual yield of grain

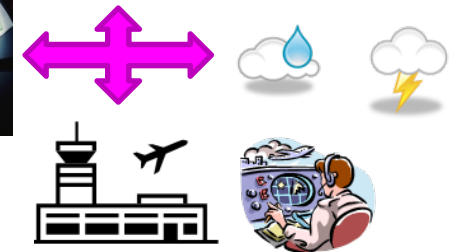
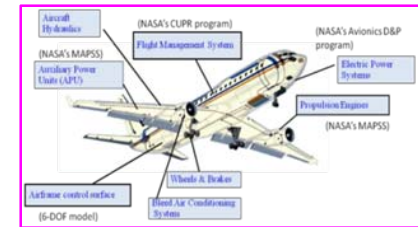


Source: Climate Central.



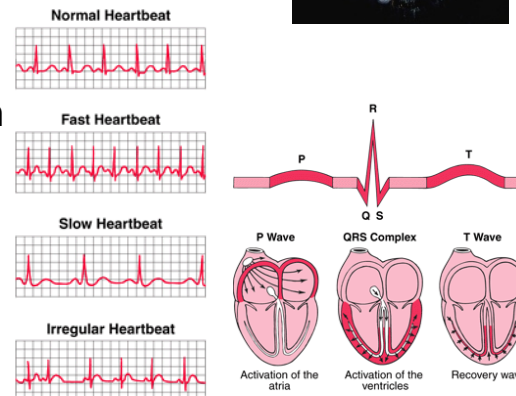
- **Vehicle Health Management**

- Monitoring the health of complex systems
- Aircraft sensors
- Flight Text Logs



- **General scenarios where sensors are monitoring time-varying phenomena**

- Sensors for monitoring human health (cardiac pacemaker, etc.)
- Manufacturing processes
- Meteorological data
- Scientific experiments

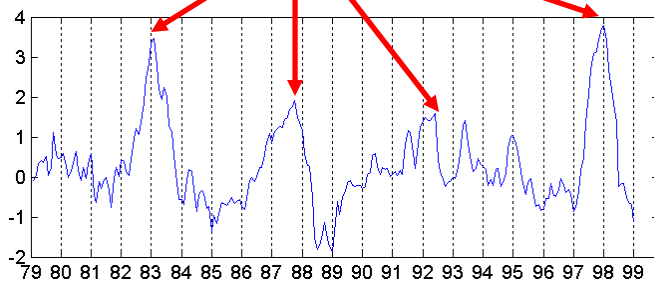


Source: Merck.

# Climate Indices: Connecting the Ocean/Atmosphere and the Land

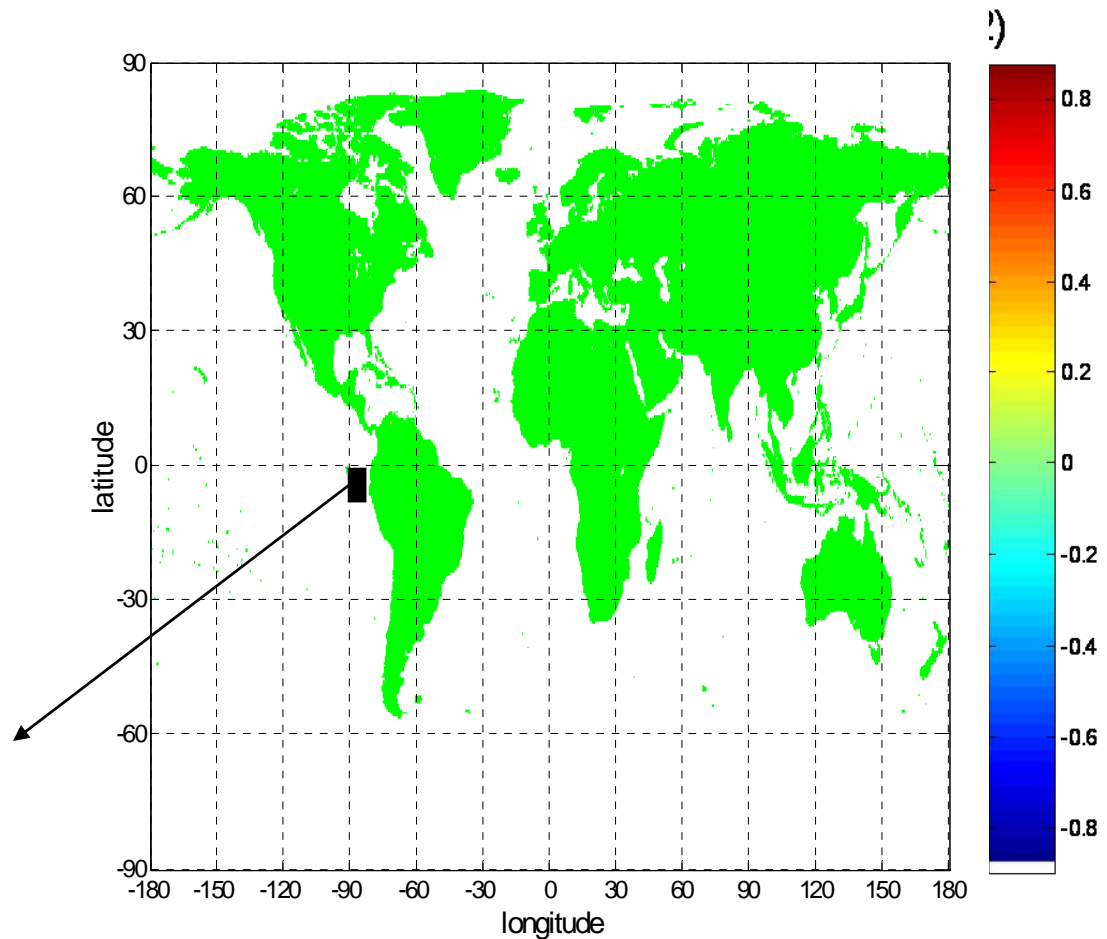
- A climate index is a time series of sea surface temperature or sea level pressure
- Climate indices capture teleconnections
  - The simultaneous variation in climate and related processes over widely separated points on the Earth

## El Nino Events



## Nino 1+2 Index

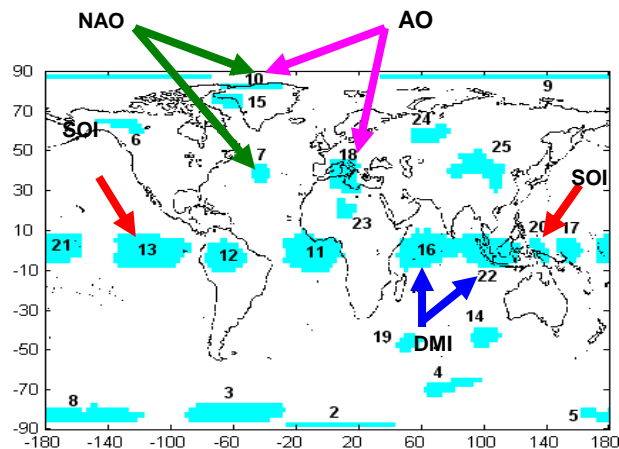
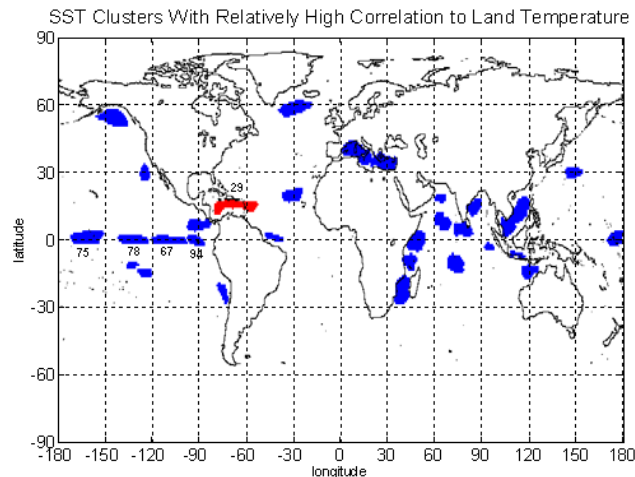
Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S



# List of Well Known Climate Indices

Index	Description
SOI	<b>Southern Oscillation Index:</b> Measures the SLP anomalies between Darwin and Tahiti
NAO	<b>North Atlantic Oscillation:</b> Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	<b>Arctic Oscillation:</b> Defined as the _first principal component of SLP poleward of 20° N
PDO	<b>Pacific Decadal Oscillation:</b> Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	<b>Quasi-Biennial Oscillation Index:</b> Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	<b>Cold Tongue Index:</b> Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	<b>Western Pacific:</b> 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
<b>NINO1+2</b>	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
<b>NINO3</b>	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
<b>NINO3.4</b>	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
<b>NINO4</b>	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

# Discovery of Climate Indices Using Clustering



- 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.
- Clusters are found using the Shared Nearest Neighbor (SNN) method that eliminates “noise” points and tends to find regions of “uniform density”.
- Clusters are filtered to eliminate those with low impact on land points

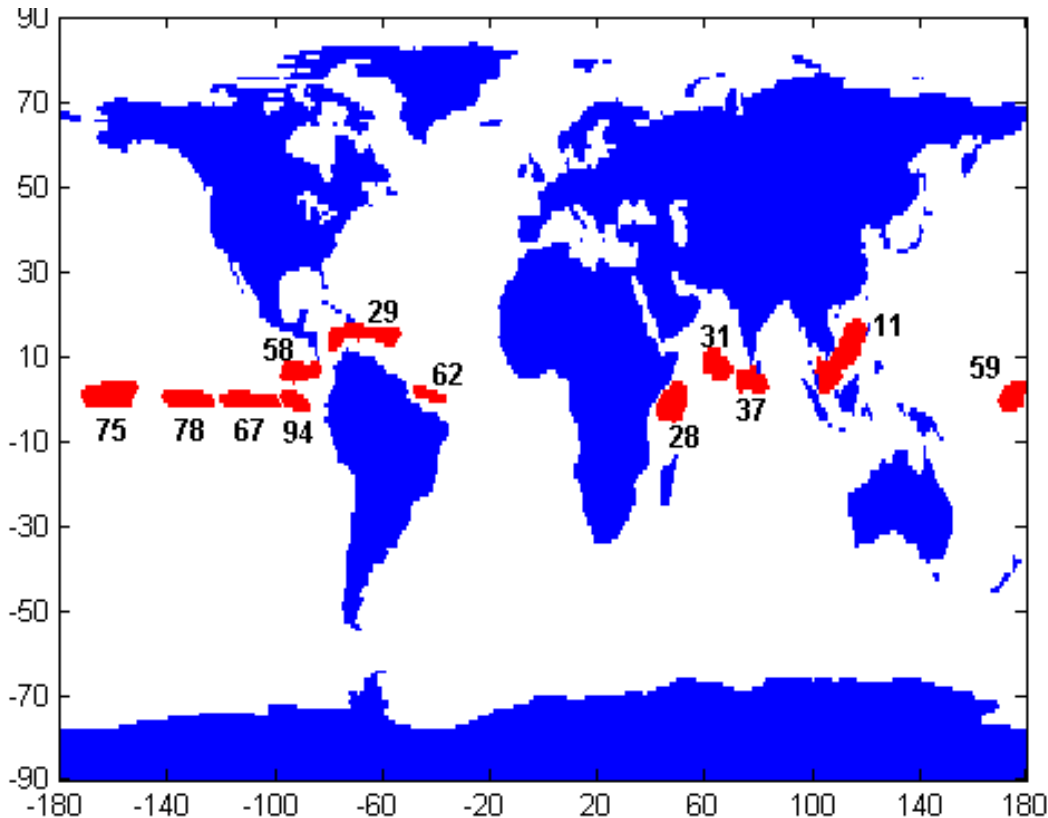
**Result:** A cluster-based approach for discovering climate indices provides better physical interpretation than those based on the SVD/EOF paradigm, and provide candidate indices with better predictive power than known indices for some land areas.



# SST Clusters that Reproduce Known Indices

# grid points: 67K Land, 40K Ocean      Current data size range: 20 – 400 MB

Monthly data over a range of 17 to 50 years



Cluster	Nino Index	Correlation
94	NINO 1+2	0.9225
67	NINO 3	0.9462
78	NINO 3.4	0.9196
75	NINO 4	0.9165

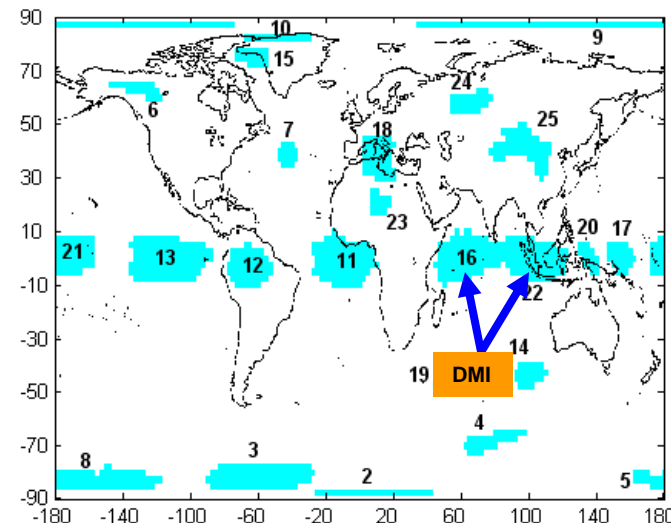
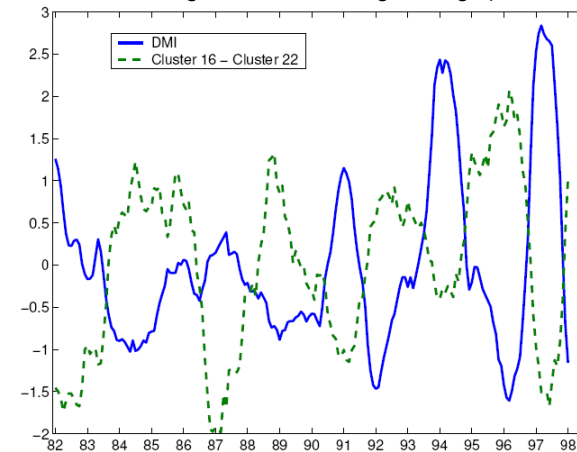
Some SST clusters reproduce well-known climate indices for El Niño.

Clusters of SST that have high impact on land temperature

# Finding New Patterns: Indian Monsoon Dipole Mode Index

- Recently a new index, the Indian Ocean Dipole Mode index (DMI), has been discovered\*.
- DMI is defined as the difference in SST anomaly between the region 5S-5N, 55E-75E and the region 0-10S, 85E-95E.
- DMI and is an indicator of a weak monsoon over the Indian subcontinent and heavy rainfall over East Africa.
- We can reproduce this index as a difference of pressure indices of clusters 16 and 22.

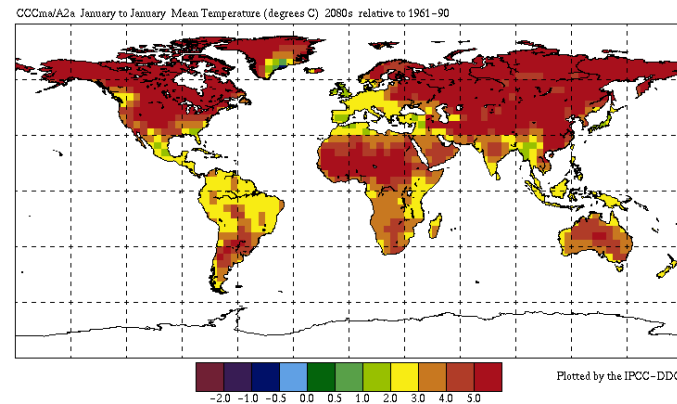
Plot of cluster 16 – cluster 22 versus the Indian Ocean Dipole Mode index. (Indices smoothed using 12 month moving average.)



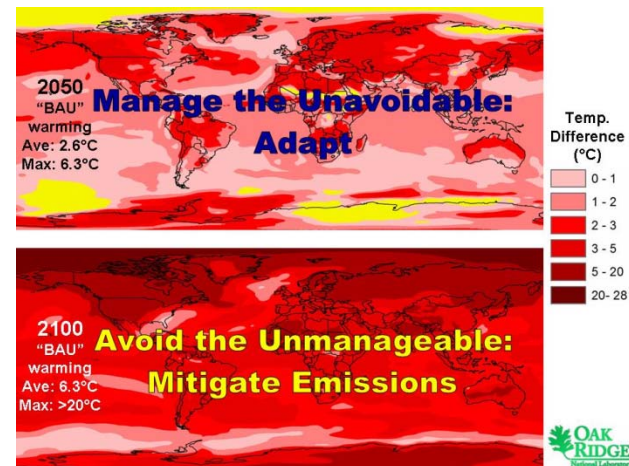
\* N. H. Saji, B. N. Goswami, P. N. Vinayachandran and T. Yamagata, "A dipole mode in the tropical Indian Ocean," Nature 401, 360-363 (23 September 1999).

# Applications of Climate Indices: Planning for Climate Change and Extreme Events

- Extract climate indices and features for extreme events from past observations.
- Develop predictive capabilities for extreme events using these features
- Generate climate forecasts using climate indices and Global Circulation Models (GCMs)

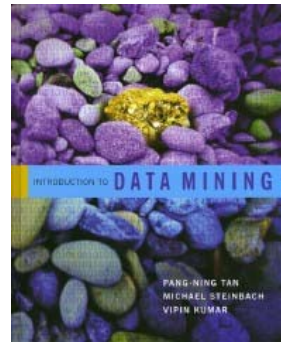
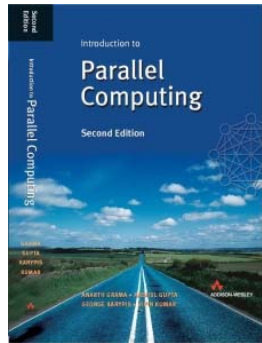


GCM Prediction for Mean Temperature in 2080s relative to 1961-90



Collaborators: Auroop Ganguly (ORNL), Fredrick Semazzi (NC State), Abdollah Homaifar (North Carolina A & T)

# Bibliography ([www.cs.umn.edu/~kumar](http://www.cs.umn.edu/~kumar))

- Introduction to Data Mining  
Pang-Ning Tan, Michael Steinbach, Vipin Kumar  
Addison-Wesley April 2006
  - Introduction to Parallel Computing (2nd Edition)  
A. Grama, A. Gupta, G. Karypis, and Vipin Kumar  
Addison-Wesley, 2003
- 
- 
- S. Boriah, V. Kumar, M. Steinbach, C. Potter, and S. Klooster  
Land Cover Change Detection: A Case Study  
*Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.
  - S. Boriah, V. Kumar, M. Steinbach, P. Tan, C. Potter, and S. Klooster  
Detecting Ecosystem Disturbances and Land Cover Change using Data Mining.  
In H. Kargupta, J. Han, P. Yu, R. Motwani, and V. Kumar, editors, *Next Generation of Data Mining*, CRC Press, 2008.
  - C. Potter, V. Genovese, P. Gross, S. Boriah, M. Steinbach, and V. Kumar  
Revealing Land Cover Change in California With Satellite Data  
*Eos Trans. AGU*, 88(26):269,2007.
  - M. Steinbach, P. Tan, V. Kumar, C. Potter and S. Klooster  
Discovery of Climate Indices Using Clustering, *Proceedings of KDD 2003*.
  - S. Boriah and V. Kumar, Time Series Change Detection: A Survey, *Technical Report*.