Mining climate and ecosystem data: challenges, opportunities, and some early results.

Vipin Kumar

University of Minnesota

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

Collaborators and Group Members:

Shyam Boriah, Sudipto Banerjee, Marc Dunham, Ashish Garg, Joe Knight, Varun Mithal, Michael Steinbach
University of Minnesota

Pang-Ning Tan
Michigan State University

Chris Potter
NASA Ames

Steve Klooster
California State University

Fred Semazzi, Nagiza Samatova
NC State

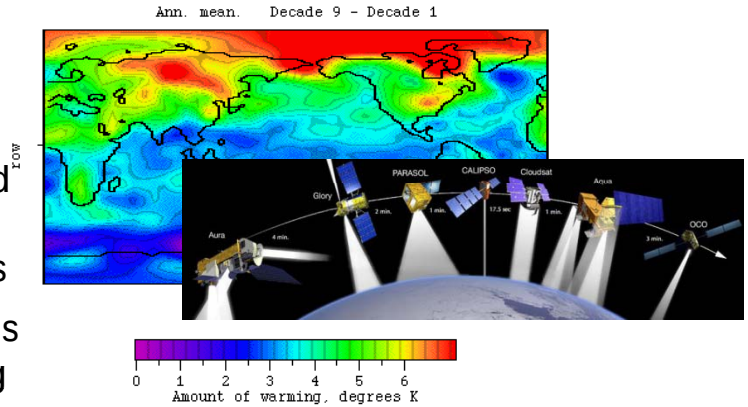
Auroop Ganguly
ORNL

Research funded by ISET-NOAA, NSF, NASA, Planetary Skin Institute, MN Futures program

© Vipin Kumar

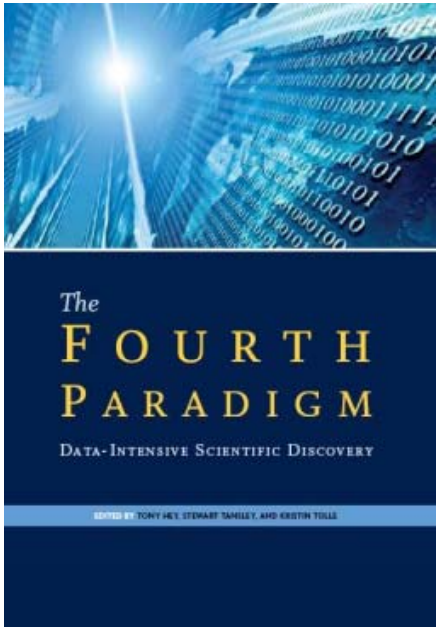
Data-Driven Knowledge Discovery in Climate and Ecosystem Sciences

- **Transformation from data-poor to data-rich domain**
  - **Sensor Observations**: Remote sensors like satellites and weather radars as well as in-situ sensors and sensor networks like weather station and radiation measurements
  - **Model Simulations**: IPCC climate or earth system models as well as regional models of climate and hydrology, along with observed data based model reconstructions

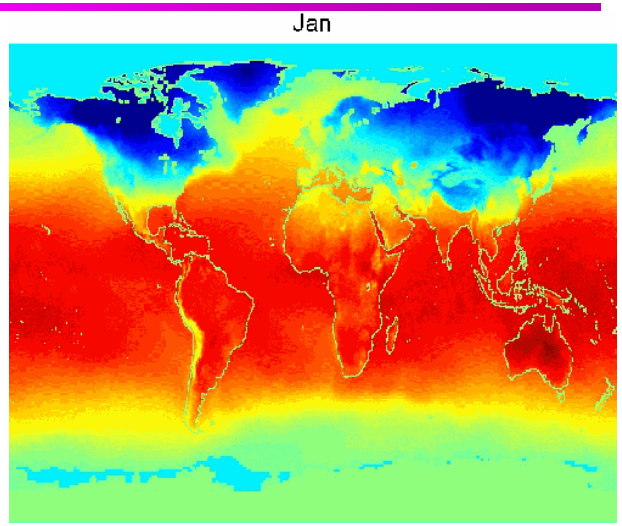


- Data guided discovery can complement hypothesis guided data analysis to develop predictive insights for use by climate scientists, policy makers and community at large.

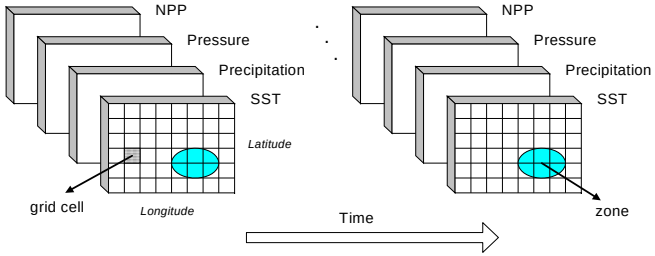
"***The world of science has changed … data-intensive science [is] so different that it is worth distinguishing [it] … as a new, fourth paradigm for scientific exploration.***" - **Jim Gray**



Challenges in Mining Climate and Ecosytem Data

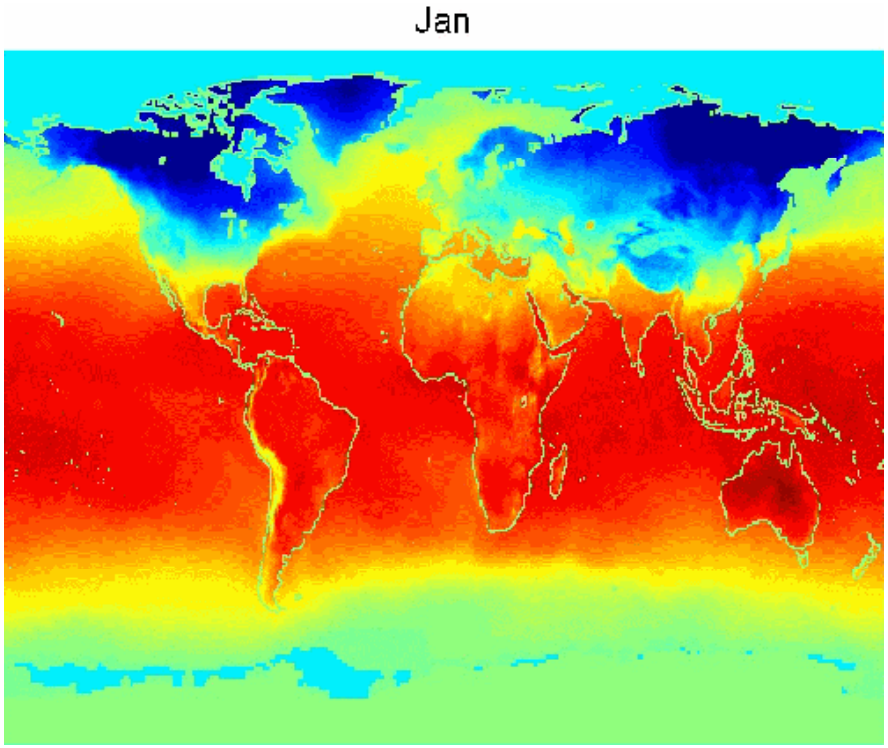


- **Spatio-temporal nature of data**
  - spatial and temporal autocorrelation.
  - Multi-scale/Multi-resolution nature
- **Scalability**
  - Size of Earth Science data sets can be very large, For example, for each time instance,
    - 2.5° x 2.5° :10K locations for the globe
    - 250m x 250m: ~10 billion
    - 50m x 50m : ~250 billion
- **High-dimensionality**
- **Noise and missing values**
- **Long-range spatial dependence**
- **Long memory temporal processes**
- **Nonlinear processes, Non-Stationarity**
- **Fusing multiple sources of data**



Challenges in Analyzing Eco-Climate Data

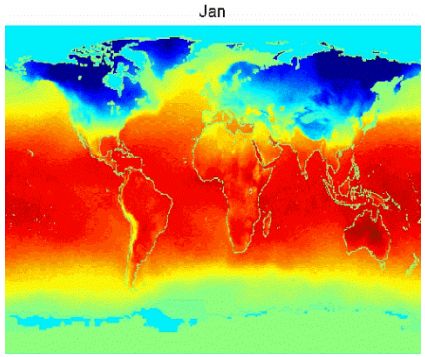
Global Sea Surface Temperature

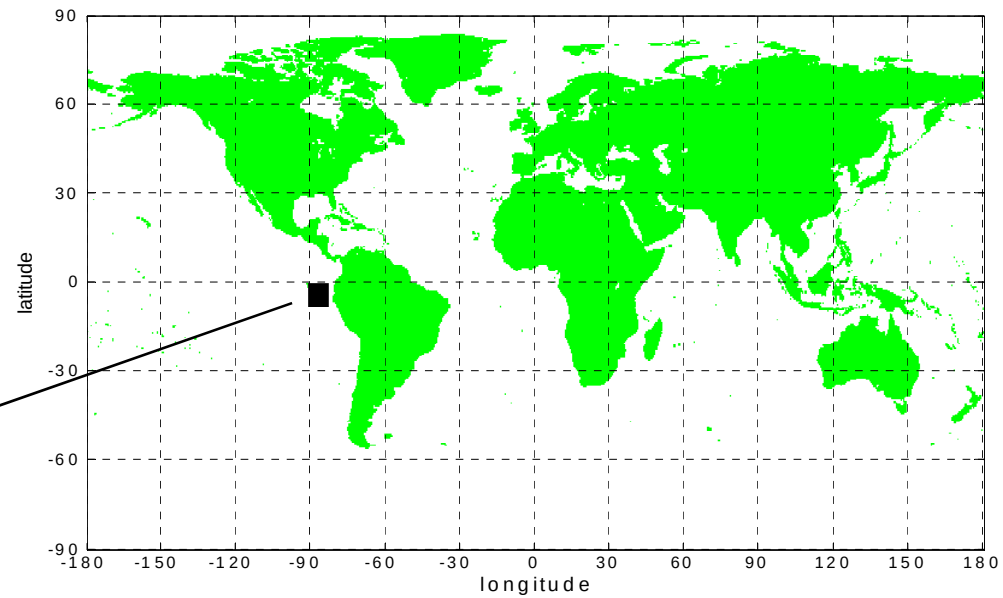


Challenges in Analyzing Eco-Climate Data

Global Sea Surface Temperature

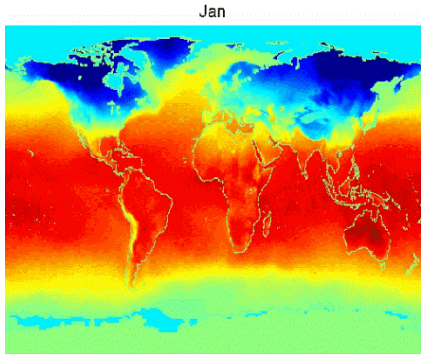
Discovering teleconnections



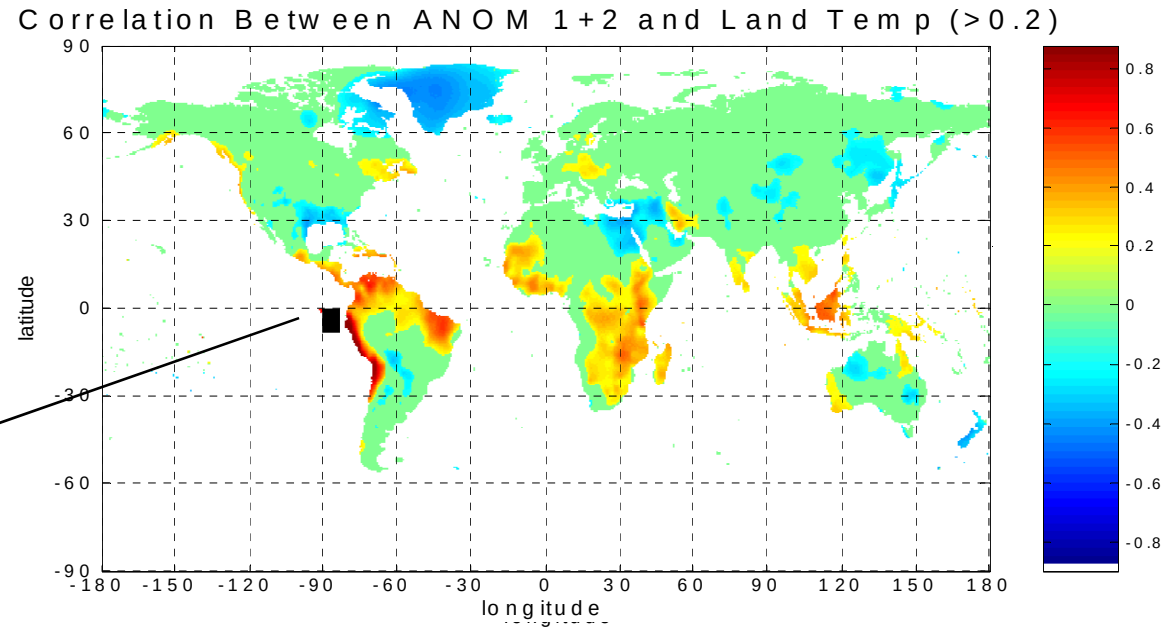


Challenges in Analyzing Eco-Climate Data

Global Sea Surface Temperature

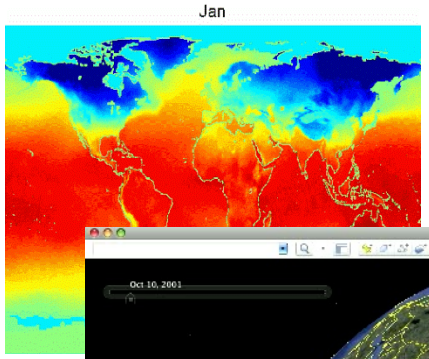


Discovering teleconnections

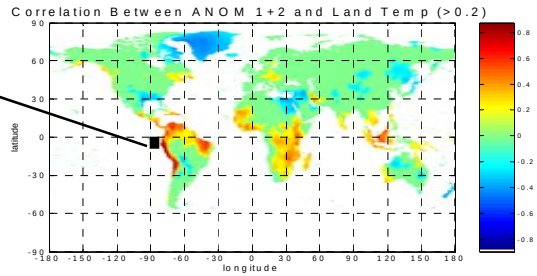


Challenges in Analyzing Eco-Climate Data

Global Sea Surface Temperature



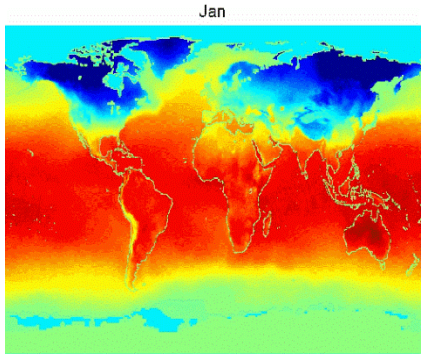
Discovering teleconnections



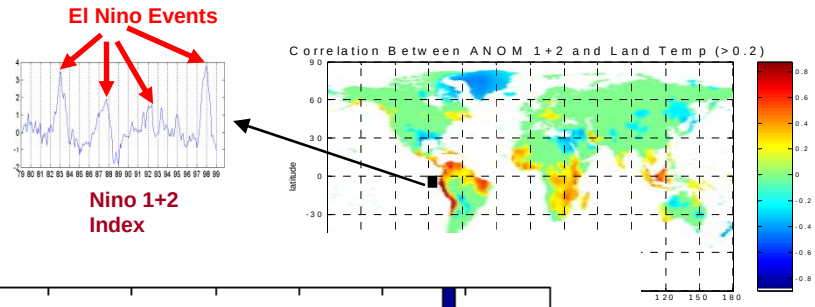
Changes in Global Forest Cover

Challenges in Analyzing Eco-Climate Data

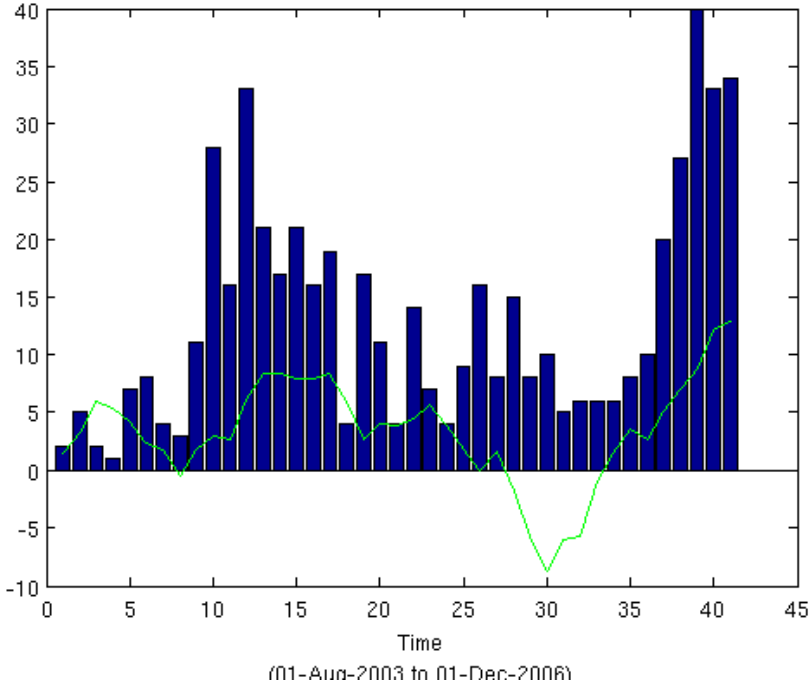
Global Sea Surface Temperature



Discovering teleconnections



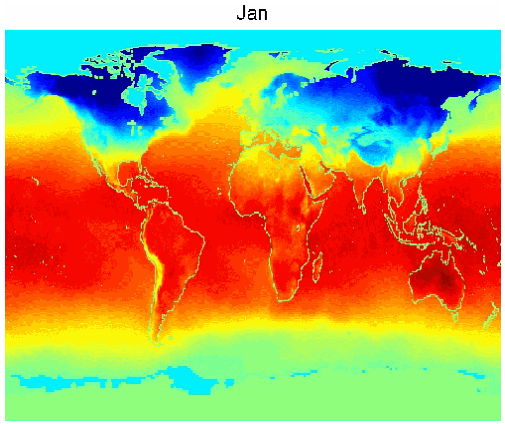
Changes in Global Forest Cover



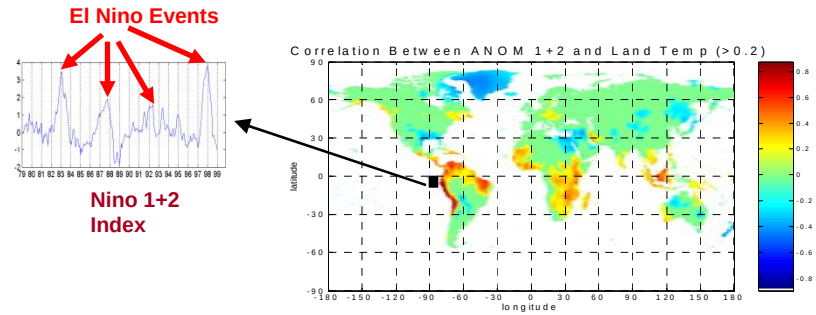
Relationship between El Nino and Fires in Indonesia

Challenges in Analyzing Eco-Climate Data

Global Sea Surface Temperature



Discovering teleconnections

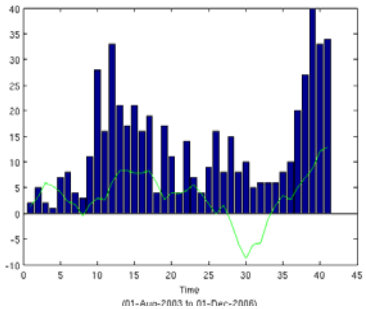


Challenges due to data characteristics

- Spatiotemporal, non-stationary, non-i.i.d.
- Massive data sets

  2.5°x 2.5°:10K locations for the globe
  250m x 250m: ~10 billion
  50m x 50m : ~250 billion
- Long range spatial dependencies
- Long memory temporal dependencies
- Nonlinear multiscale dependencies
- Low frequency variability
- ...

Changes in Global Forest Cover



Relationship between El Nino and Fires in Indonesia

Illustrative Applications of Data Mining

- Monitoring of global forest cover
- Discovering teleconnections among climate variables
- Predicting the impacts of climate change

Monitoring Forest Cover Change: Motivation

- Changes in forests account for over **20%** of the greenhouse gas emissions
  - 2nd 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 CO2.

SFO San Francisco International Airport

Good to Go Green: SFO Unveils Carbon Offset Kiosks

npr

'Carbon Offset' Business Takes Root
by Martin Kaste

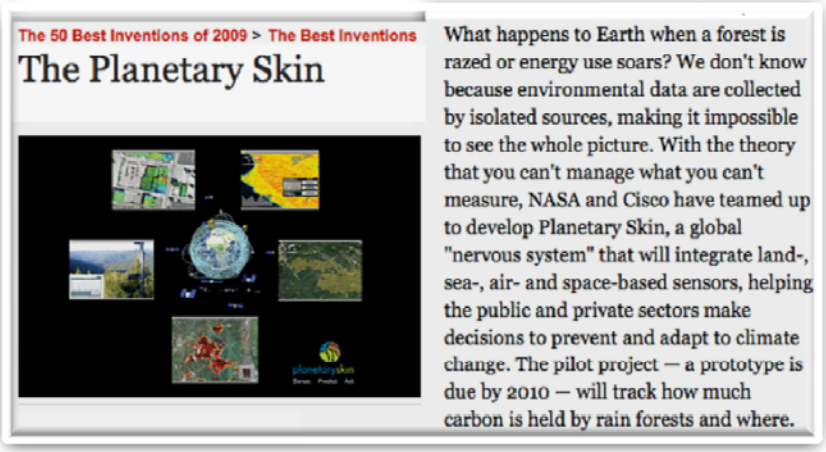
Monitoring of Global Forest Cover

- Planetary Information System for assessment of ecosystem disturbances
  - Forest fires
  - Droughts
  - Floods
  - Logging/deforestation
  - Conversion to agriculture

- This system will help
  - quantify the carbon impact of these changes
  - Understand the relationship to global climate variability and human activity

- Provide **ubiquitous web-based access** to changes occurring across the globe, creating public awareness

TIME The 50 Best Inventions of 2009



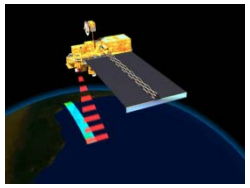
The 50 Best Inventions of 2009 > The Best Inventions

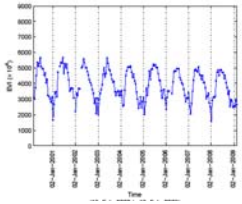
The Planetary Skin

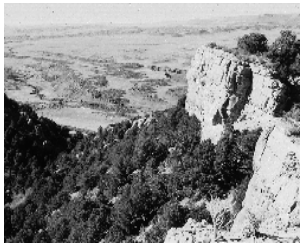
What happens to Earth when a forest is razed or energy use soars? We don't know because environmental data are collected by isolated sources, making it impossible to see the whole picture. With the theory that you can't manage what you can't measure, NASA and Cisco have teamed up to develop Planetary Skin, a global "nervous system" that will integrate land-, sea-, air- and space-based sensors, helping the public and private sectors make decisions to prevent and adapt to climate change. The pilot project — a prototype is due by 2010 — will track how much carbon is held by rain forests and where.

Novel Algorithms for Monitoring Global Eco-system

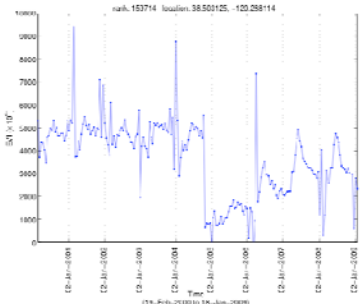
- State of the art algorithm for land cover change detection do not scale
- Alternate Approach: Use remote sensing vegetation data

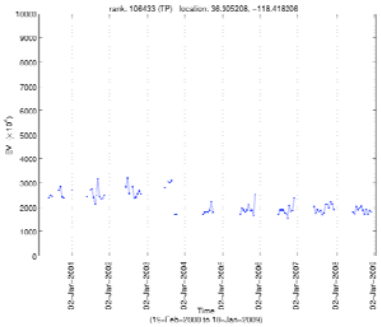






- Existing Time series change detection algorithms do not address unique characteristics of eco-system data
  - noise, missing values, outliers, high degree of variability across regions, vegetation types, and time.
- We have developed new algorithms that build non-parametric models of different segments of the time series and use them to capture the degree of change
- Data sets used in our study :
  - EVI: Enhanced Vegetation Index (250m x 250m)
  - FPAR : Fraction of Photosynthetically Active Radiation (1km and 4 km)



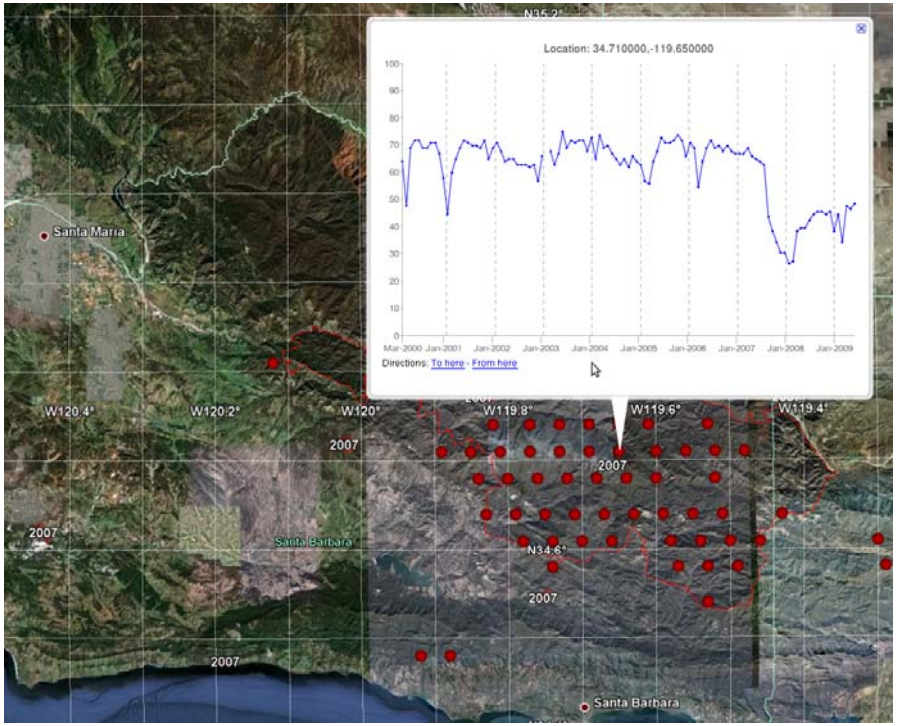


S. Boriah, V. Kumar, M. Steinbach, et al., *Land cover change detection: a case study*, *KDD* 2008.
V. Mithal, S. Boriah, A. Garg, M. Steinbach, V. Kumar et al., *Monitoring global forest cover using data mining. Technical Report Department of Computer Science, University of Minnesota* 2010.

Case Study 1:

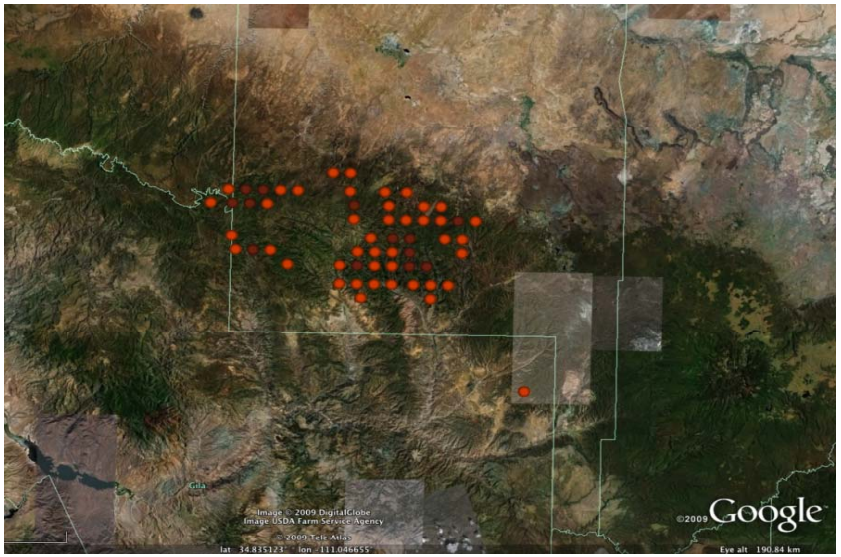
Monitoring Global Forest Cover

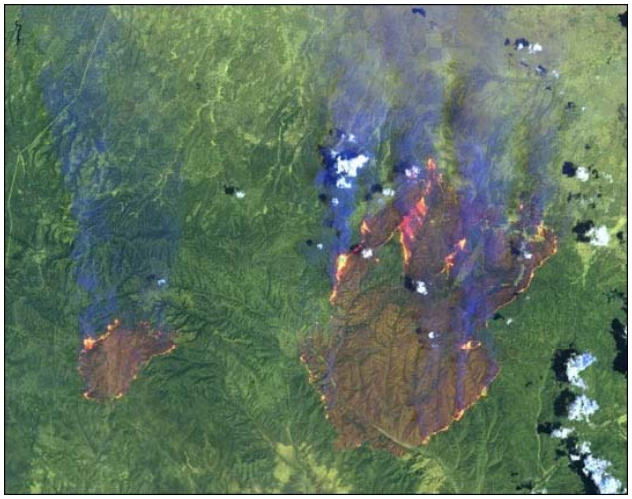
California Fires: 2007 Santa Barbara Fire

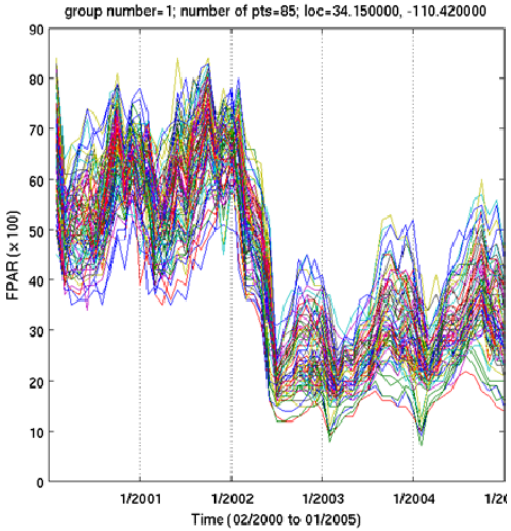


- Fire detected is the well documented Zaca Fire. It began burning about 15 miles northeast of Buellton, California. The fire started on July 4, 2007 and by August 31, it had burned over 240,207 acres (972.083 km^2), making it California's second largest fire and Santa Barbara's county largest fire.
- The fire was human induced and started as a result of sparks from a grinding machine on private property which was being used to repair a water pipe. The fire cost $118.3 million to fight and involved 21 fire crews.

Arizona





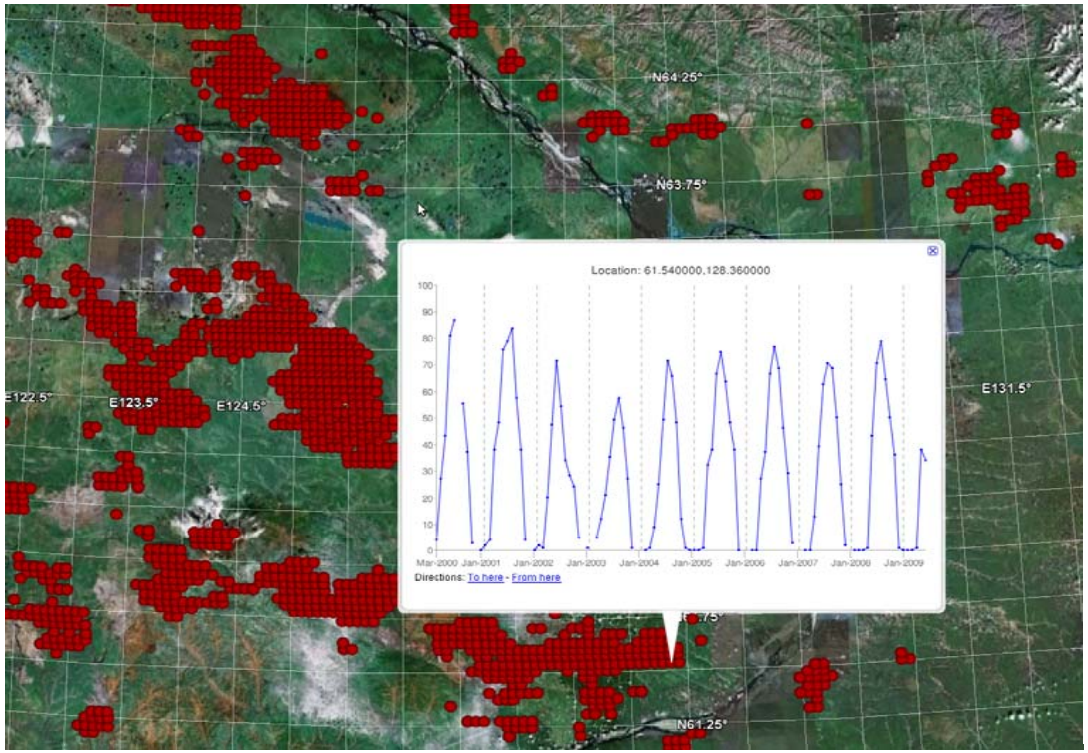


Two huge forest fires have become one giant inferno sweeping across the American state of Arizona.

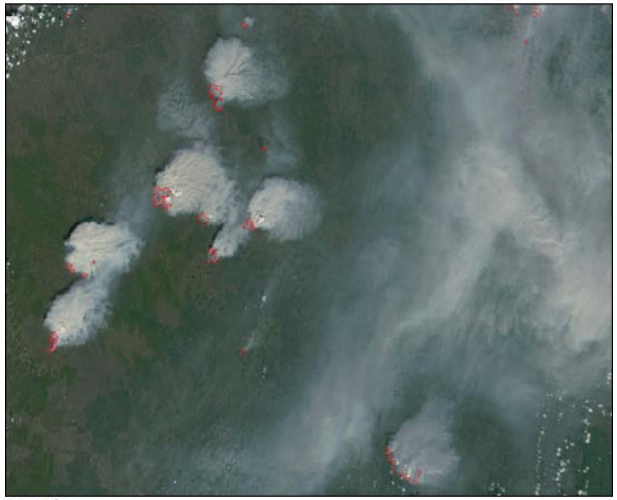
http://news.bbc.co.uk/cbbcnews/hi/world/newsid_2061000/2061402.stm

June 2002

Large Outbreak of Fires near Yakutsk, Russia



2:40 UTC

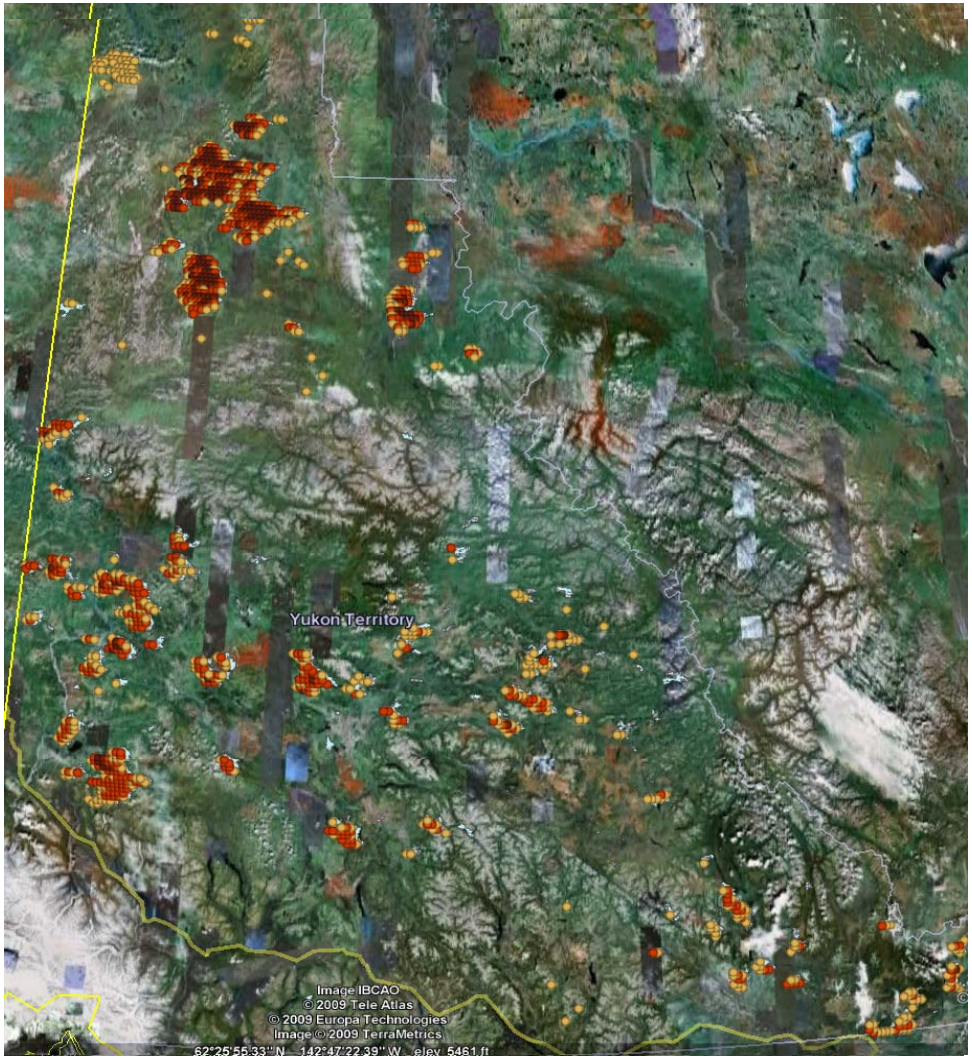


4:15 UTC

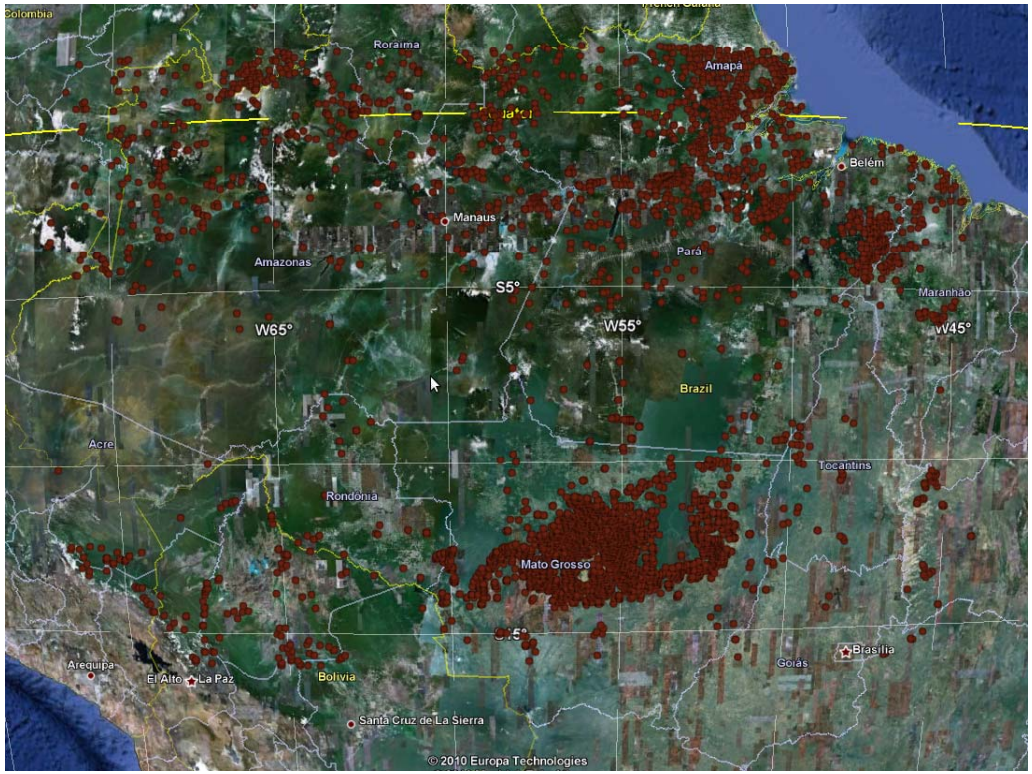
During the summer months in the Northern Hemisphere, many fires are ignited in the boreal forests of Canada and Russia by lightning striking the surface

Image courtesy Jacques Descloitres, MODIS Land Rapid Response Team

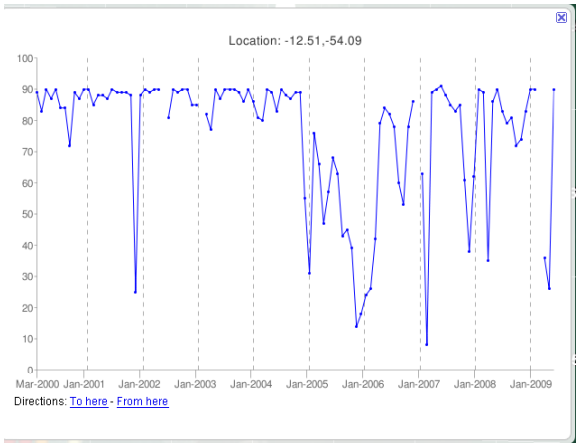
Canada: Fires in Yukon Province



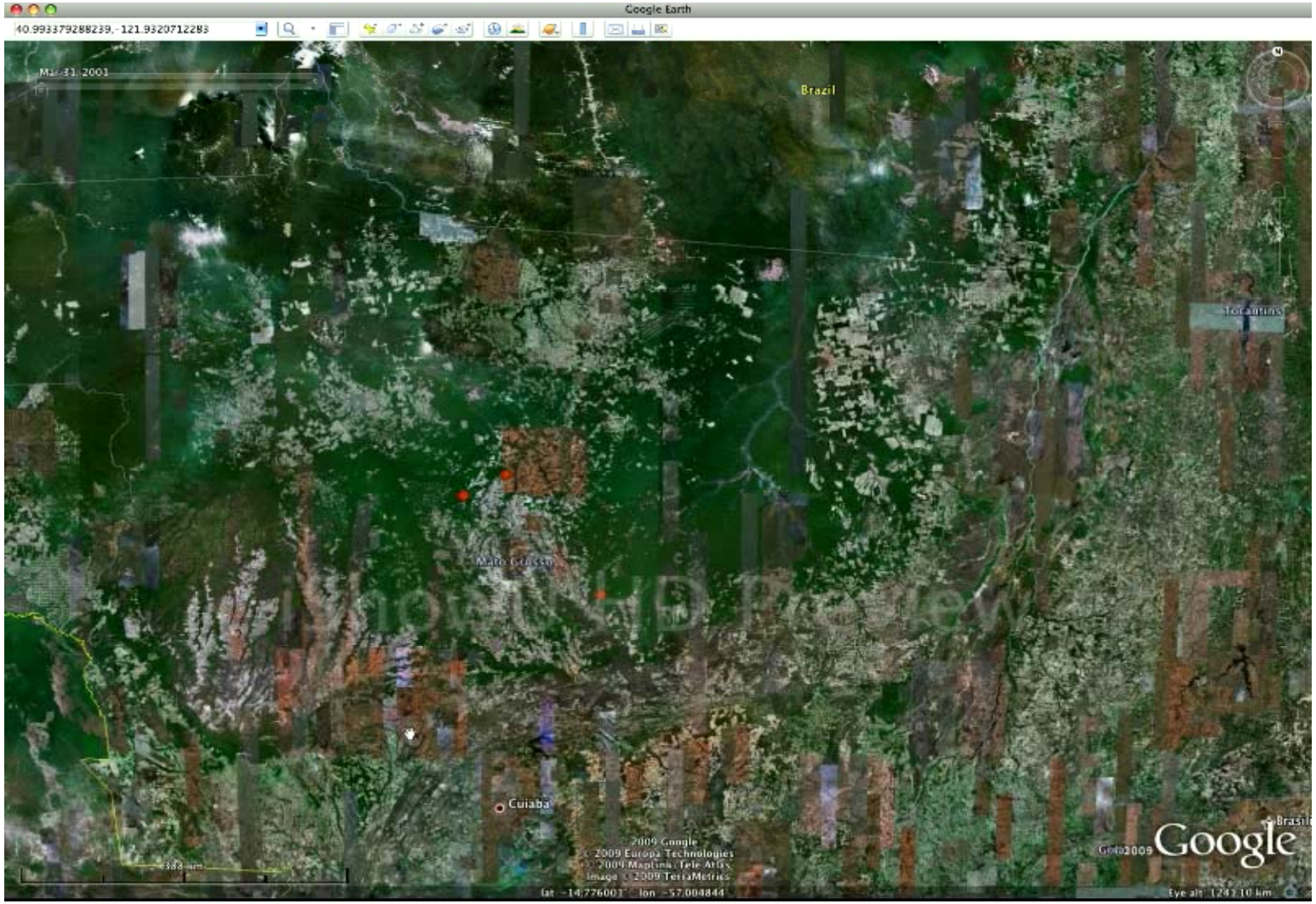
Amazon Rainforest



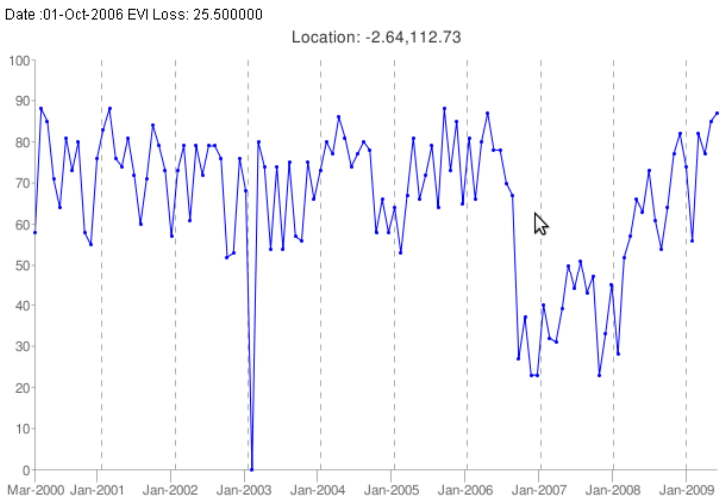
Brazil Accounts for almost 50% of all humid tropical forest clearing, nearly 4 times that of the next highest country, which accounts for 12.8% of the total.



Amazon Animation



Indonesia

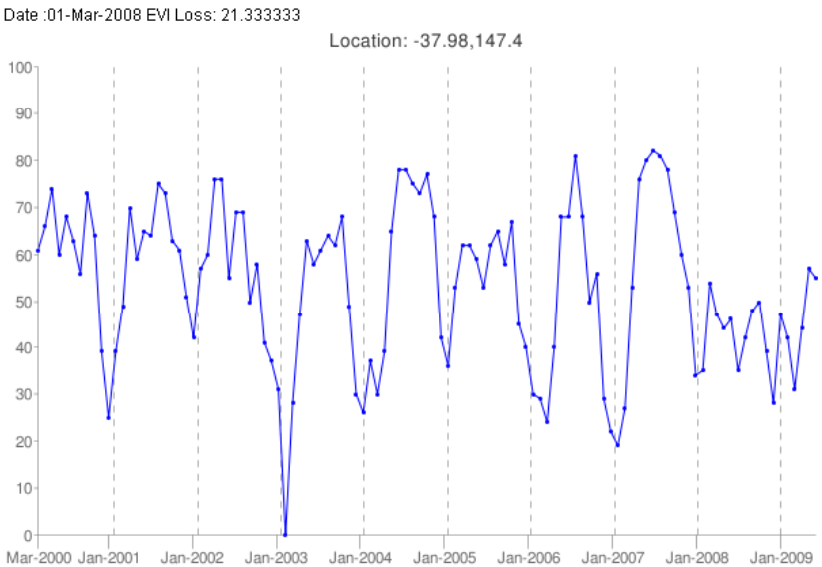


newsVOAcom

11 September 2006

Forest Fires Sweep Indonesia Borneo and Sumatra.
Officials in Indonesia say illegal burning to clear land has caused rampant wildfires across Borneo and Sumatra ... eight million hectares have gone up in smoke over the last month, and fires are still burning out of control on the island of Borneo.

Victoria (Australia)



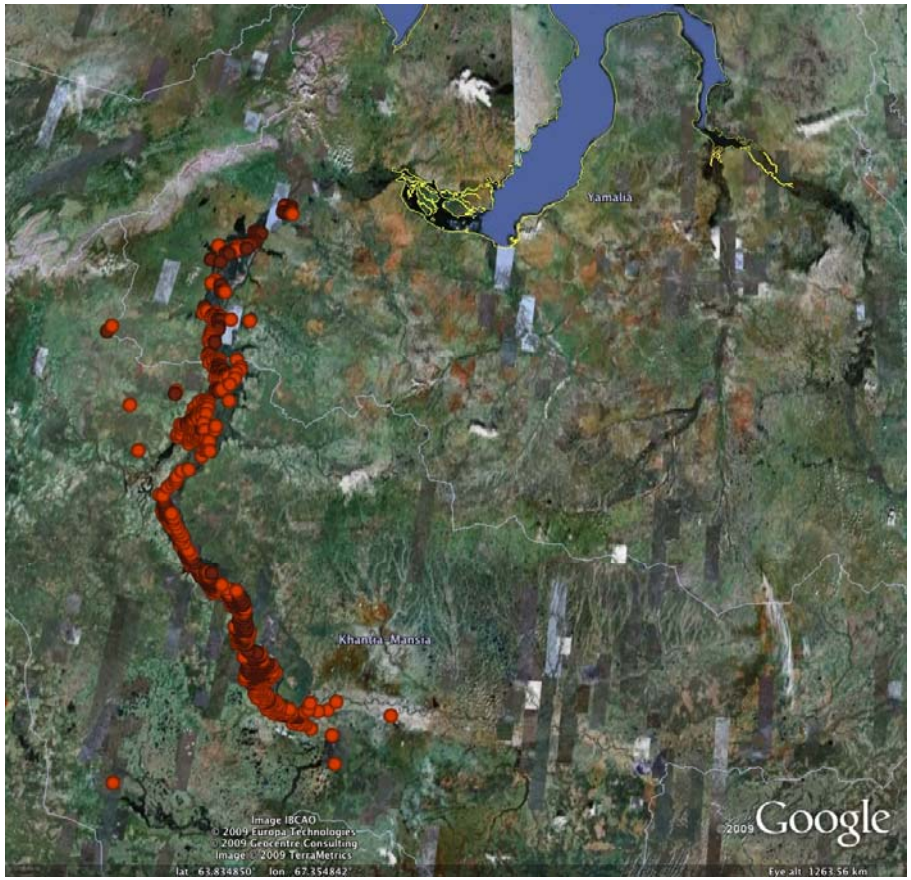
Drought in southern Australia declared 'worst on record'

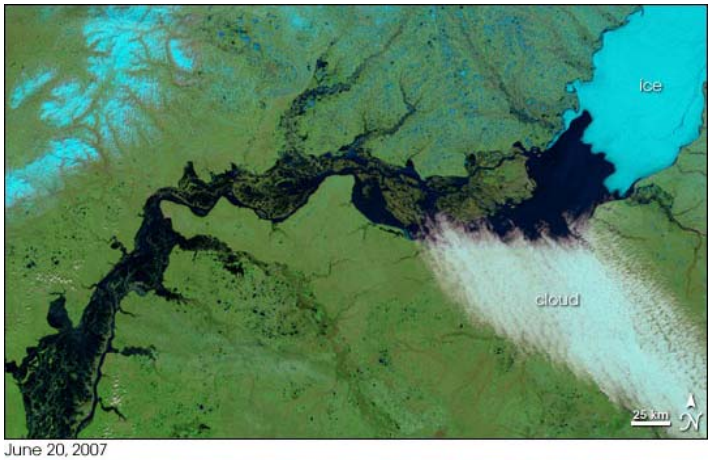
October 10, 2008

David Jones, the head of climate analysis at the Bureau of Meteorology, said the drought affecting south-west Western Australia, south-east South Australia, Victoria and northern Tasmania "is now very severe and without historical precedent".

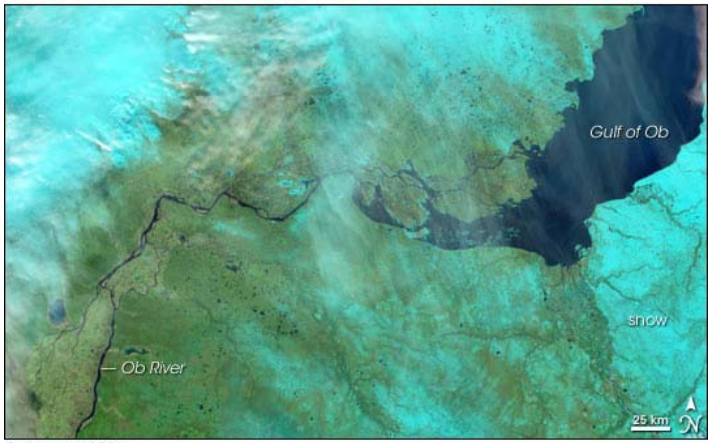
Source: climateprogress.org

Flooding along Ob River, Russia





June 20, 2007

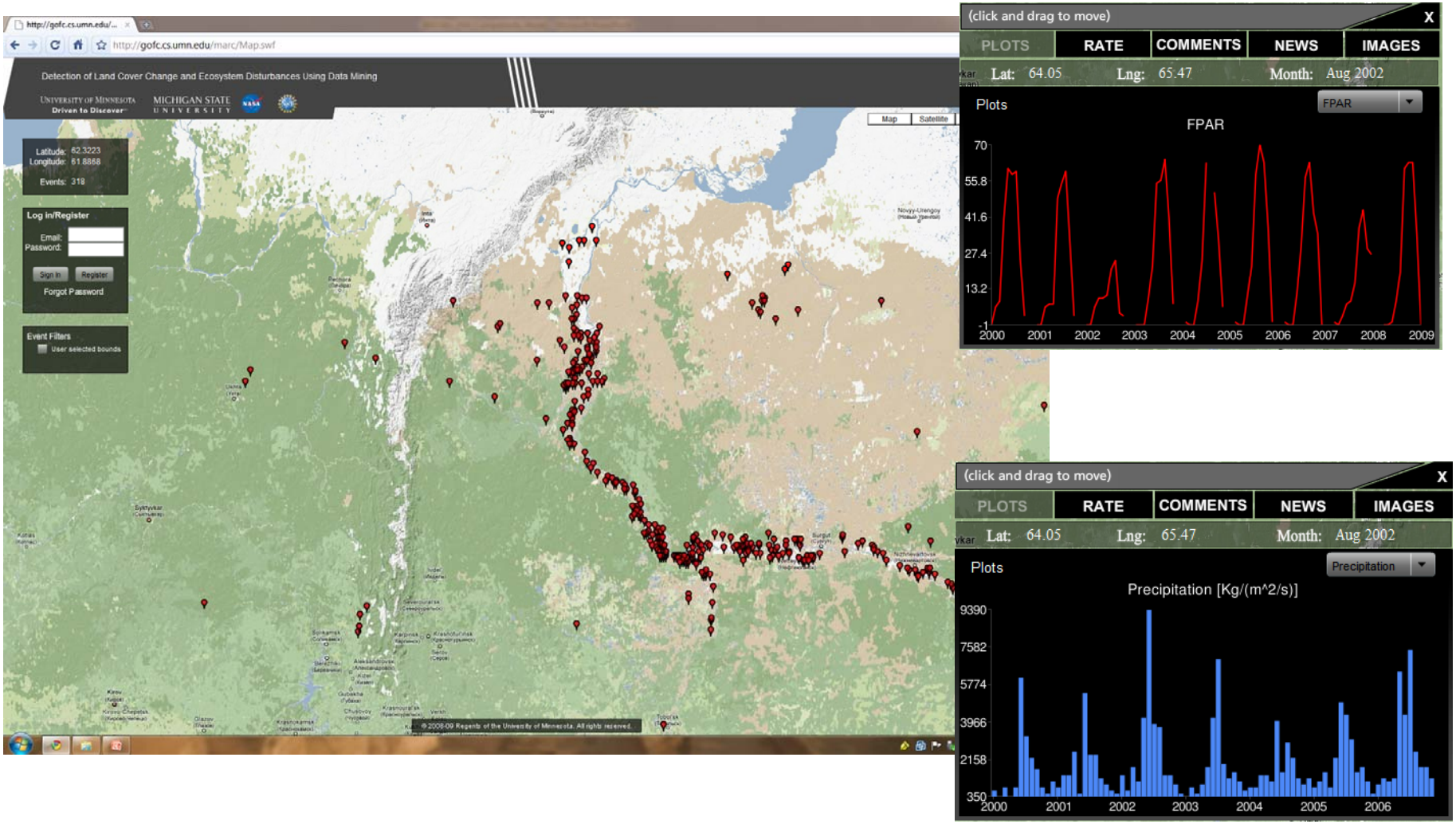


October 5, 2006

The river flows north and is blocked by ice (top right), which causes flooding. Under normal circumstances the river flows into the Gulf of Ob.

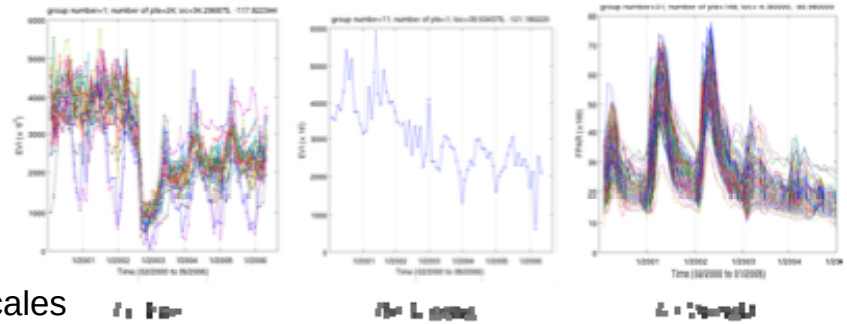
Source: NASA Earth Observatory

Web 2.0 interface for planetary information system

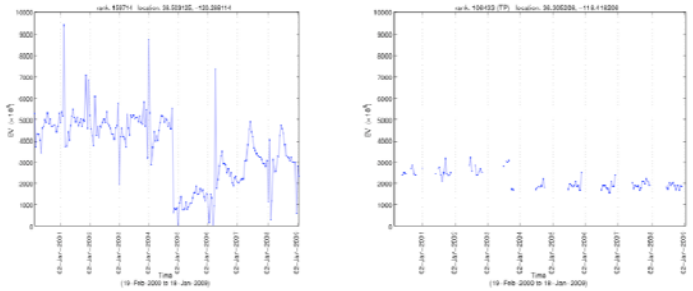


Monitoring Forest Cover Change: Challenges Ahead

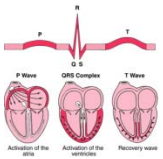
- Designing robust change detection algorithms
- Characterization of land cover changes
- Multi-resolution analysis (250m vs 1km vs 4km)
  - Different kinds of changes are visible at different scales

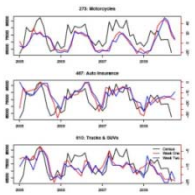


- Multivariate analysis
  - Detecting some types of changes (e.g. crop rotations) will require additional variables.



- Data quality improvement
  - Preprocessing of data using spatio-temporal noise removal and smoothing techniques can increase performance of change detection.
- Incremental update and Real-time detection





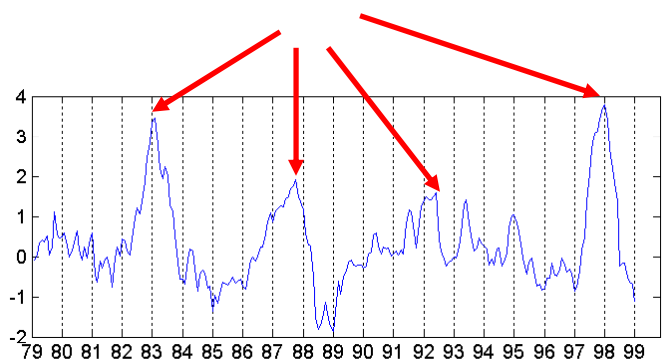
- Spatial event identification
- Applications in variety of domains:
  - Climate, agriculture, energy
  - Economics, health care, network traffic

Source: Merck, Google.

Case Study 2:
Discovering teleconnections:
Relationship among ocean/atmosphere and the land

- Climate indices capture teleconnections (in both space and time)
  - The simultaneous variation in climate and related processes over widely separated points on the Earth

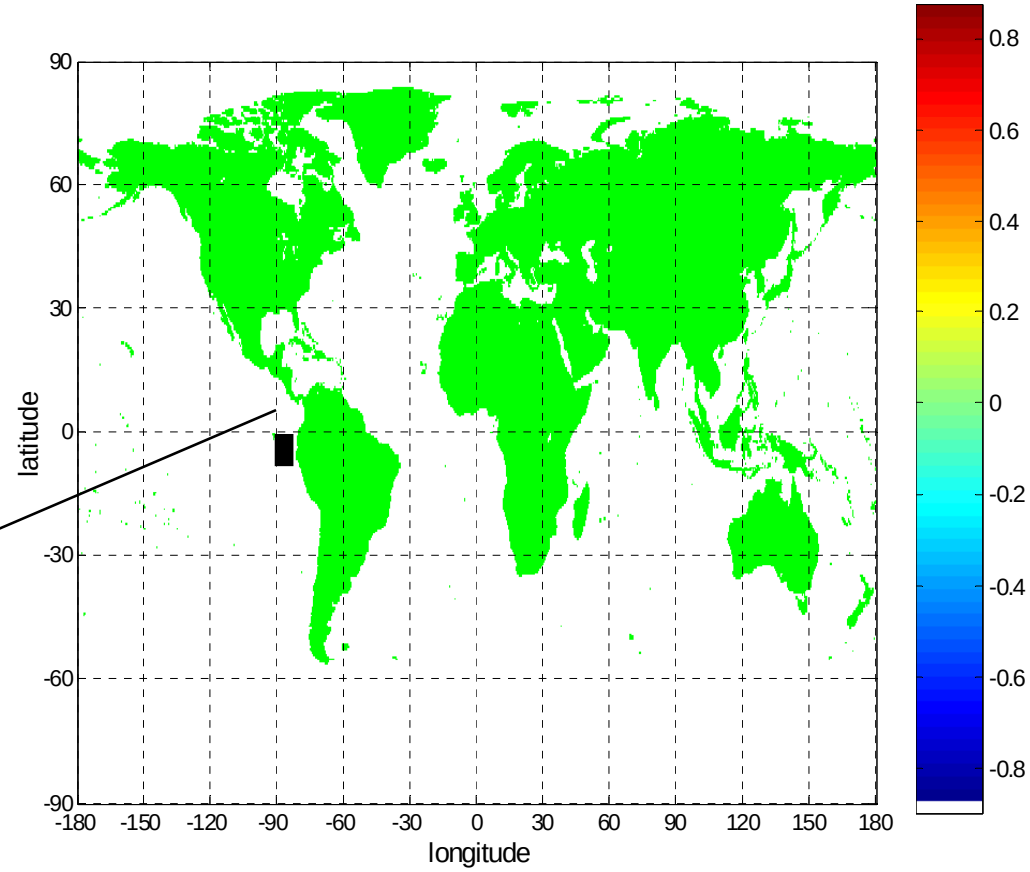
El Nino Events



Nino 1+2 Index

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

Correlation Between ANOM 1+2 and Land Temp (>0.2)

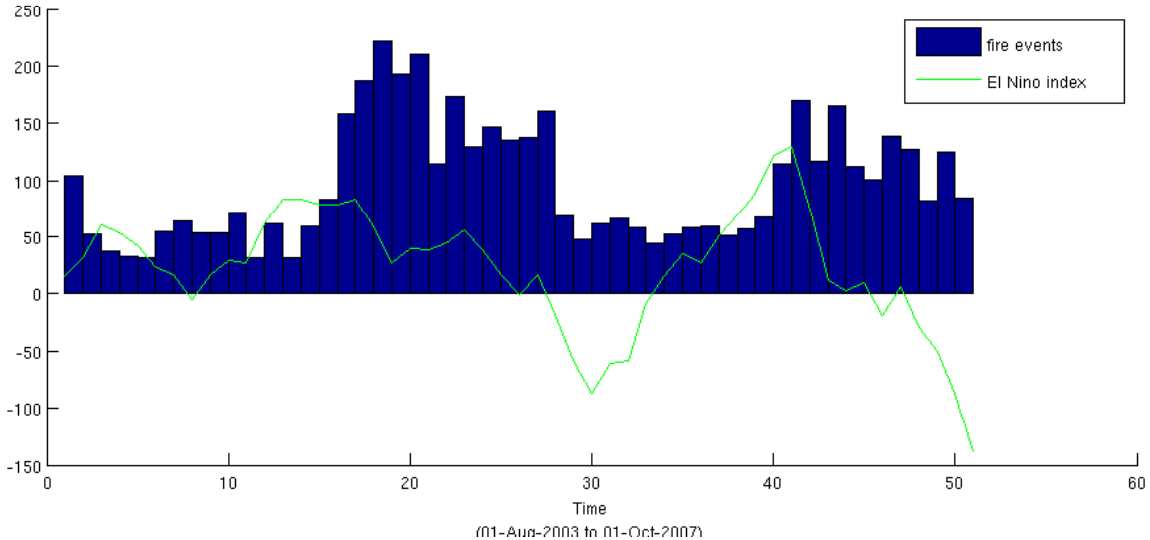


Effects: Drought in Australia, warmer winter in North America, flooding in coastal Peru, increased rainfall in East Africa

Relationship between El Nino and Fires in Indonesia

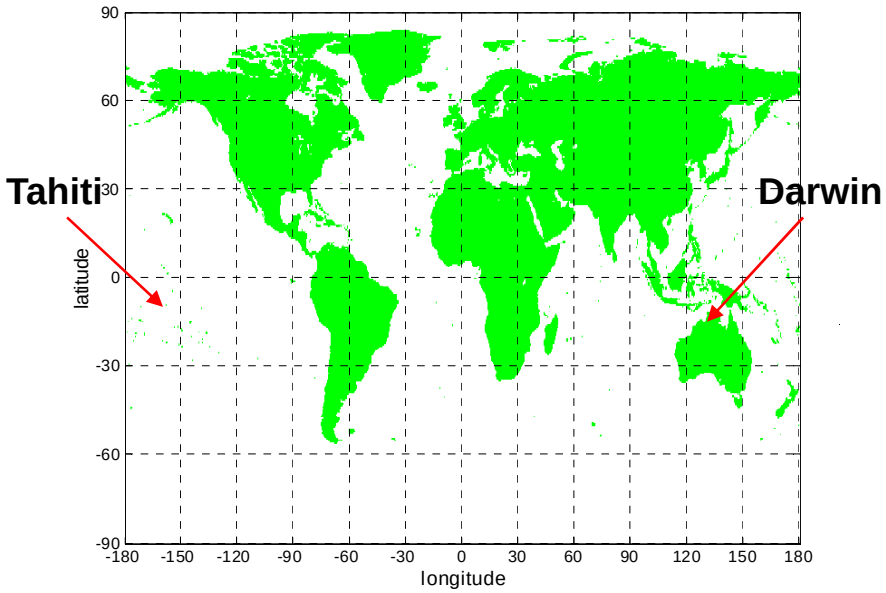
El Nino induced drought conditions worsen the forest fires in Indonesia.

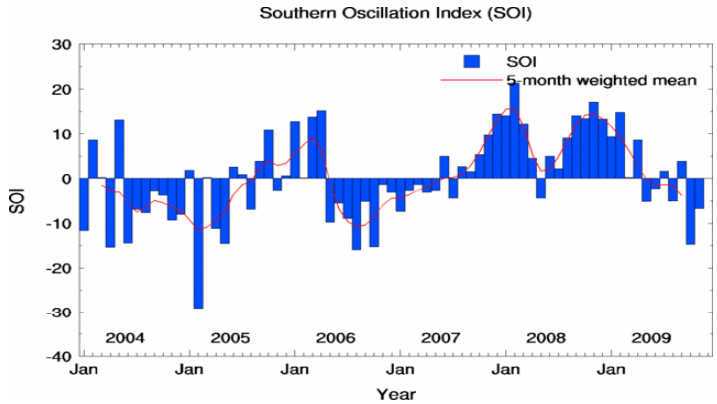
The figure shows a positive correlation of El Nino with the number of forest fires detected in Indonesia



A Pressure Based El Niño 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

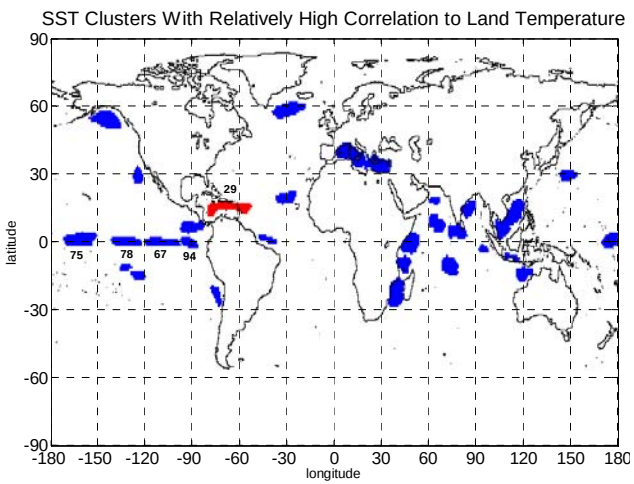
Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP 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

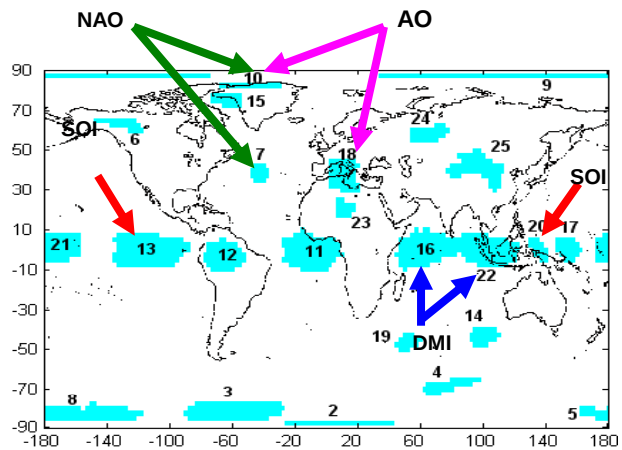
Discovered primarily by human observation

Discovery of Climate Indices Using Clustering

- Clustering provides an alternative approach for finding candidate indices.
- Clusters are found using the Shared Nearest Neighbor (SNN) method that eliminates “noise” points and tends to find homogeneous regions of “uniform density”.
- Clusters are filtered to eliminate those with low impact on land points

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



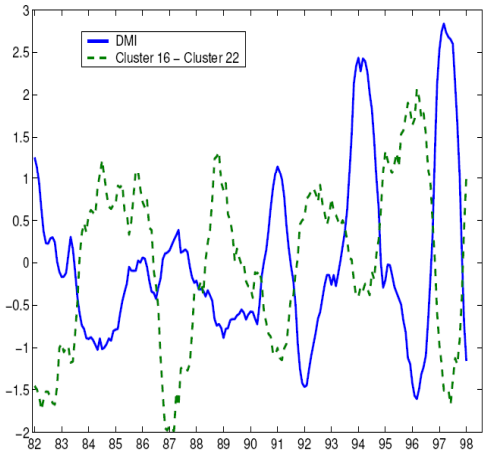


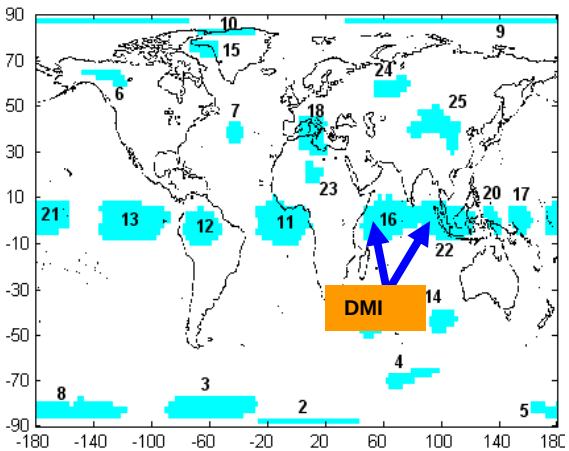
M. Steinbach, P. Tan, V. Kumar, C. Potter and S. Klooster. Discovery of Climate Indices Using Clustering, *Proceedings of KDD 2003*.

Automated Discovery of Climate Indices: Opportunities and Challenges

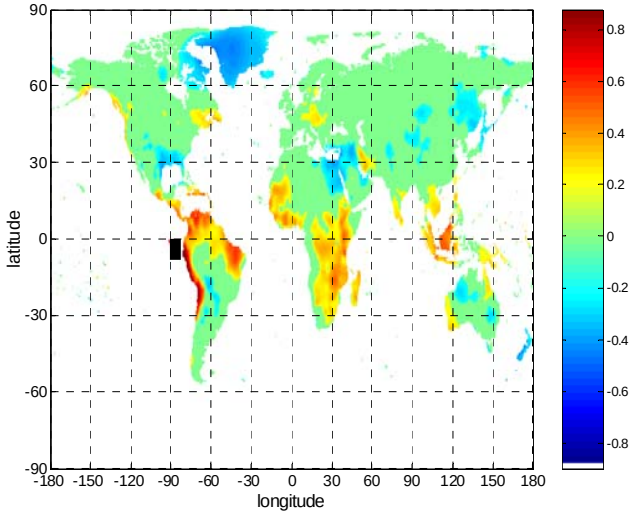
Opportunities:

- Discover new relationships that are difficult to find manually
- Example:
  - **DMI** is a temperature based index which is an indicator of weak mansoon over Indian subcontinent and heavy rainfall over east Africa.
  - **Clustering** finds a pressure based surrogate

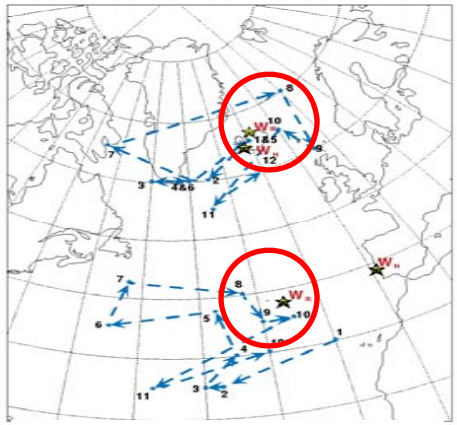




Correlation Between ANOM 1+2 and Land Temp (>0.2)



Phenomenon underlies NAO is dynamic



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

Challenges:

- Distinguishing spurious relationships from real
- Nonlinear, dynamic relationships
- Finding relationship between climate variables at Multi-scale/Multi-resolution

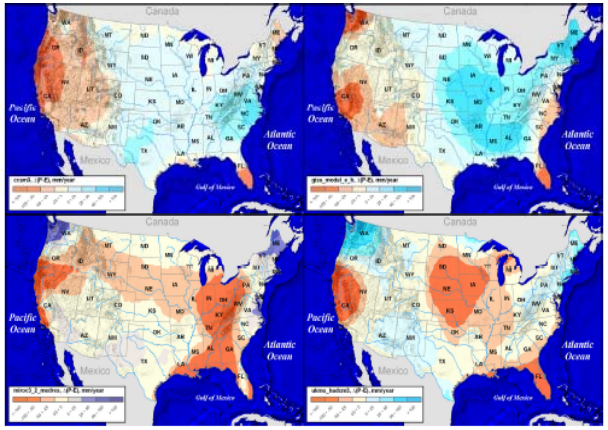
Case Study 3:

Planning for Climate Change and Extreme Events

- **Physics-based Models are Essential but Not Adequate**
- **Models make relatively reliable predictions at global scale for ancillary variables:**
  - Sea Surface Temperature (SST)
  - Temperature/humidity profiles over land
  - Wind spread at different heights
- **They provide least reliable predictions for variables that are crucial for impact assessment:**
  - Regional precipitation and extremes
  - Hurricane intensity and frequency
  - Droughts and floods

"The sad truth of climate science is that the most crucial information is the least reliable" (Nature, 2010)

Disagreement between IPCC models



Regional hydrology ("P–E" changes in 2030s) exhibits large variations among major IPCC model projections

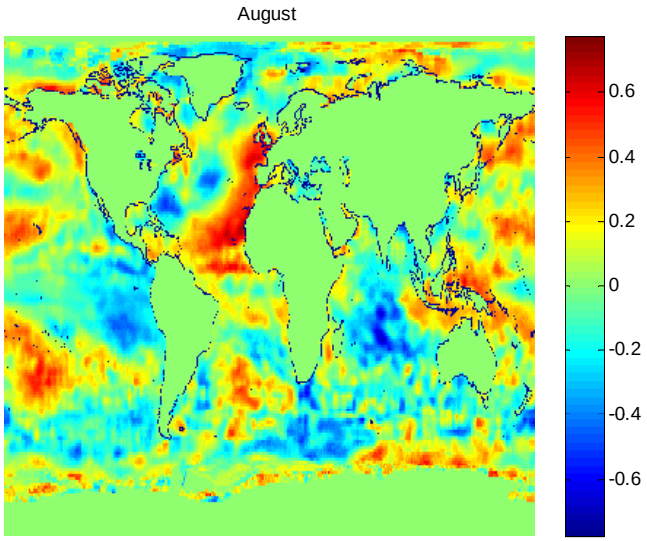
Hypothesis-driven "manual" conceptual models try to address this gap:

- Hurricane models (Emanuel et al, BAMS, 2008)
- Regional-scale precipitation extremes (O'Gorman & Schneider, PNAS, 2008; Sugiyama et al, PNAS, 2010)

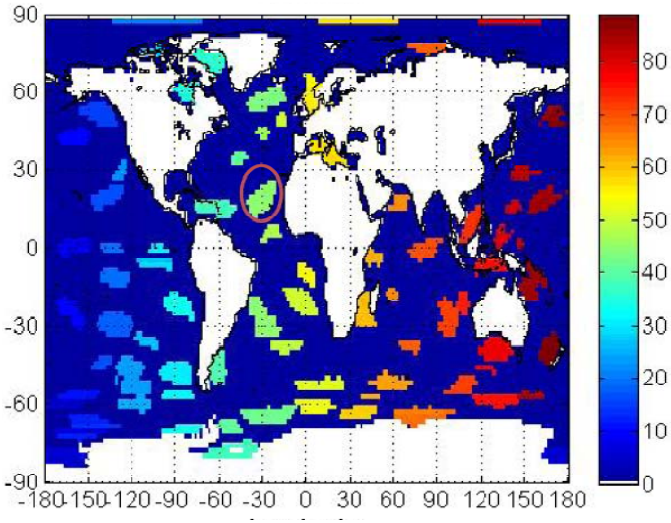
We need a systematic approach to semi-automatic data-driven model inference.

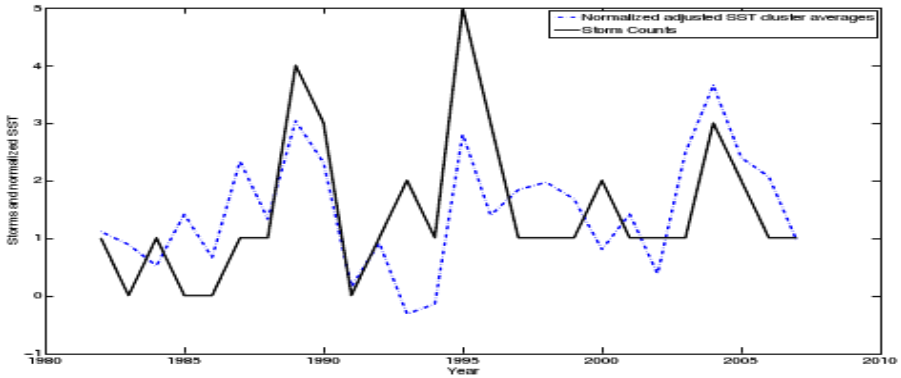
Example: Connection of Tropical Cyclones and Sea Surface Temperature

- There is a strong relationship between SST and the number of tropical cyclones off the west coast of Africa.(Semazzi, Diaz)



Correlation between Sea Surface Temperature and the number of tropical cyclones off the western coast of Africa from 1982 to 2007.

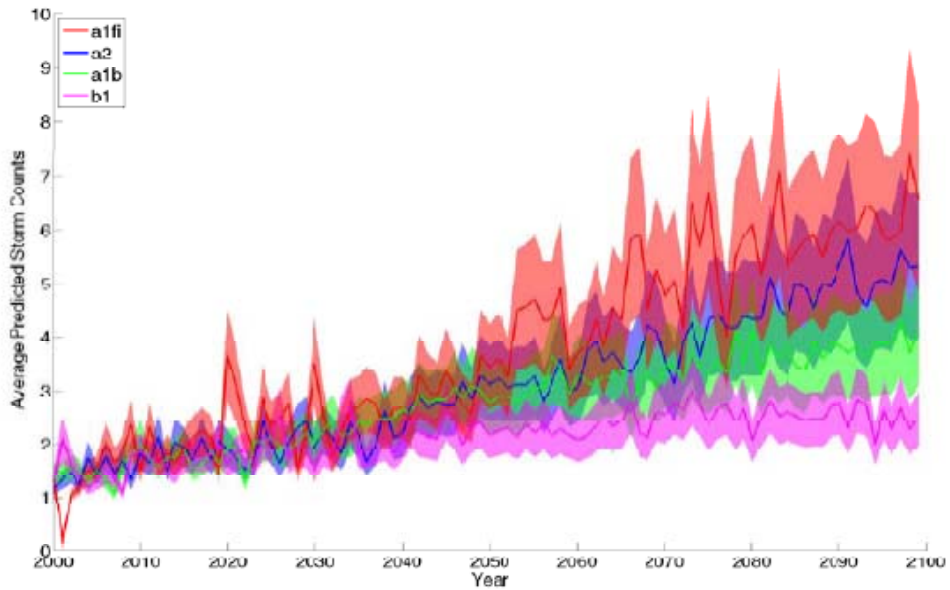




Correlation between Sea Surface Temperature and the number of tropical cyclones off the western coast of Africa from 1982 to 2007.

Predicting Tropical Storm Counts from Climate Model Projections for SST

- Built a regression model that relates August SST values off the western coast with the August tropical storm counts.
- Used predicted SST from climate scenarios produced by Global Climate Models (GCMs) to compute projected cyclones.

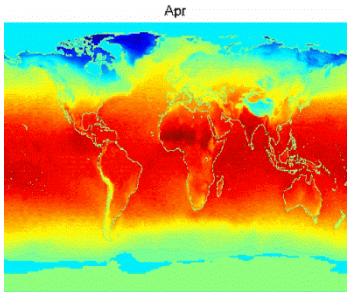


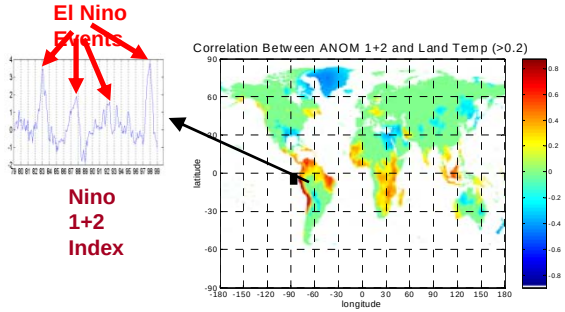
10 year moving average of predictions based on linear regression, and SST from four IPCC climate scenarios. Confidence intervals are based on estimate variability, not uncertainty. (Joint work with Ganguly and Semazzi).

- **Challenges**
  - How to model non-linear relationships?
  - How to incorporate other climate parameters e.g. Wind speed, Vegetation over near by land?
  - How to account for multiple hypothesis testing?

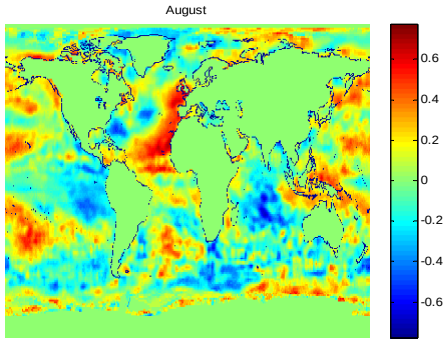
Summary

- Data driven discovery methods hold great promise for advancement in the mining of climate and eco-system data.
- Scale and nature of the data offer numerous challenges and opportunity for the data mining and HPC community.





- Multi-scale/Multi-resolution nature
- Spatio-temporal autocorrelation
- Long-range spatial dependence
- Long memory temporal processes
- Nonlinear processes, Non-Stationarity
- Variability, noise, and missing values
- Massive Size
- Fusing multiple sources of data



Team Members and Collaborators

Michael Steinbach, Shyam Boriah, Rohit Gupta, Varun Mithal, Ashish Garg, Jaya Kawale, Marc Dunham, Divya Alla, Matt Kappel, Ivan Brugere, Vikrant Krishna, Caitlin Race, James Faghmous, Dhruv Goel

Chris Potter, Auroop Ganguly, Fred Semazzi, Nagiza Samatova, Steve Klooster, Pang-Ning Tan, Joe Knight, Arindam Banerjee, Sudipto Banerjee

Project website
Climate and Eco-system: www.cs.umn.edu/~kumar/nasa-umn