Monday, September 8, 2014, 10:10 am - 12:00 pm

Opportunities and Challenges in Mining Earth System Data

Prof. Vipin Kumar

The abundance of Earth Science data from global observing satellites, models, and in-situ measurements combined with data offers an unprecedented opportunity for understanding and predicting Earth System phenomena on a global scale. Due to the large amount of data that are available, data mining techniques are needed to facilitate the automatic extraction and analysis of interesting patterns from Earth Science data. However, this is a difficult task due to the spatio-temporal nature of the data. This talk will discuss various challenges involved in analyzing large noisy and heterogeneous spatio-temporal datasets, and present some of our work on the design of efficient algorithms for finding spatio-temporal patterns from such data. We will conclude the talk with an application of identifying global mesoscale eddy trajectories in sea surface height anomaly data. For more info please visit: www.ucc.umn.edu/eddies

About Vipin Kumar

Vipin Kumar is currently William Norris Professor and Head of the Computer Science and Engineering Department at the University of Minnesota. Kumar received the B.E. degree in Electronics & Communication Engineering from Indian Institute of Technology Roorkee (formerly, University of Roorkee), India, in 1977, the M.E. degree in Electronics Engineering from Philips International Institute, Eindhoven, Netherlands, in 1979, and the Ph.D. degree in Computer Science from University of Maryland, College Park, in 1982. Kumar's current research interests include data mining, high-performance computing, and their applications in Climate/Ecosystems and Biomedical domains. Kumar is the Lead PI of a 5-year, $10 Million project, "Understanding Climate Change - A Data Driven Approach", funded by the NSF's Expeditions in Computing program that is aimed at pushing the boundaries of computer science research. He also served as the Director of Army High Performance Computing Research Center (AHPCRC) from 1998 to 2005. His research has resulted in the development of the concept of isoefficiency metric for evaluating the scalability of

parallel algorithms, as well as highly efficient parallel algorithms and software for sparse matrix factorization (PSPASES) and graph partitioning (METIS, ParMetis, hMetis). He has authored over 300 research articles, and has coedited or coauthored 11 books including widely used text books ``Introduction to Parallel Computing'' and ``Introduction to Data Mining'', both published by Addison Wesley. Kumar has served as chair/co-chair for many international conferences and workshops in the area of data mining and parallel computing, including IEEE International Conference on Data Mining (2002) and International Parallel and Distributed Processing Symposium (2001). Kumar co-founded SIAM International Conference on Data Mining and served as a founding co-editor-in-chief of Journal of Statistical Analysis and Data Mining (an official journal of the American Statistical Association). Currently, Kumar serves on the steering committees of the SIAM International Conference on Data Mining and the IEEE International Conference on Data Mining, and is series editor for the Data Mining and Knowledge Discovery Book Series published by CRC Press/Chapman Hall. Kumar is a Fellow of the ACM, IEEE and AAAS. He received the Distinguished Alumnus Award from the Indian Institute of Technology (IIT) Roorkee (2013), the Distinguished Alumnus Award from the Computer Science Department, University of Maryland College Park (2009), and IEEE Computer Society's Technical Achievement Award (2005). Kumar's foundational research in data mining and its applications to scientific data was honored by the ACM SIGKDD 2012 Innovation Award, which is the highest award for technical excellence in the field of Knowledge Discovery and Data Mining (KDD).

Opportunities and Challenges in Mining Earth System Data

Vipin Kumar

University of Minnesota

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

PLANETARY SKIN

Outline

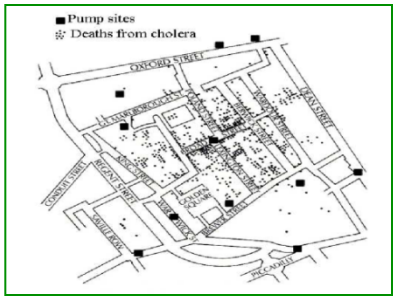
- Motivation: Brief overview of data mining
- Opportunities and challenges in Earth science
- Case Studies:
  1. Monitoring ecosystem distrubances
  2. Data-driven discovery of atmospheric dipoles
  3. Monitoring mesoscale ocean eddies
- Concluding Remarks

Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

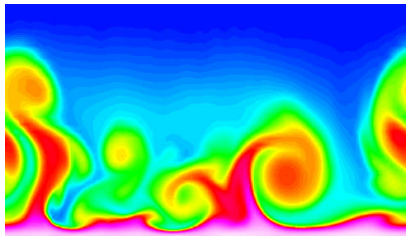
Homeland Security

Business Data



Geo-spatial data

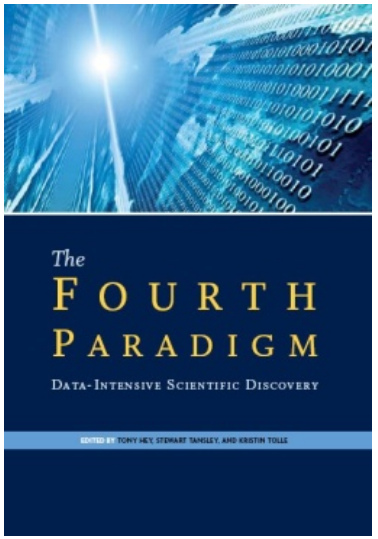
Sensor Networks



Computational Simulations

Data guided discovery - A new paradigm

"... data-intensive science [is] …a new, fourth paradigm for scientific exploration." - Jim Gray



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

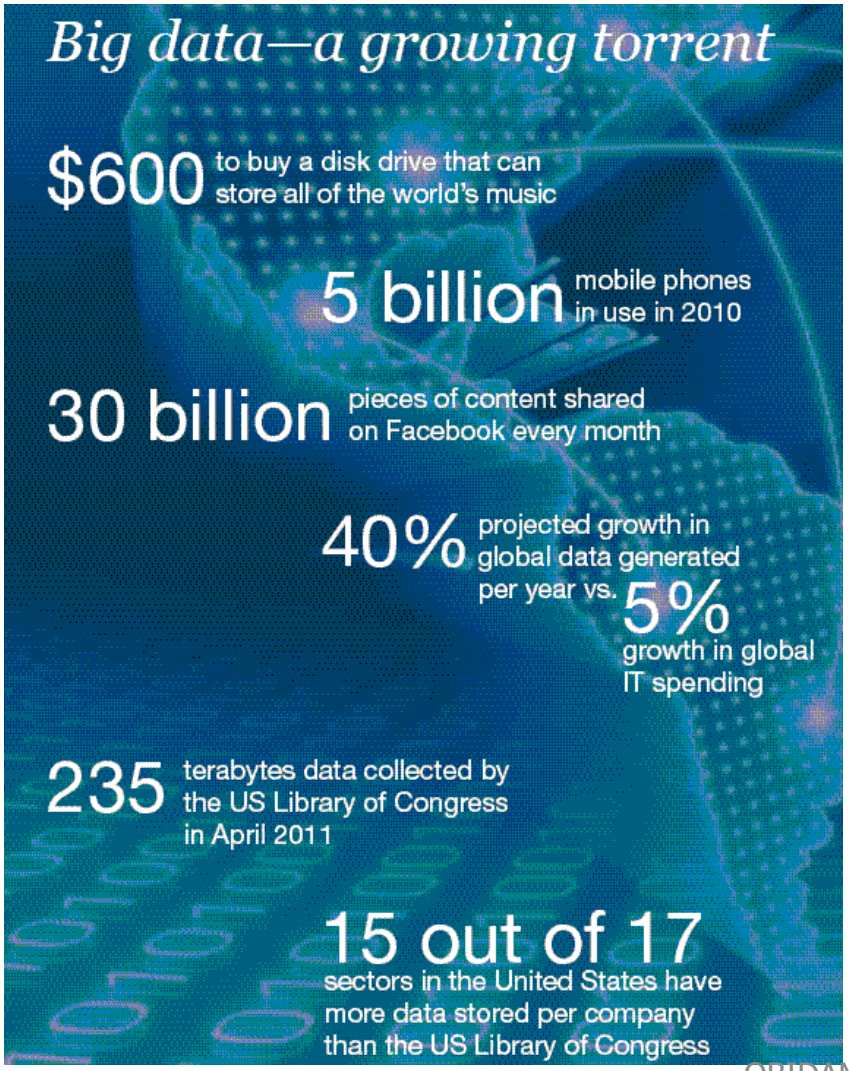
By Chris Anderson 06.23.08

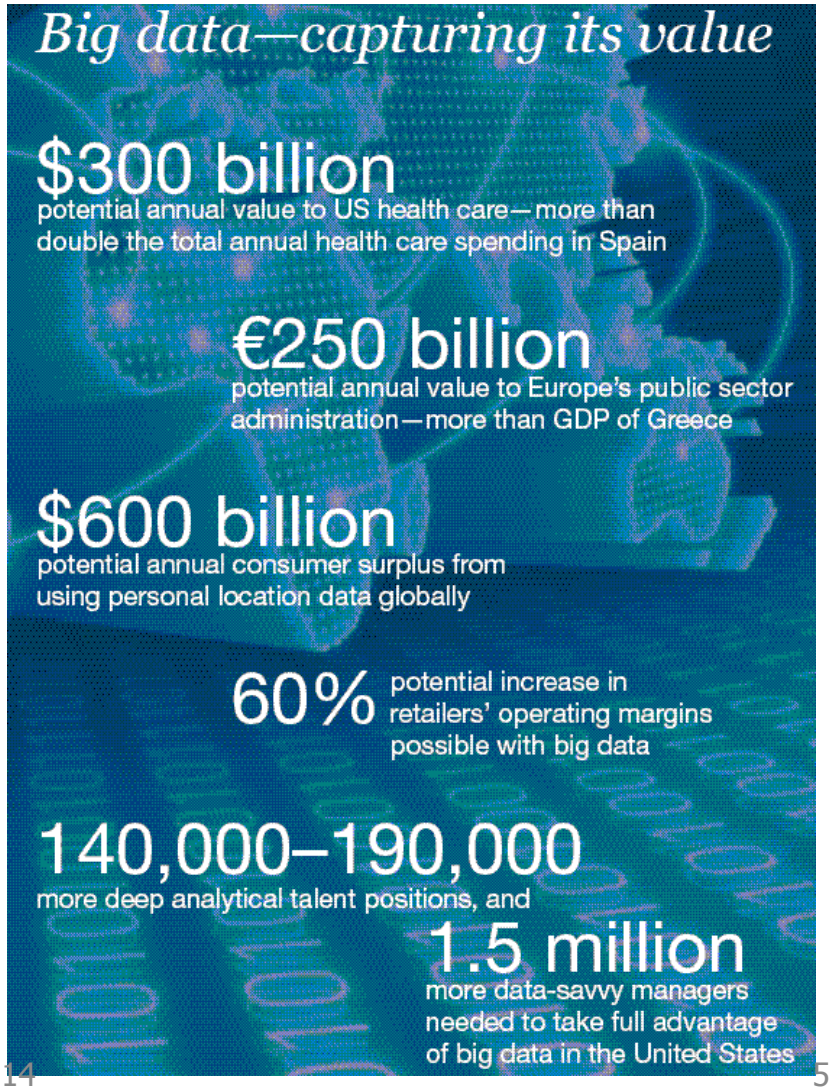
McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

McKinsey Global Institute – Report on Big Data

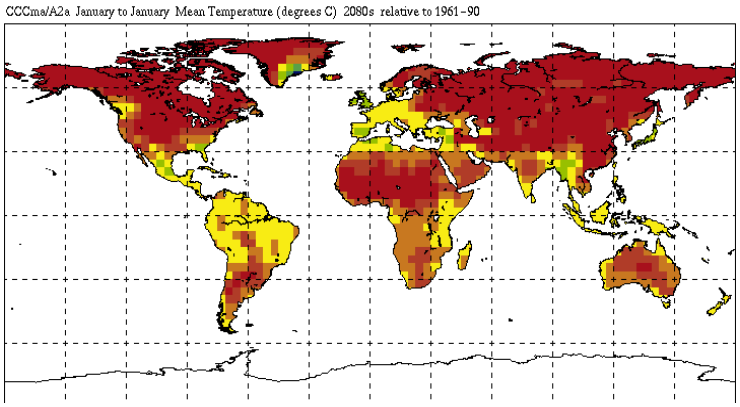
June 2011





Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

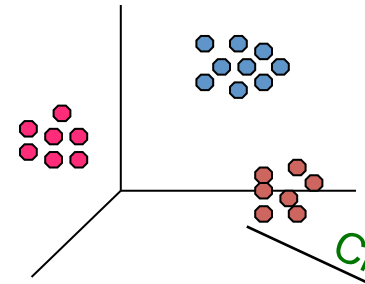
Reducing hunger and poverty by increasing agriculture production

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks ...

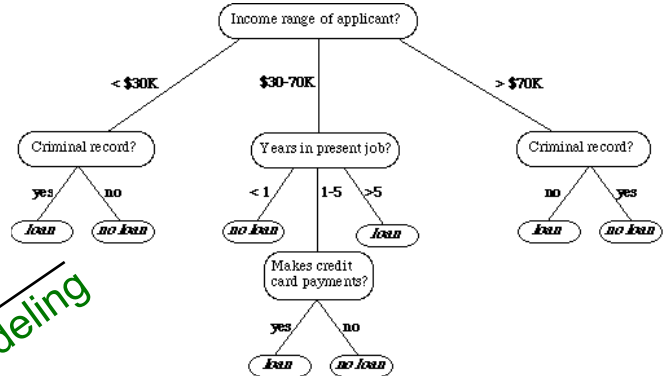


Clustering

Data

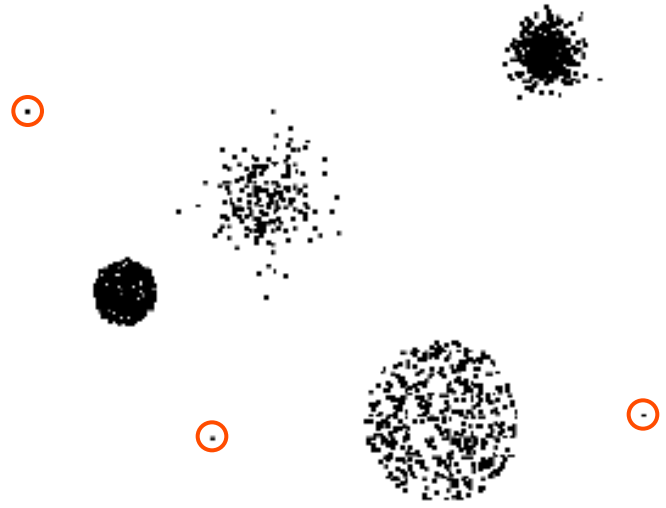
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling



Association Rules

Anomaly Detection



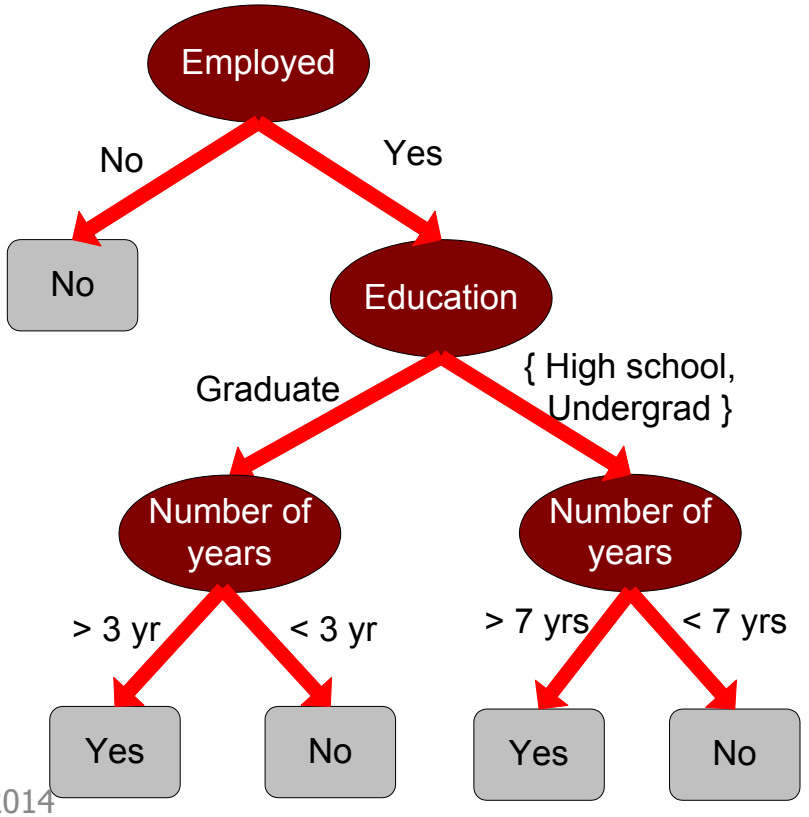
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

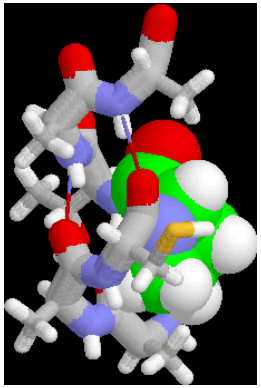
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness



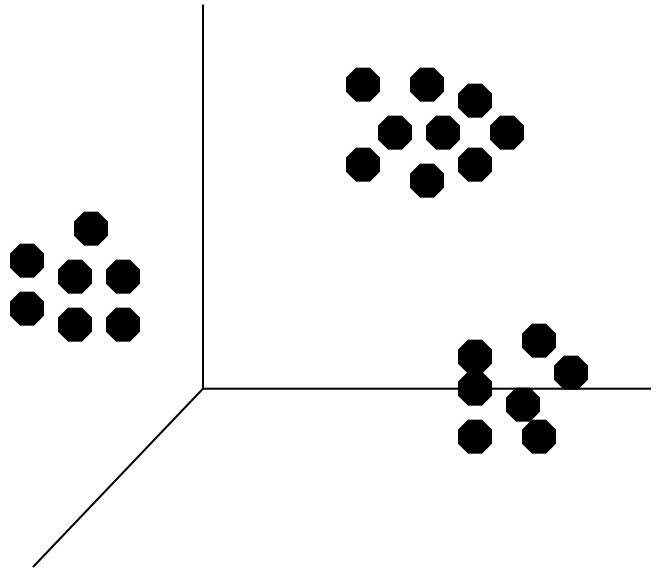
Examples of Classification Tasks

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil



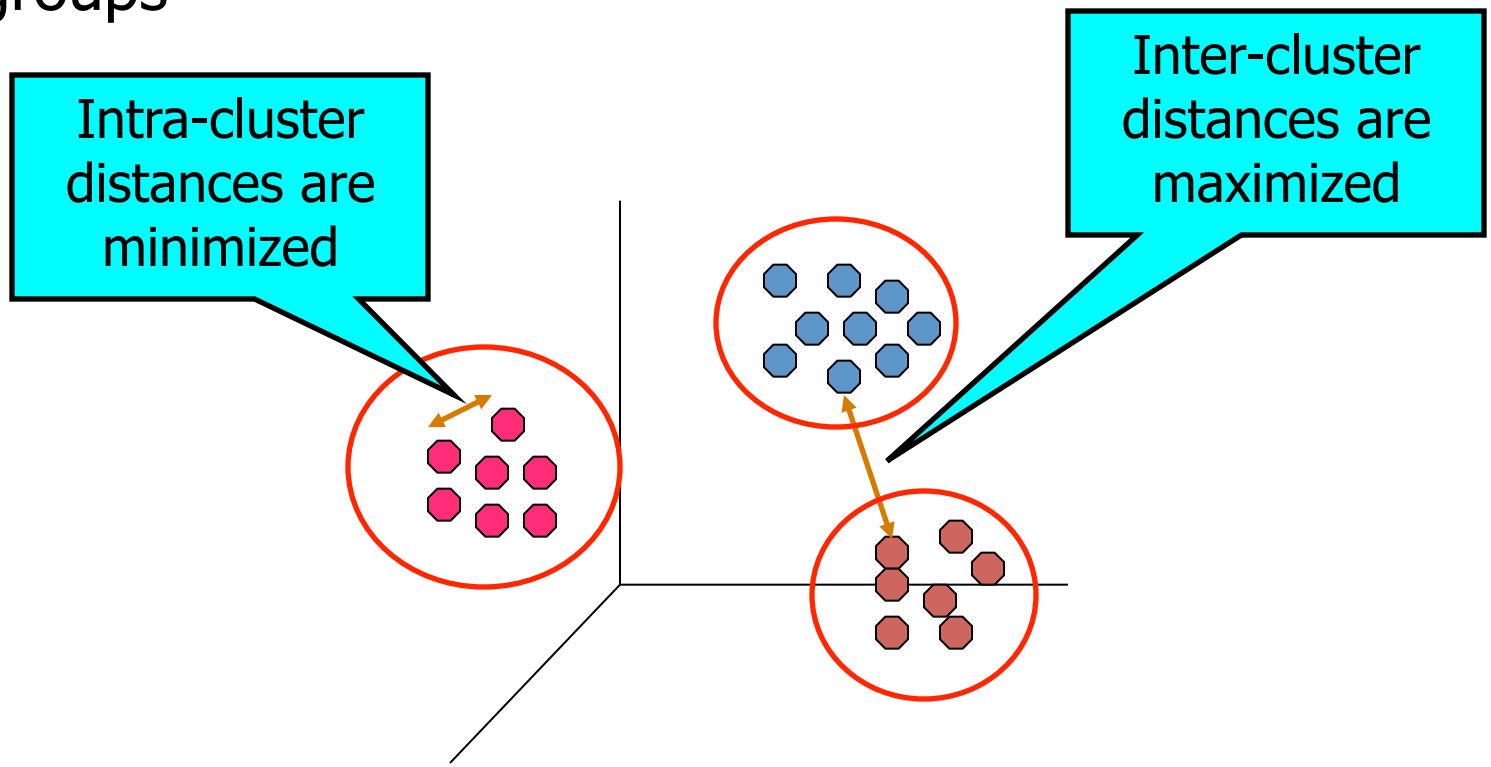
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



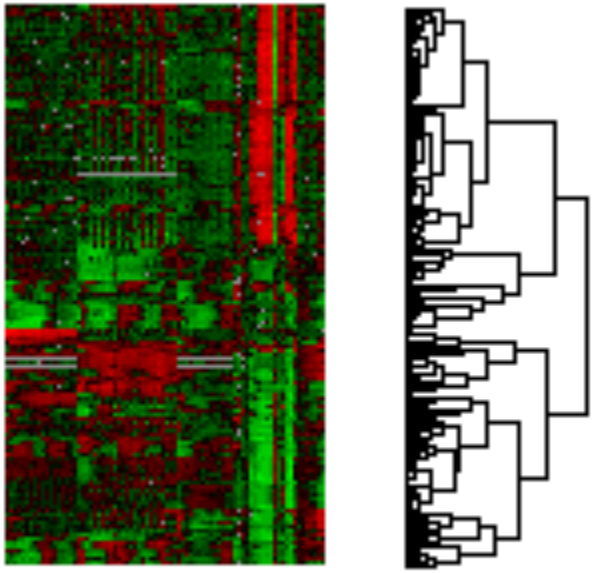
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

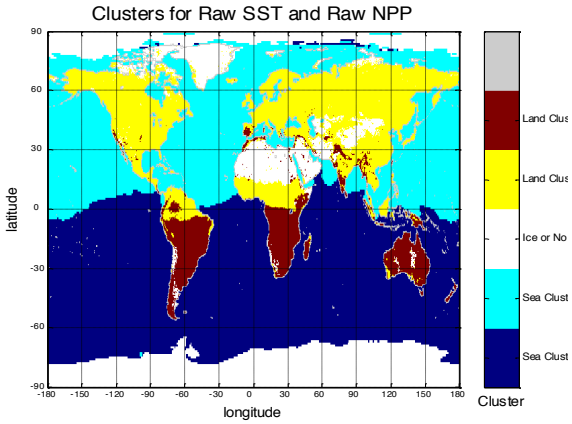


Applications of Cluster Analysis

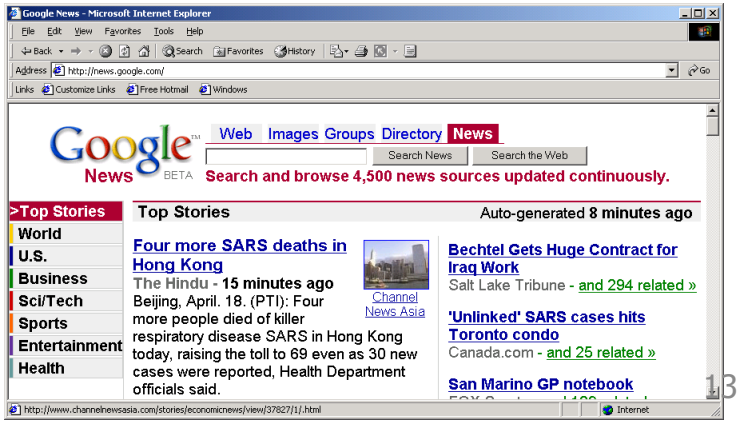
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

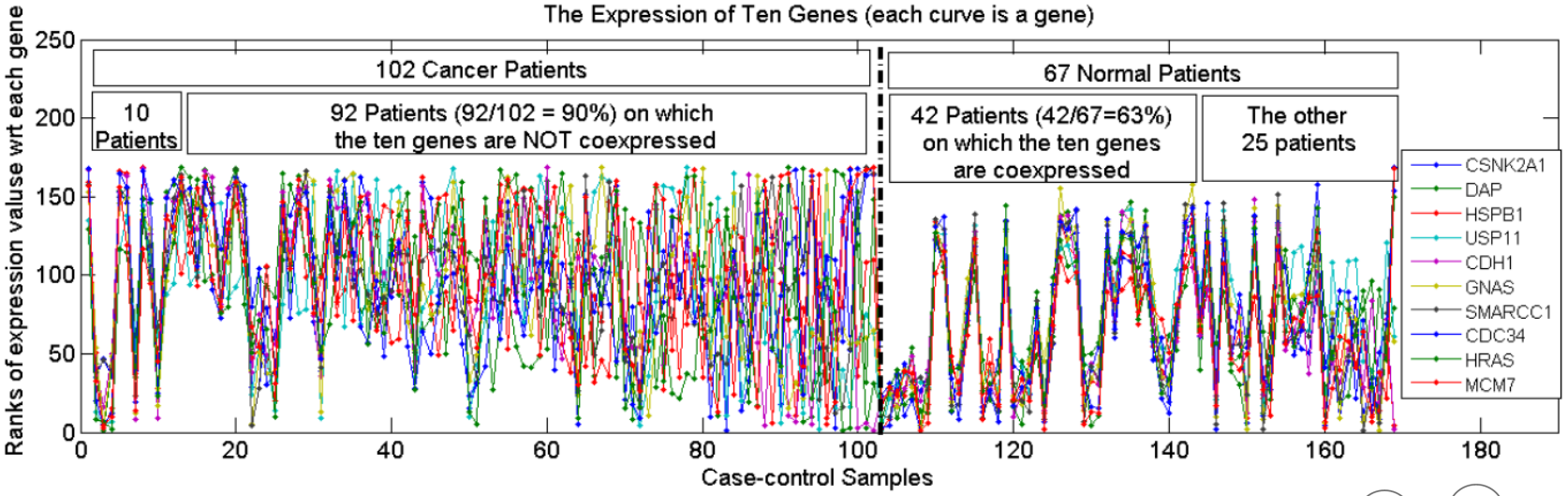
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

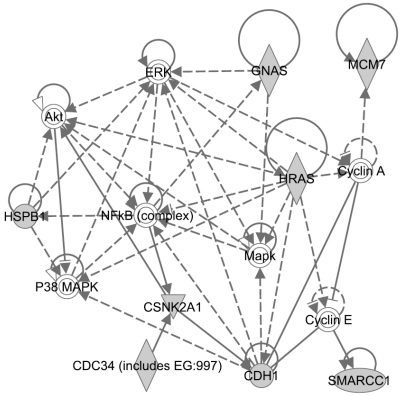
Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



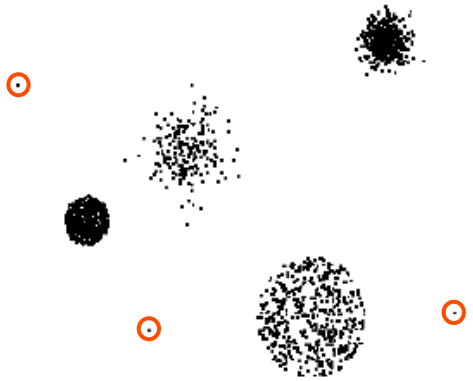
Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

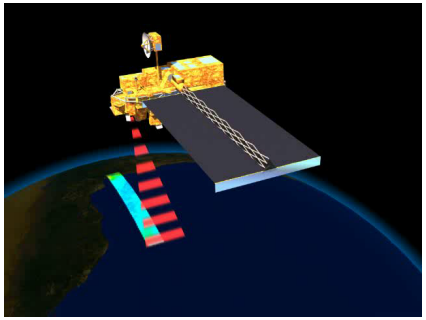


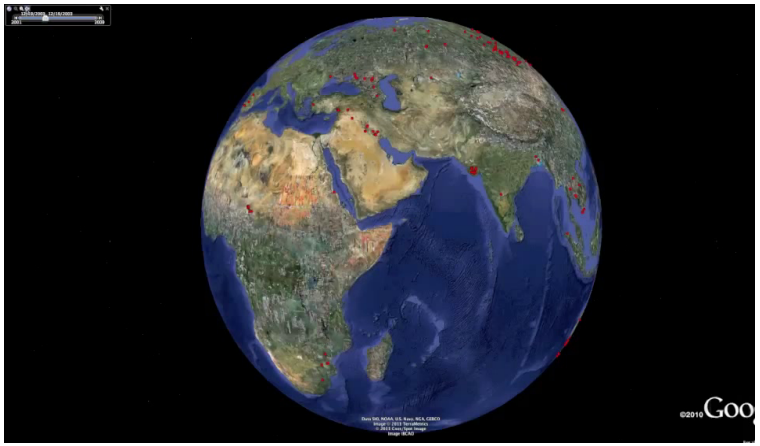
[Fang et al PSB 2010]

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.





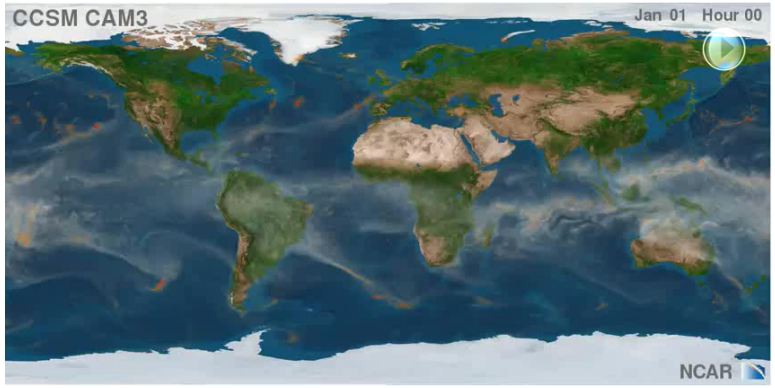


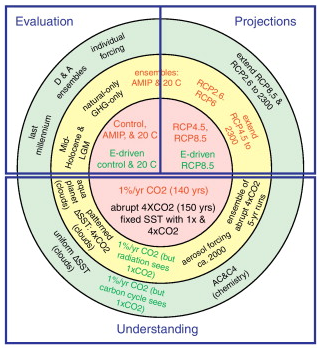
Opportunities and Challenges in Earth science

Big Data in Earth Science

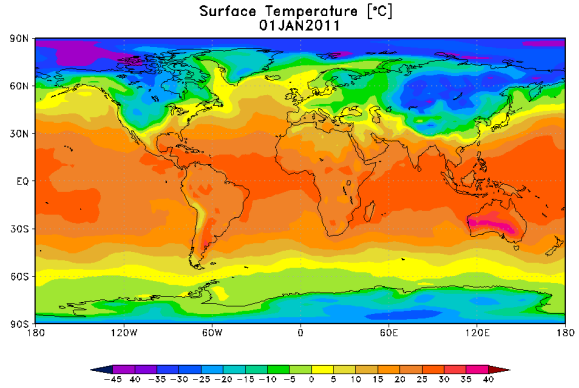
- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation
- Climate Models
- Reanalysis Data
- River Discharge
- Agricultural Statistics
- Population Data
- Air Quality
- …

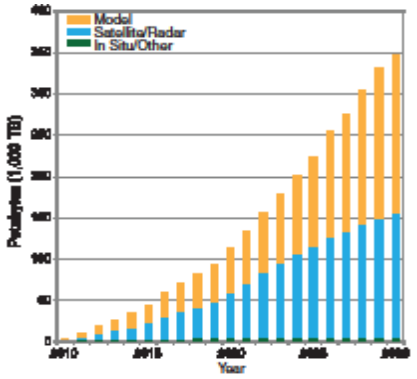
Source: NCAR





Source: NASA



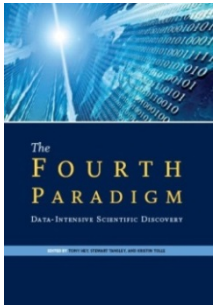


Big Data in Earth Science

- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation
- Climate Models
- Reanalysis Data
- River Discharge
- Agricultural Statistics
- Population Data
- Air Quality
- …

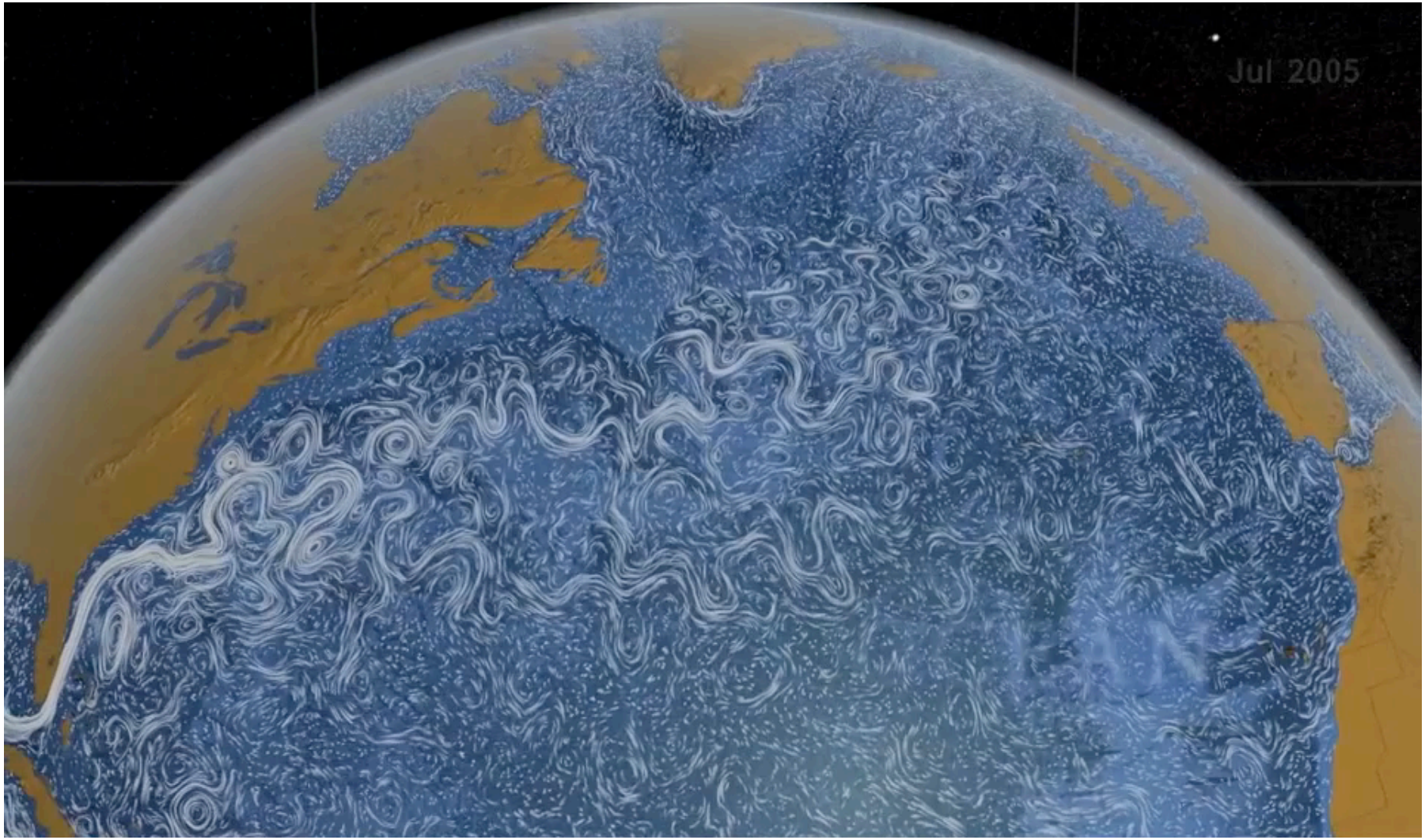
"Climate change research is now 'big science,' comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics."
(Nature Climate Change, Oct 2012)

- Scale and nature of the data offer numerous challenges and opportunities for understanding global change.



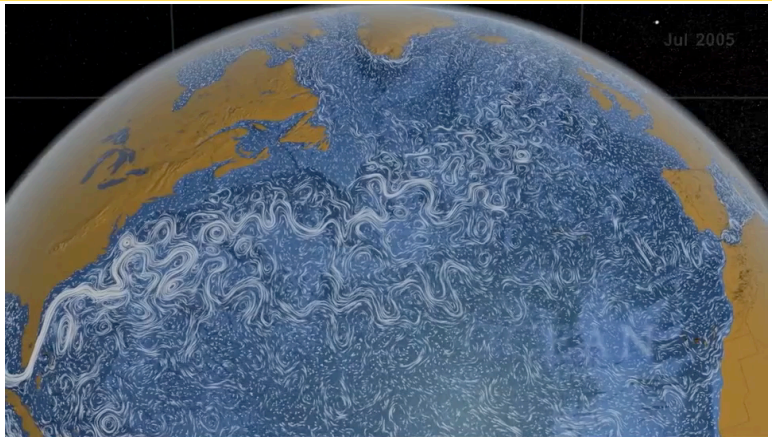
"data-intensive science [is] so different that it is worth distinguishing [it] … as a new, fourth paradigm for scientific exploration." – Jim Gray

Illustrative Research Tasks:
Understanding Global Ocean Dynamics



Source: NASA

Illustrative Research Tasks:
Understanding Global Ocean Dynamics



Source: NASA

Earth Science Research Tasks

- Understand global ocean dynamics
- Identify and monitor mesoscale ocean eddies as they impact global ocean kinetic energy, heat, and nutrients
- Assess the impact of climate change on global ocean dynamics

Data

Global weekly sea surface height anomalies (1992-2010) at 25 km resolution

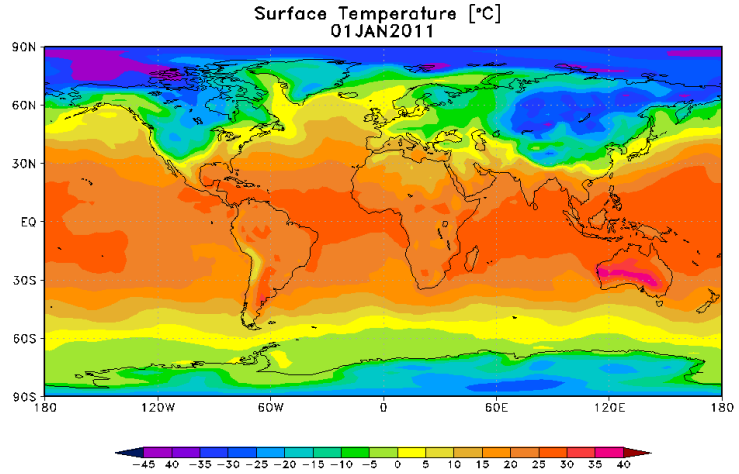
Data Mining Research Tasks

- Autonomously identify uncertain objects (no clear boundaries) in spatio-temporal field
- Autonomously track unlabeled dynamic spatio-temporal objects

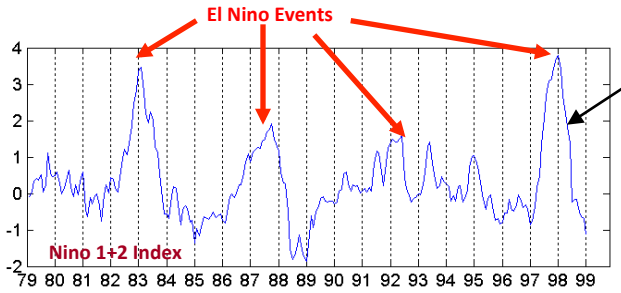
Challenges

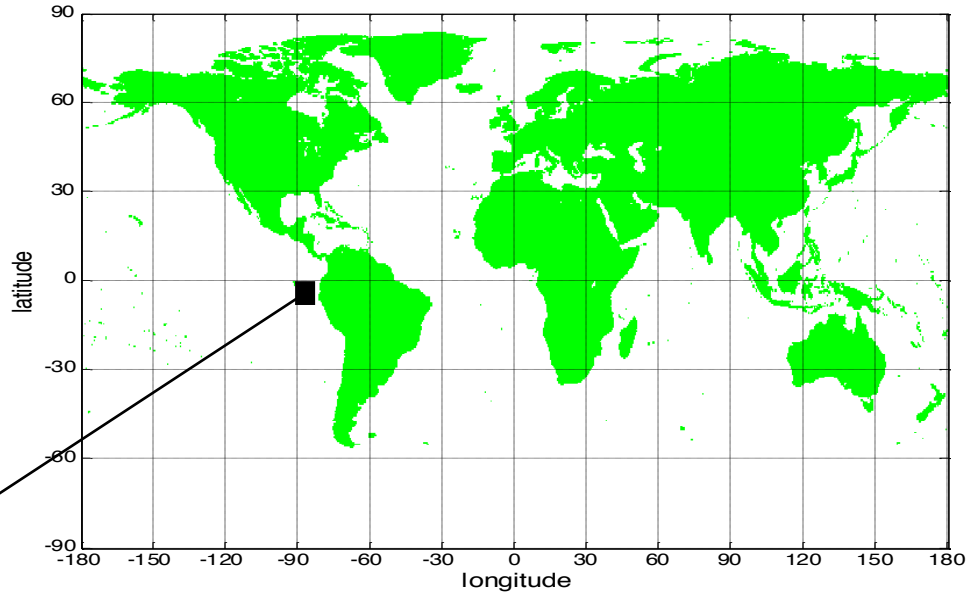
- Highly variable, noisy, and uncertain data
- Lack of ground truth
- Post-processed data makes data artificially smooth (difficult to identify boundaries)
- Objects are dynamic in space and time (size, shape, properties change abruptly)

Illustrative Research Tasks:
Relationship Mining in Climate Data

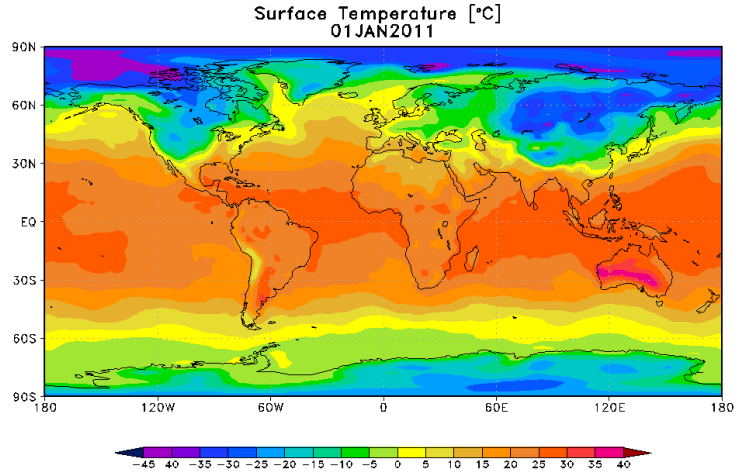


Identifying Atmospheric Teleconnections

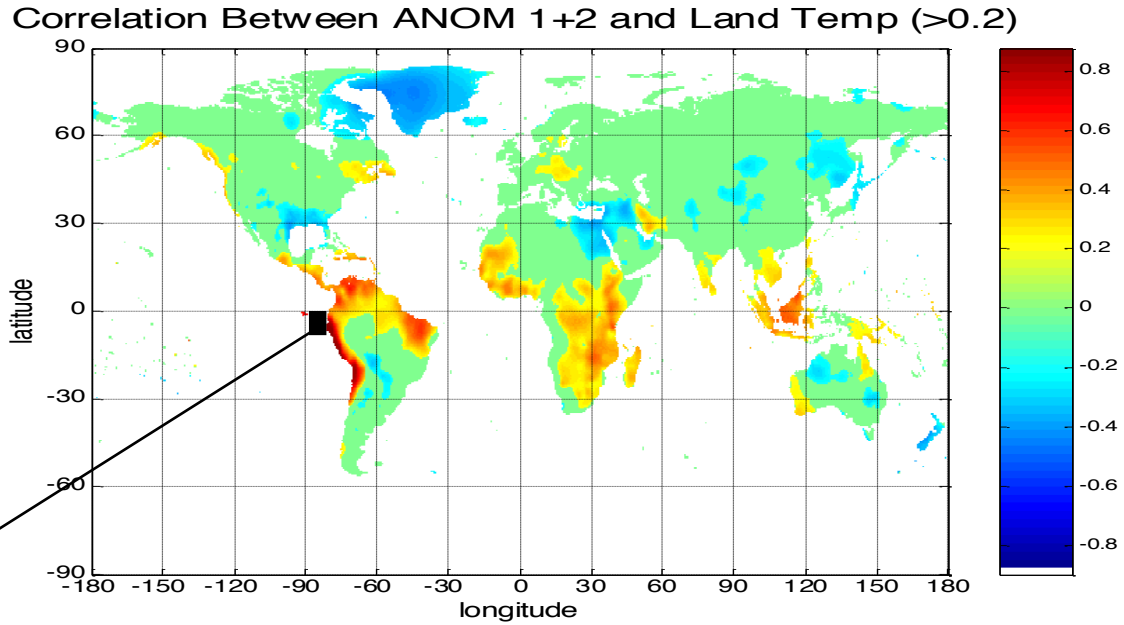


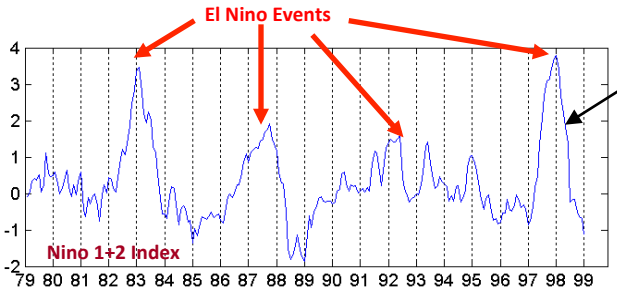


Illustrative Research Tasks:
Relationship Mining in Climate Data

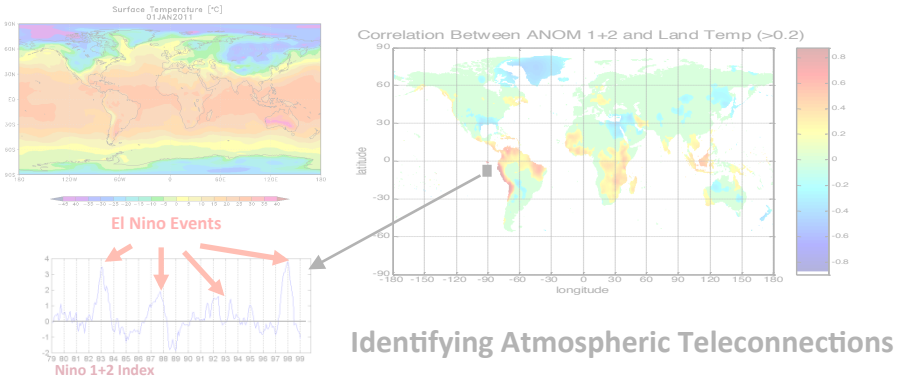


Identifying Atmospheric Teleconnections

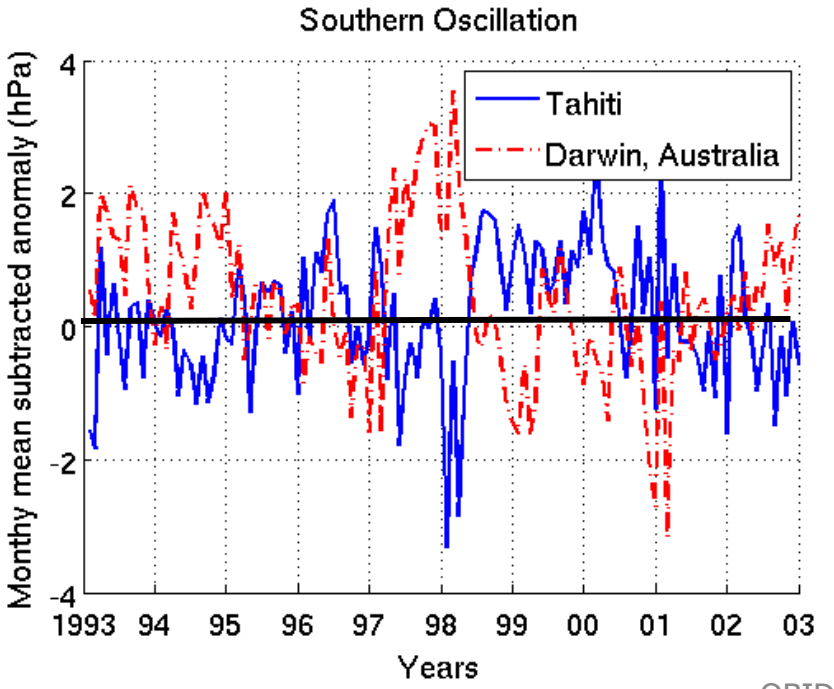


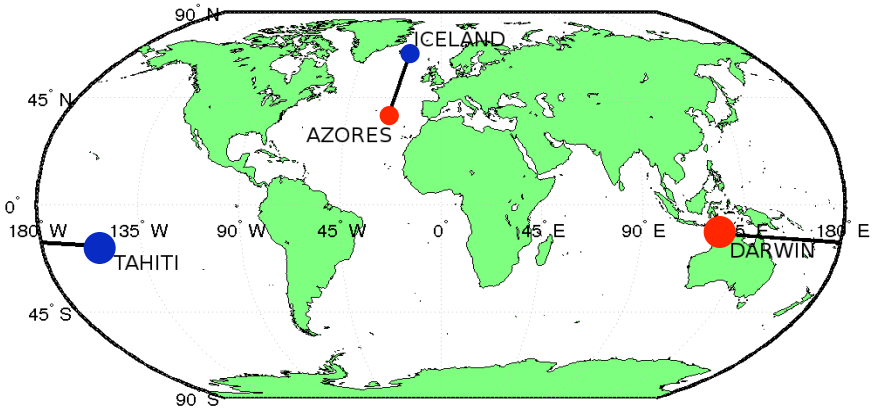


Illustrative Research Tasks:
Relationship Mining in Climate Data

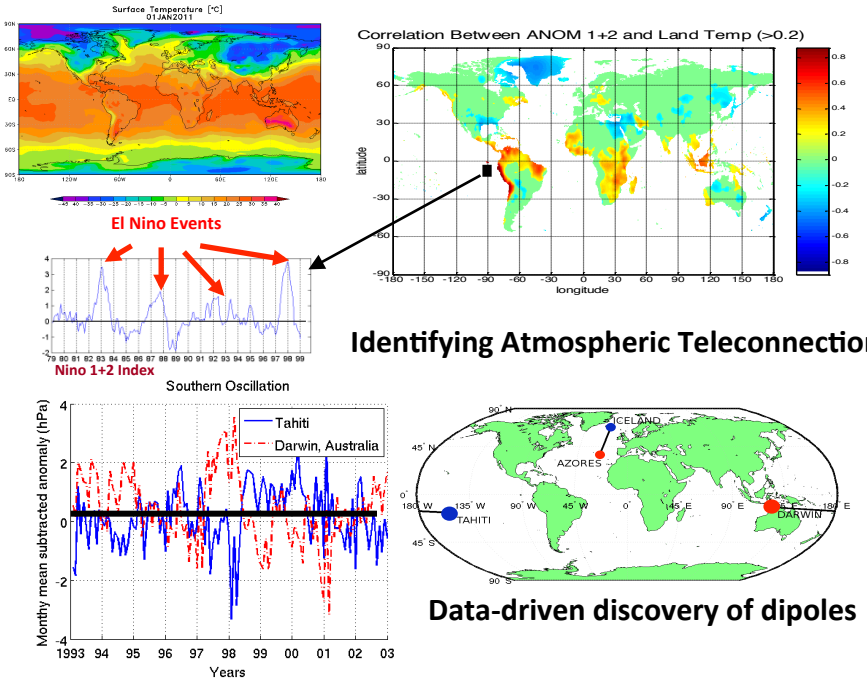


Data-driven discovery of dipoles





Illustrative Research Tasks:
Relationship Mining in Climate Data



Identifying Atmospheric Teleconnections

Data-driven discovery of dipoles

Data Mining Research Tasks

- Identify statistically significant relationships between a variable at a location and a spatio-temporal field
- Detect pairs of regions in space that participate in a spatially coherent and temporally consistent relationship

Challenges

- Non-i.i.d. data (due to spatio-temporal auto-correlation)
- Relationships often exist only in a small number of geographic regions and time intervals
- Massive search space
- Unknown, non-linear, and long-range dependency structure
- High false discovery rate

Earth Science Research Tasks

- Study interactions among land and ocean processes
- Develop predictive insights about terrestrial ecosystem disturbances and weather events

Data

- Reanalysis and climate model datasets
- Observational data from Earth-observing satellites

Illustrative Research Tasks:
Land Cover Change Detection

Illustrative Research Tasks:
Land Cover Change Detection

Earth Science Research Tasks

- Monitor the state of global **forest ecosystems** and identify changes happening as a result of logging and natural disasters.
- Determine the impact of a growing population on **agriculture**, e.g., via creation of new farmlands, changes in cropping patterns, etc.
- Understand the effects of **urbanization** on surrounding ecosystem resources and water supply

Data

- MODIS: Available daily at 250m resolution since Feb 2000
- LANDSAT: Bi-weekly at 30m resolution, since 1972
- Other high resolution datasets with limited spatial and temporal coverage

Data Mining Research Tasks

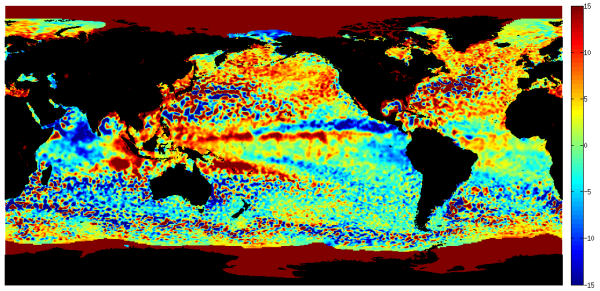
- Change detection in multi-variate spatio-temporal data
- Classify land cover changes occurring in space and time

Challenges

- Presence of noise, missing values, and poor-quality data
- Lack of ground truth
- High temporal variability
- Spatio-temporal auto-correlation
- Spatial and temporal heterogeneity
- Class imbalance (changes are rare events)
- Multi-resolution, multi-scale nature of data

Sample of Research Projects:

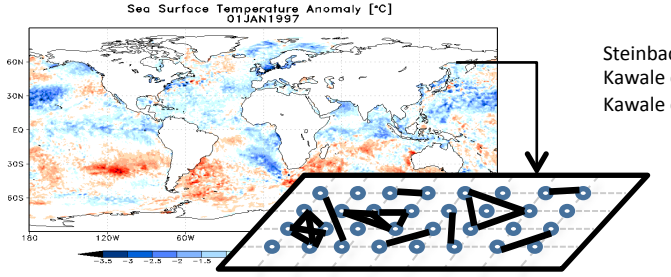
NSF Expeditions Project on Understanding Climate Change: A Data-driven Approach



Faghmous et al. 2012 a,b,
Faghmous et al. 2013 a,b

Pattern Mining:
Monitoring Ocean Eddies

- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



Steinbach et al. 2003
Kawale et al., 2011 a,b
Kawale et al., 2012

http://climatechange.cs.umn.edu/
http://gopher.cs.umn.edu/

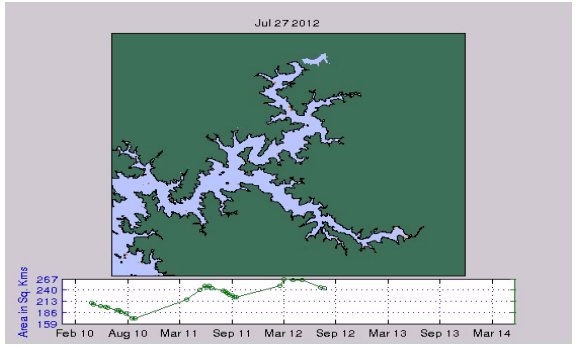
Network Analysis:
Climate Teleconnections

- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections
- Also applicable in analyzing brain fMRI data

Chen et al. 2013 a,b, 2014
Karpatne et al. 2012, 2014
Mithal et al. 2011, 2013, 2014
Chamber et al. 2011
Garg et al. 2011 a,b
Boriah et al. 2008, 2010 a,b
Potter et al. 2003, 2004 a,b, 2005
Potter et al. 2006, 2007, 2008

Change Detection:
Monitoring Ecosystem Distrubances

- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in vegetation, e.g. fires, deforestation, and insect damage



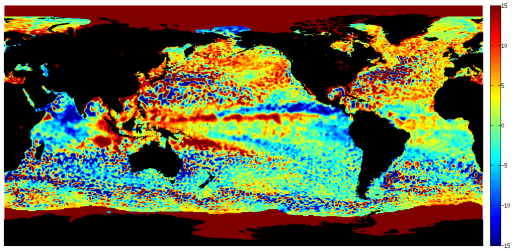
Chen et al. 2014

Classification:
Mapping Water Dynamics

- Physics-guided classification approaches that adhere to physical constraints
- Global scale mapping of lake water bodies

Sample of Research Projects:

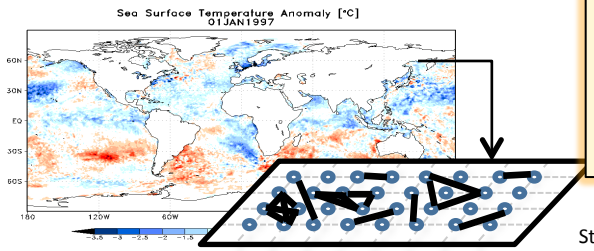
NSF Expeditions Project on Understanding Climate Change: A Data-driven Approach



Faghmous et al. 2012 a,b,
Faghmous et al. 2013 a,b

Pattern Mining:
Monitoring Ocean Eddies

- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



Network Analysis:
Climate Teleconnections

- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections
- Also applicable in analyzing brain fMRI data

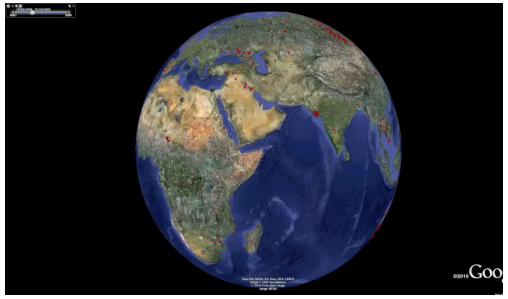
Steinbach et al. 2003
Kawale et al., 2011 a,b
Kawale et al., 2012

Highlights:

- Highly inter-displicinary
  - Computer science, hydrology, Earth sciences, statistics, civil engineering
- Dozens of publications (journals, conferences, and workshops) with authors from multiple disciplines
- Public release of software & data products
- Advances in computer science driven by Earth science applications
- Advances in Earth sciences using computer science methods
- Development of physics-guided data mining paradigm

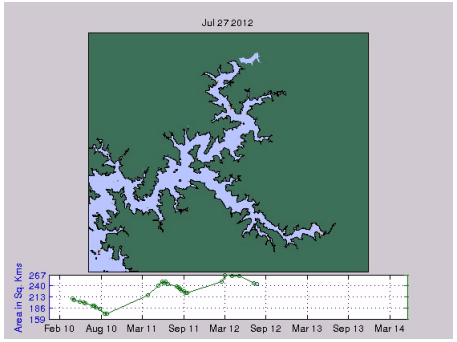
Chen et al. 2014

Chen et al. 2013 a,b, 2014
Karpatne et al. 2012, 2014
Mithal et al. 2011, 2013, 2014
Chamber et al. 2011
Garg et al. 2011 a,b
Boriah et al. 2008, 2010 a,b
Potter et al. 2003, 2004 a,b, 2005
Potter et al. 2006, 2007, 2008



Change Detection:
Monitoring Ecosystem Distrubances

- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in vegetation, e.g. fires, deforestation, and insect damage



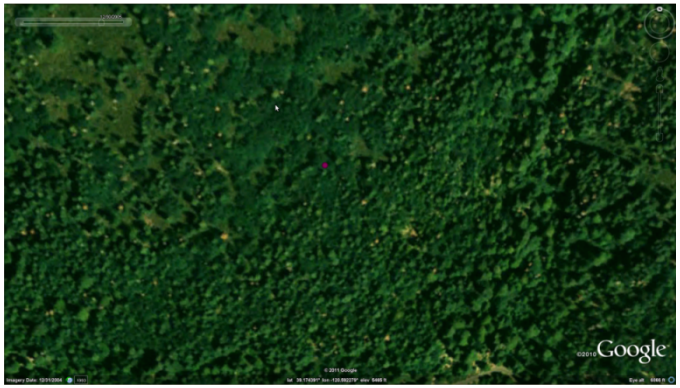
Classification:
Mapping Water Dynamics

- Physics-guided classification approaches that adhere to physical constraints
- Global scale mapping of lake water bodies

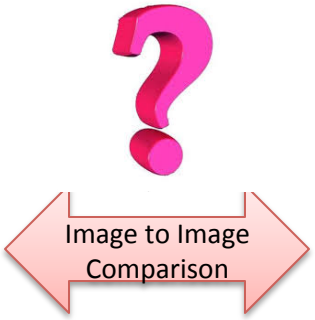
http://climatechange.cs.umn.edu/
http://gopher.cs.umn.edu/

Case Study 1:
Monitoring Ecosystem Changes

Monitoring Ecosystem Disturbances: Traditional Approach



2005



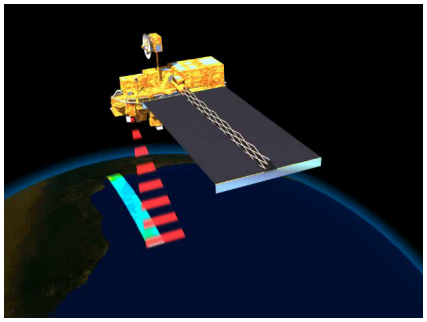
2009

- Requires high-quality imagery
  - Available infrequently
- Requires high resolution
  - No global coverage
- Requires training data
  - Must be created manually
  - Labor-intensive, time-consuming, expensive

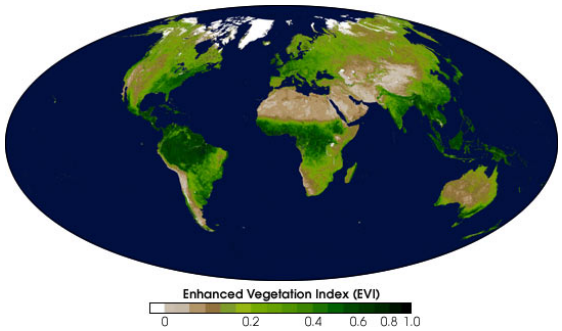
Studies are limited to small regions and unable to identify change point of rate of change

Global Monitoring of Ecosystem Disturbances

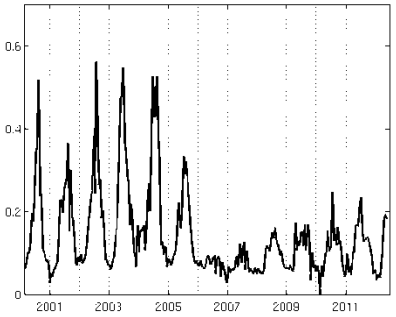
MODIS: Daily product at 250m spatial resolution available daily since February 2000



MODIS instrument on
NASA Aqua/Terra Satellites



A **vegetation index** measures the surface "greenness" – proxy for total biomass



This vegetation **time series** captures temporal dynamics around the site of the China National Convention Center

LANDSAT: Bi-weekly at 30m resolution, since 1972

Multi-spectral data

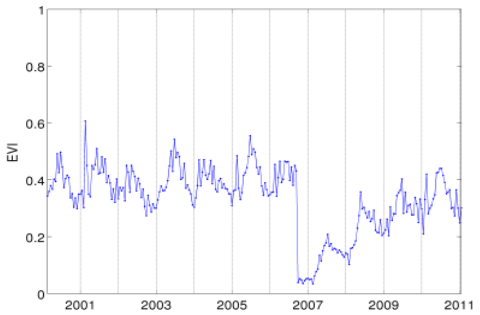
- Provides frequent global coverage
- (Relatively) coarse resolution
- Sometimes poor quality
  - Noisy
  - Missing data

Opportunities and challenges for **spatio-temporal data mining**

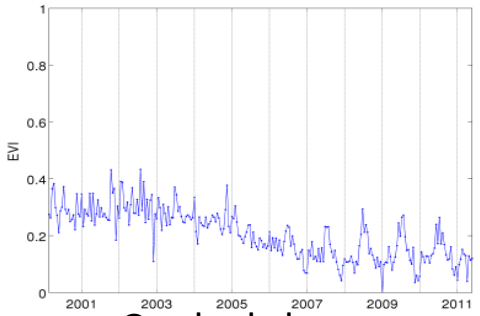
Novel Techniques For Global Change Detection

- Robust to missing data, noise and outliers
- Model variability and spatial heterogeneity
- Low false discovery rate
- Aware of spatial context
- Make use of multi-scale and multi-variate data
- Exploit limited training data available

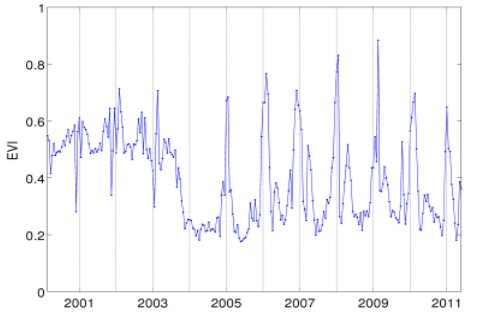
Potter 2003,2005,2007
Boriah 2008, Boriah 2010, Mithal 2011, Garg 2011,
Chamber 2011, Chen 2012, Mithal 2012, Karpatne 2012, Mithal 2013*



Abrupt change



Gradual change



Model change

ALERTS: Automated Land change Evaluation, Reporting and Tracking System

- **Planetary Information System** for interactive investigation of ecosystem disturbances discovered by GOPHER
  - Forest Fires
  - Deforestation
  - Droughts
  - Urbanization
  - …
- Helps quantify **carbon impact** of changes, understand the relationship between climate variability and human activity
- Provides **ubiquitous web-based access** to changes occurring across the globe, creating public awareness

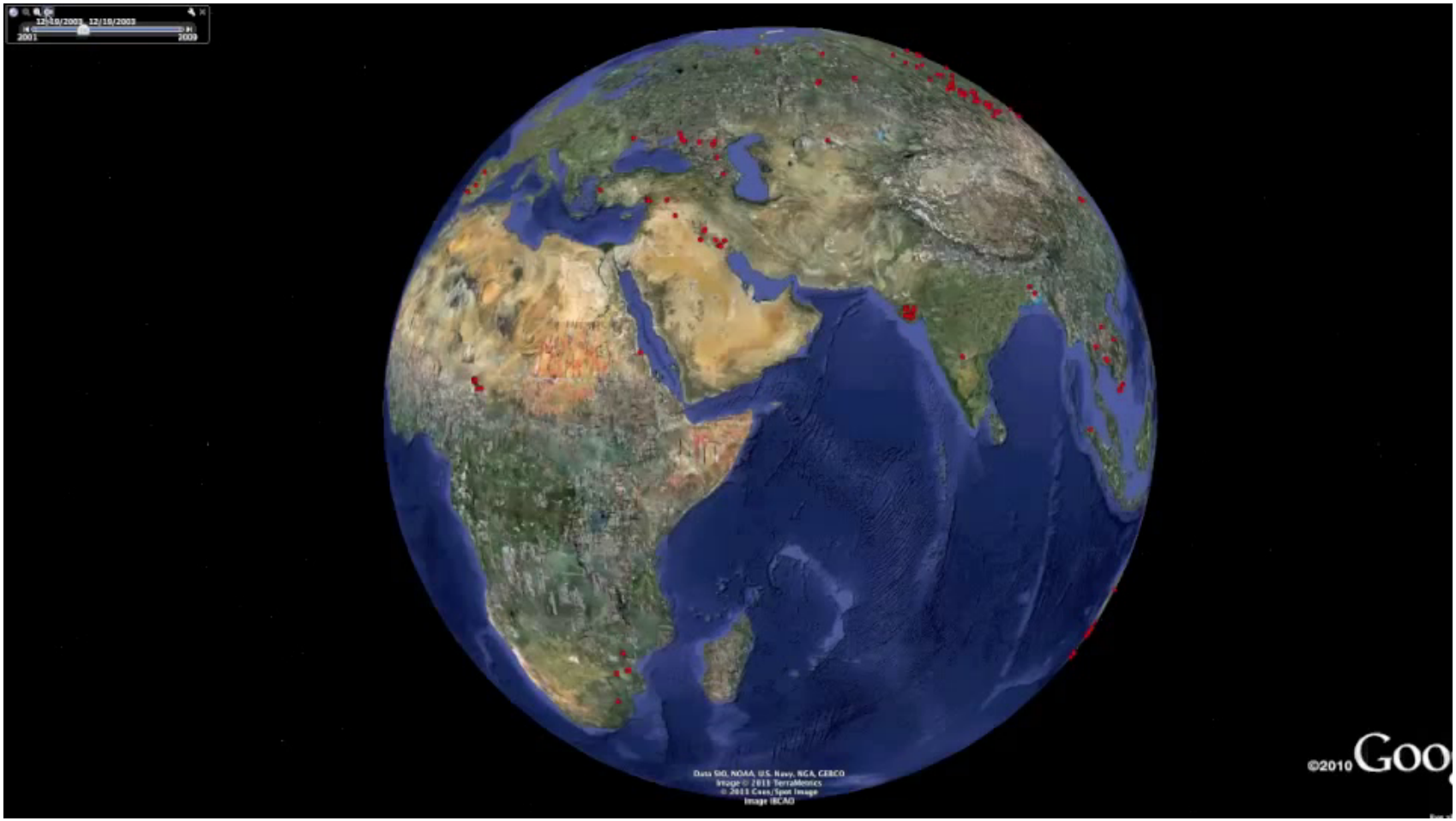
TIME The 50 Best Inventions of 2009

The 50 Best Inventions of 2009 > The Best Inventions

The Planetary Skin

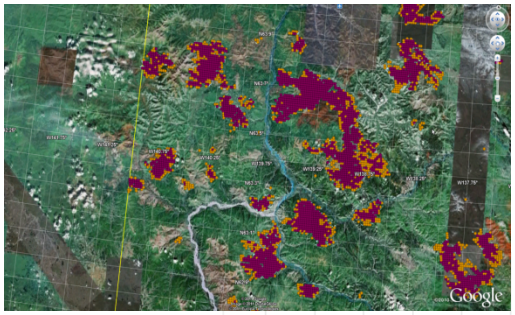
What happens to Earth when a forest is razed or energy use soars? We don't know because environmental data are collected by isolated sources, making it impossible to see the whole picture. With the theory that you can't manage what you can't measure, NASA and Cisco have teamed up to develop Planetary 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.

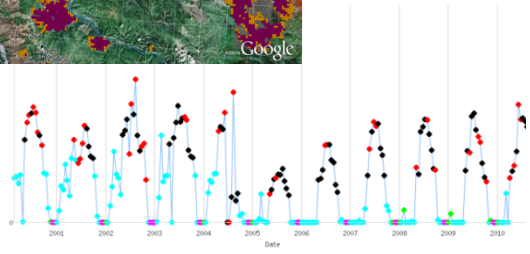
Global Change Points



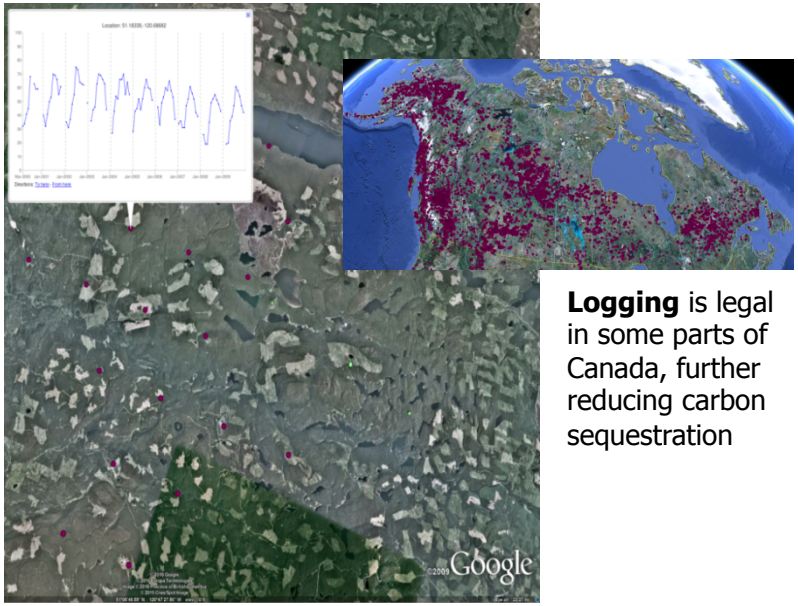
Northern Hemisphere Changes

Illustrative Examples

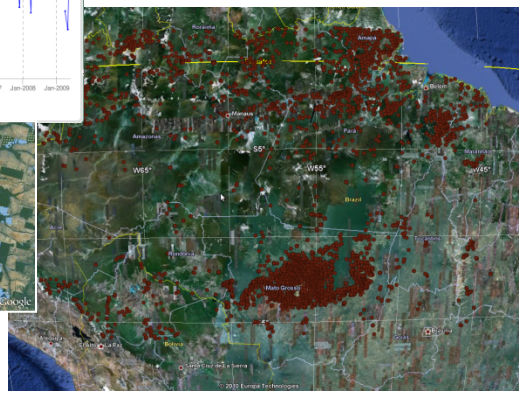




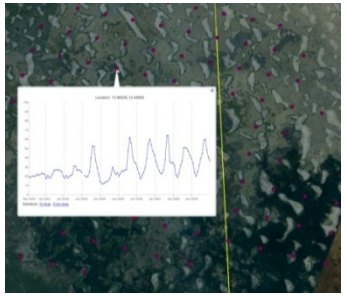
Large **forest fires in Canada** release significant amounts of carbon into the atmosphere.



Logging is legal in some parts of Canada, further reducing carbon sequestration

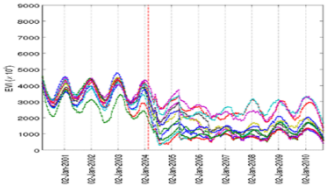


Brazil Accounts for almost 50% of all humid **tropical forest clearing**, nearly 4 times that of the next highest country.

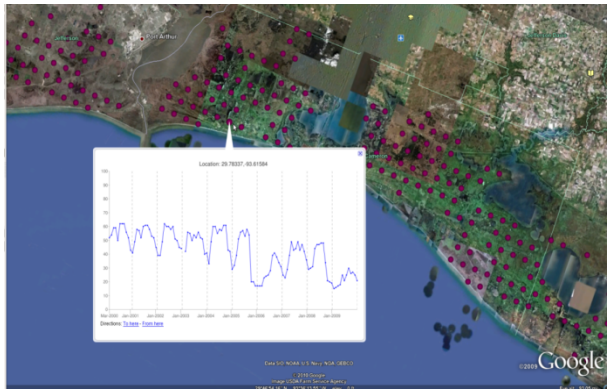


Lake Chad (Nigeria) shrunk by as much as 90% over the past two decades.

Illustrative Examples

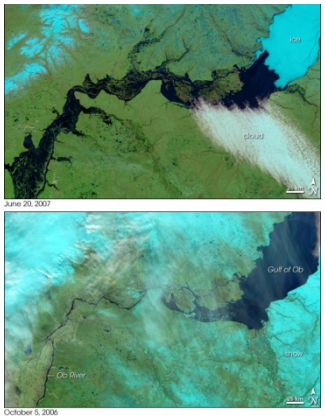


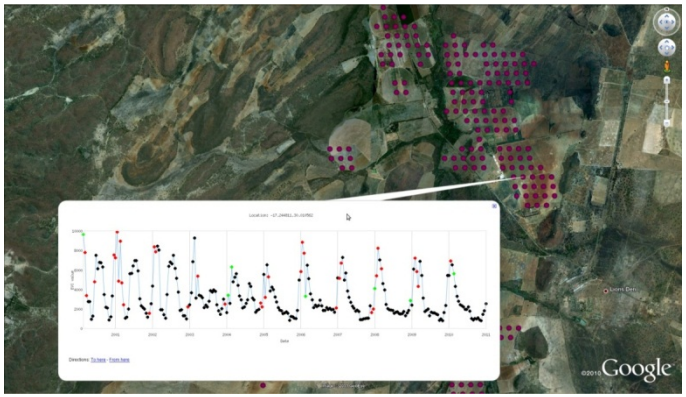
The multi-phase construction of an **illegal gold mine** inside a protected forest in Tanzania is seen quite clearly.



Hurricane Katrina caused significant damage and vegetation loss along the US Gulf Coast.

One winter the **Ob River** caused a massive **flooding** due to the unusual extent of freezing in the Bay of Ob / Kara Sea.

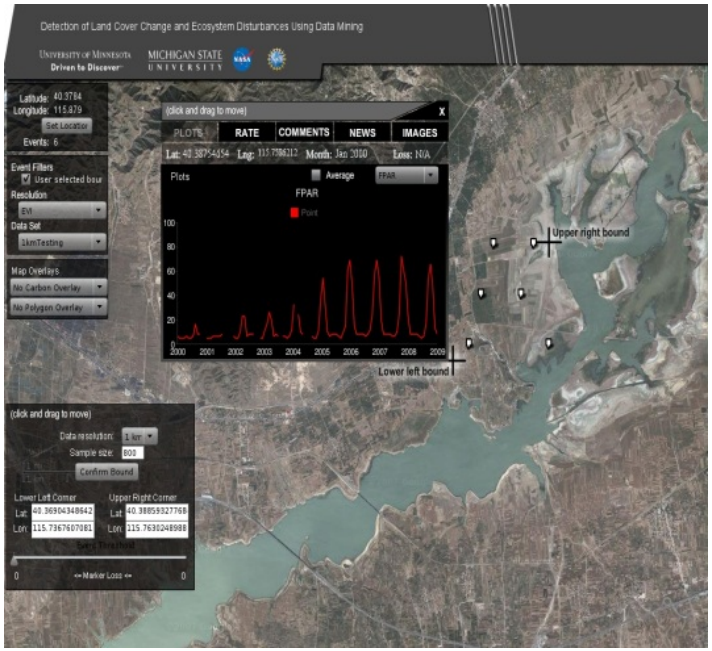


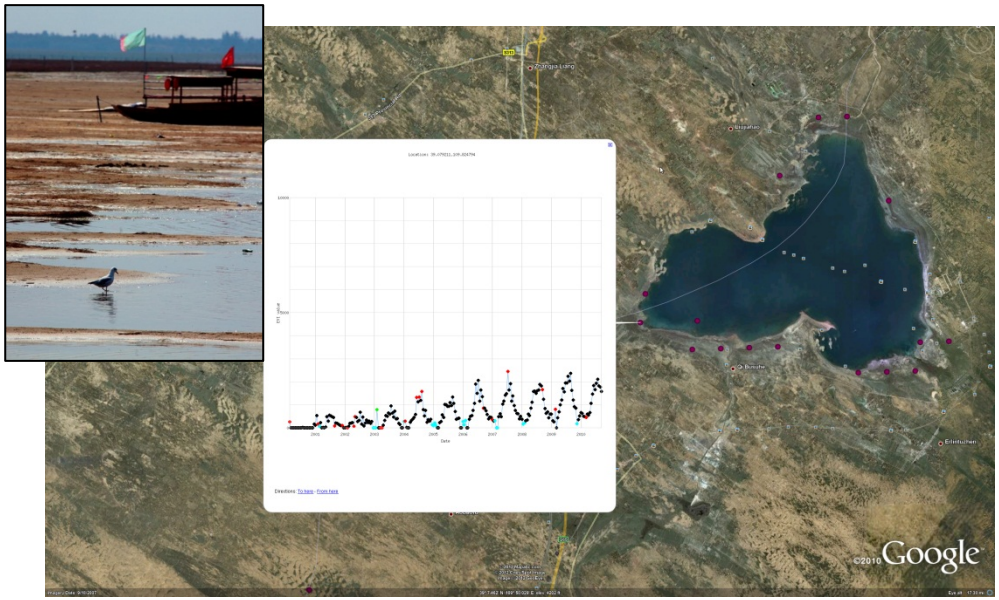


Political conflict and the ensuing "land reform" resulted in wide-spread **farm abandonment** and loss of productivity in **Zimbabwe** between 2004 and 2008.

Illustrative Examples in China

Examples of **afforestation** can be seen in several areas around the world, including this region **near Beijing** where new trees have been planted to prevent dust storms and erosion.

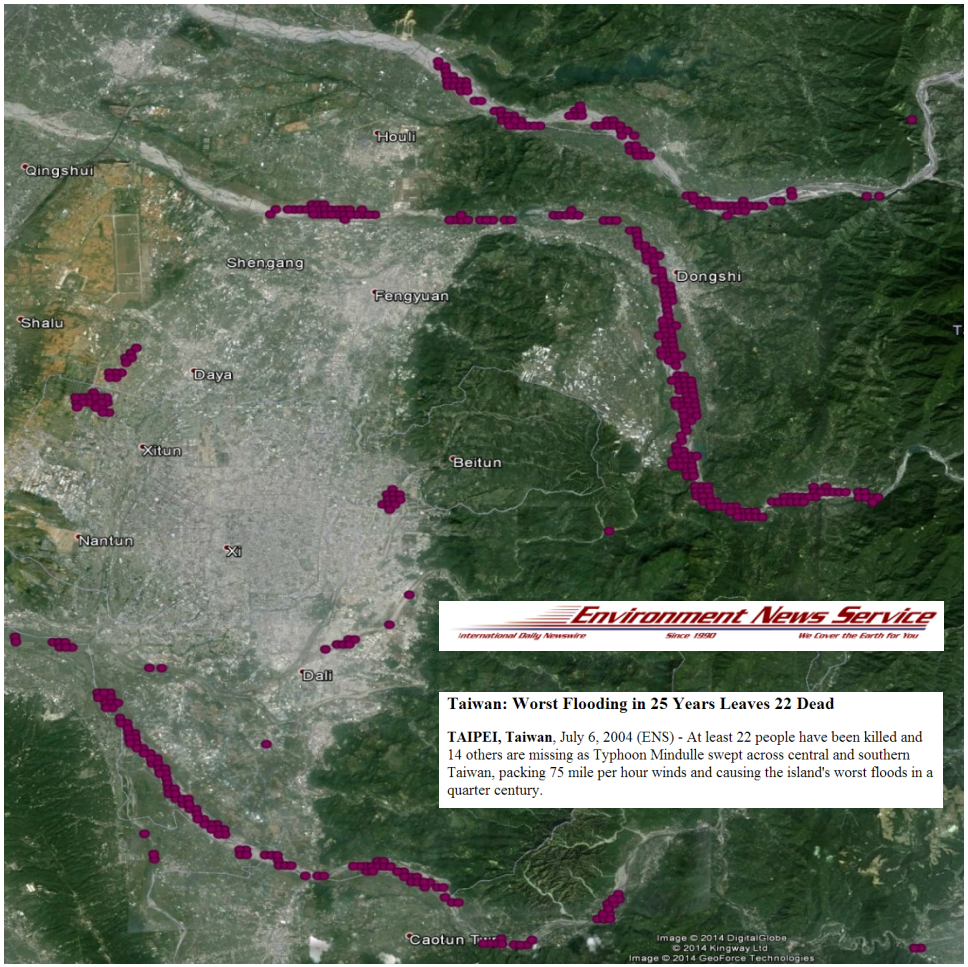




Vegetation growth is detected along the shores of a shrinking **Lake Hongjiannao**, the largest desert lake in China. Water levels have been rapidly decreasing and fish populations are dwindling; experts predict that the lake may vanish within a decade.

Source: Want China Times

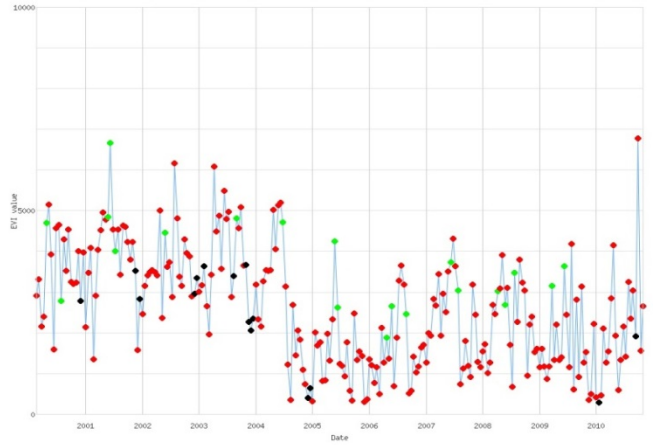
Illustrative Examples in Taiwan (2004)



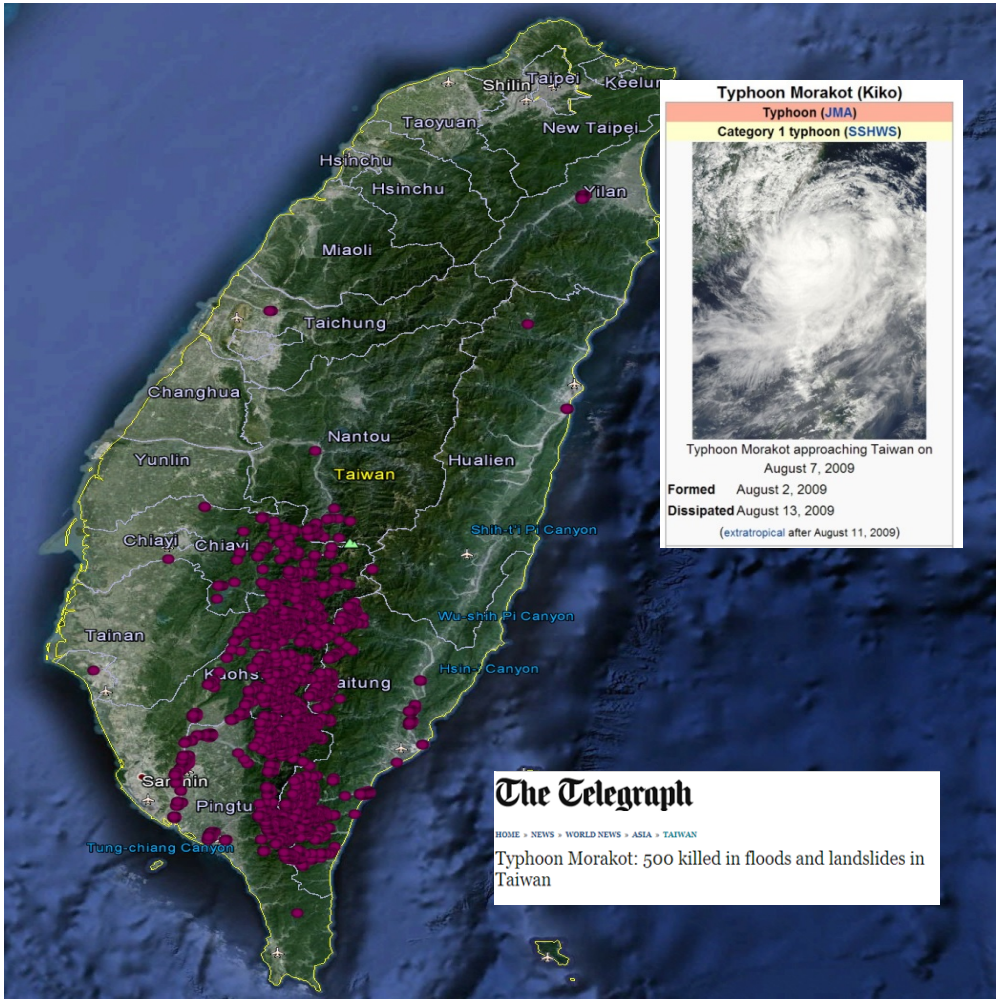
Environment News Service
International Daily Newswire Since 1990 We Cover the Earth for You

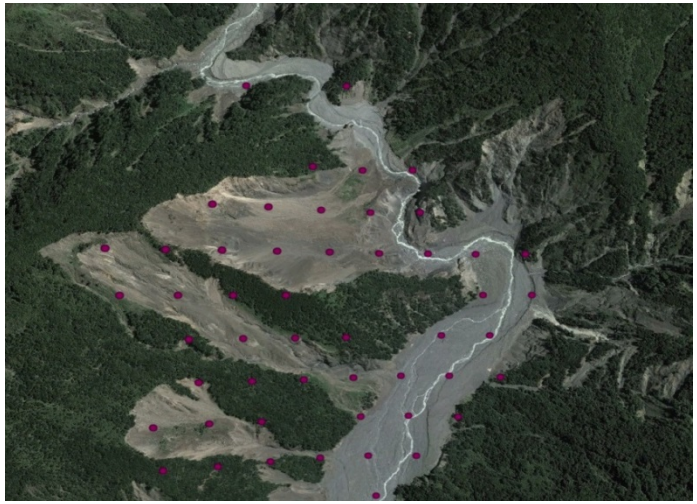
Taiwan: Worst Flooding in 25 Years Leaves 22 Dead

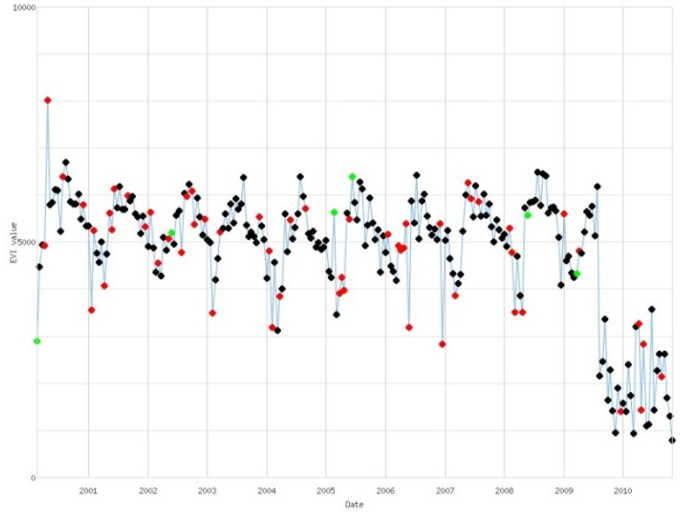
TAIPEI, Taiwan, July 6, 2004 (ENS) - At least 22 people have been killed and 14 others are missing as Typhoon Mindulle swept across central and southern Taiwan, packing 75 mile per hour winds and causing the island's worst floods in a quarter century.



Illustrative Examples in Taiwan (2009)







Impact on REDD+

"The [Peru] government needs to spend more than $100m a year on high-resolution satellite pictures of its billions of trees.
But … a computing facility developed by the Planetary Skin Institute (PSI) … might help cut that budget."

"ALERTS, which was launched at Cancún, uses … **data-mining** algorithms developed at the **University of Minnesota** and a lot of computing power … to spot places where land use changed."

(The Economist 12/16/2010)

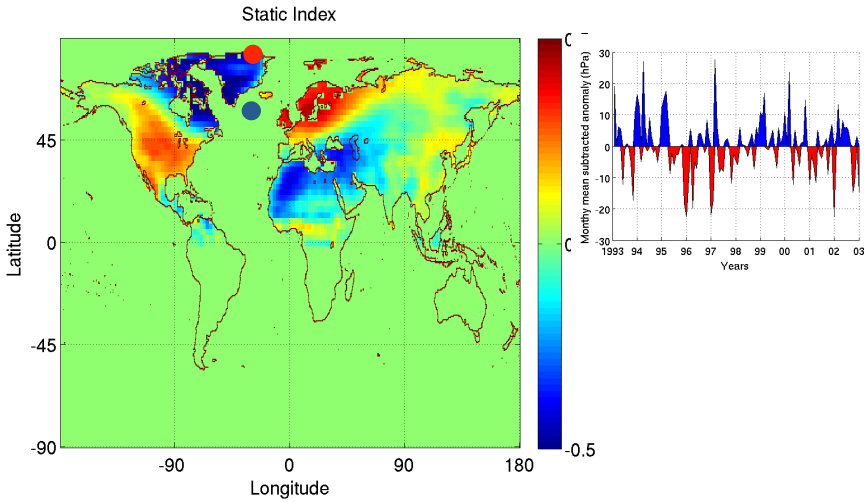
Case Study 2: Data-driven Discovery of Atmospheric Dipoles

Automated Discovery of Dipoles

Importance of Dipoles:

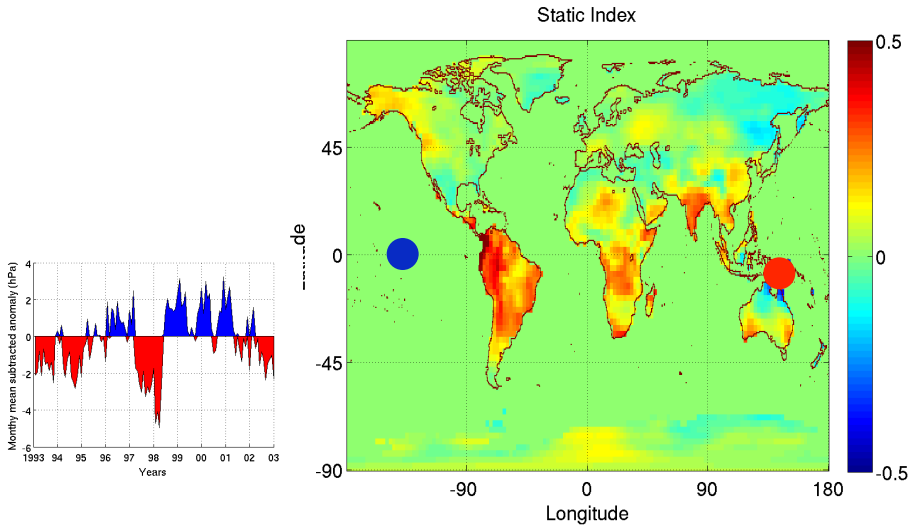
Crucial for understanding the climate system and are known to cause temperature and precipitation anomalies throughout the globe.

NAO influences sea level pressure (SLP) and temperature over the Northern Hemisphere.



Correlation of land temperature anomalies with NAO

SOI strongly influences global climate variability.



Correlation of land temperature anomalies with SOI

List of Major Climate Oscillations

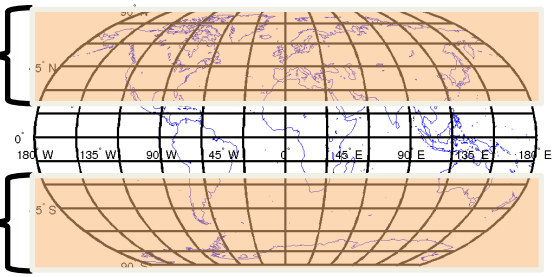
Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti. It has a period averaging 2.33 years and is analysed as a part of an ENSO event.
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the first principal component of SLP northward of 20° N
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
PNA	**Pacific North American:** SLP Anomalies over the North Pacific Ocean and the North America
AAO	**Antarctic Oscillation:** Defined as the first principal component of SLP southward of 20° S

Discovered primarily by human observation or by EOF analysis.

van Loon & Rogers, 1978
Wallace & Gutzler, 1981
von Storch & Zwiers, 2002

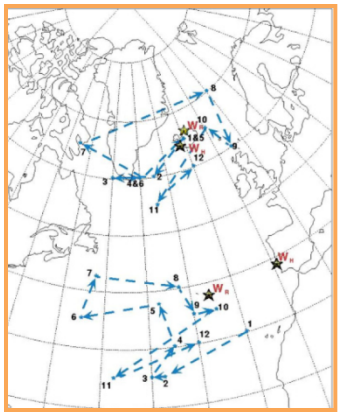
AO: EOF Analysis of 20N-90N Latitude

AAO: EOF Analysis of 20S-90S Latitude



Motivation for Automated Discovery of Dipoles

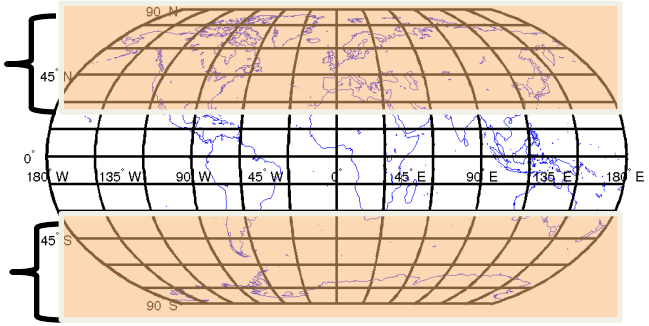
- The known dipoles are defined by static locations but the underlying phenomenon is dynamic
- Manual discovery can miss many dipoles
- EOF and other types of eigenvector analysis finds the strongest signals and the physical interpretation of those can be difficult.



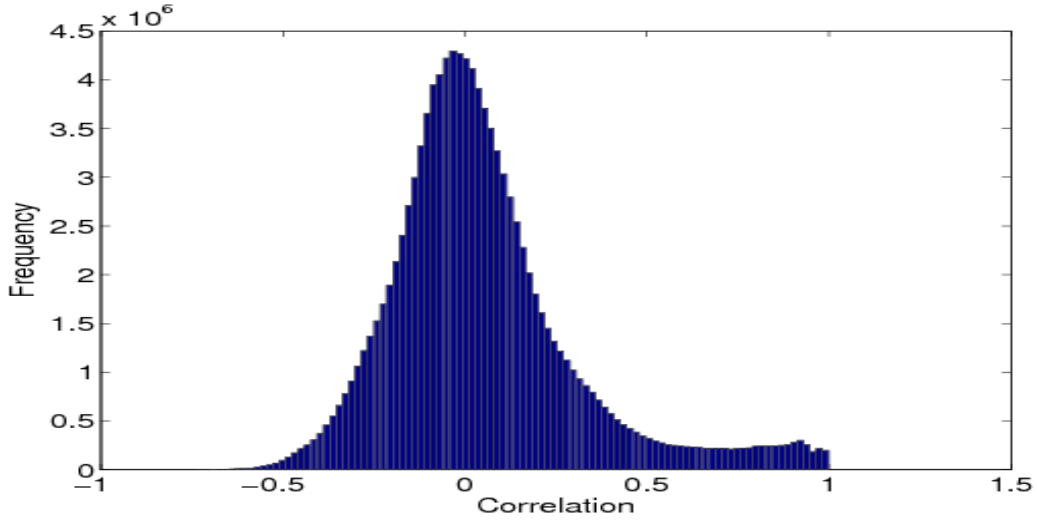
Dynamic behavior of the high and low pressure fields corresponding to NOA climate index (Portis et al, 2001)

AO: EOF Analysis of 20N-90N Latitude

AAO: EOF Analysis of 20S-90S Latitude

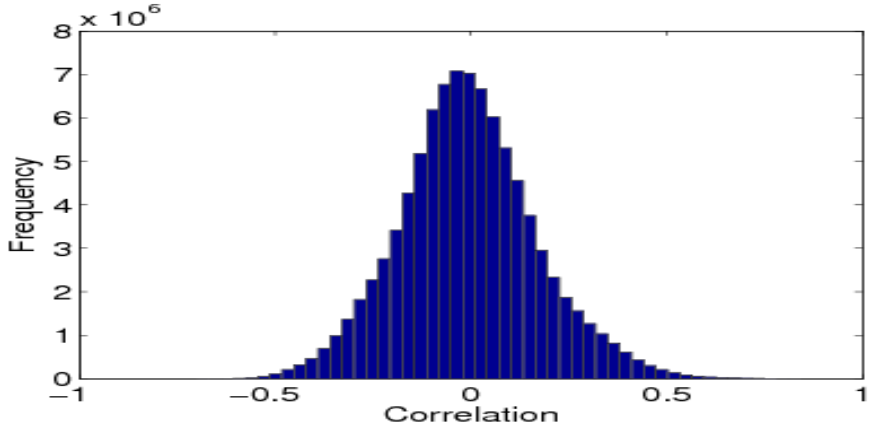


Challenges in Automated Discovery of Dipoles

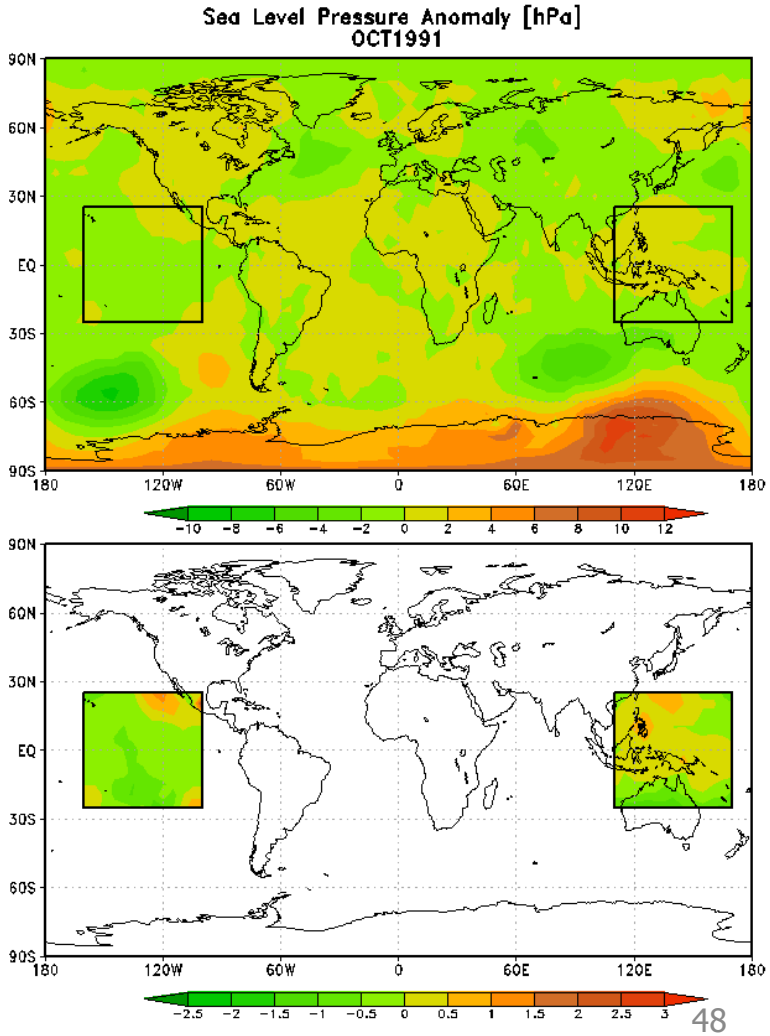


The distribution of pair-wise correlations of anomaly time series at locations in a 2.5 x 2.5 degree grid (approx. 10000 locations)

Challenges in Automated Discovery of Dipoles



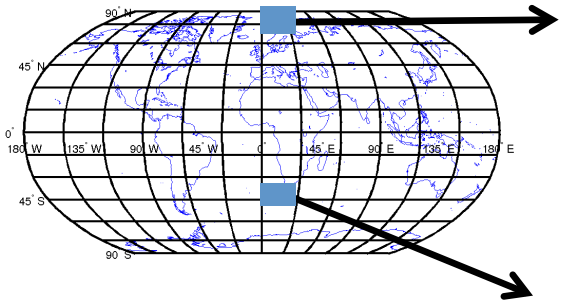
The distribution after considering only those pairs of lcoations that are at least 5000 km apart



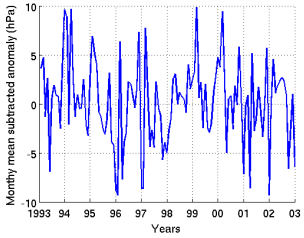
- Variability in the strength of different relationships across the globe
- Number of locations involved in significant positive and negative relationships is too large

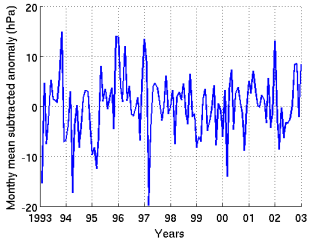
Consistency in space and time is key to reduce the search space

Graph-Based Approach for Dipole Discovery

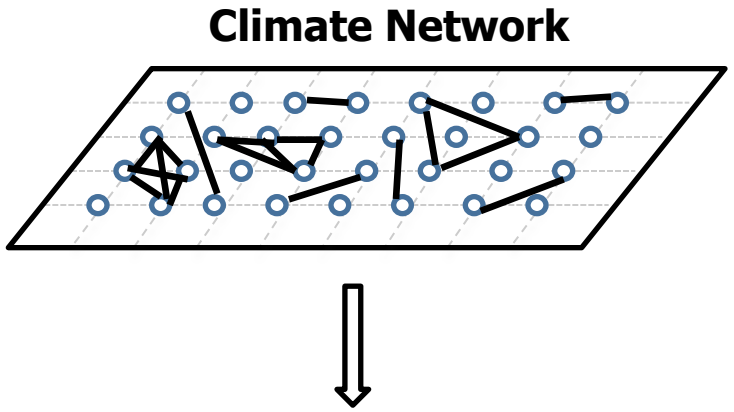


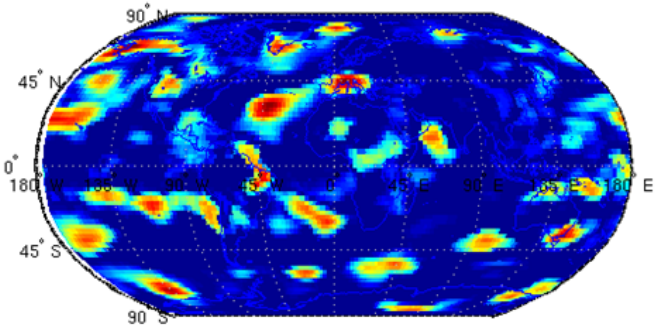
Nodes in the Graph correspond to grid points on the globe.





Edge weight corresponds to correlation between the two anomaly time series



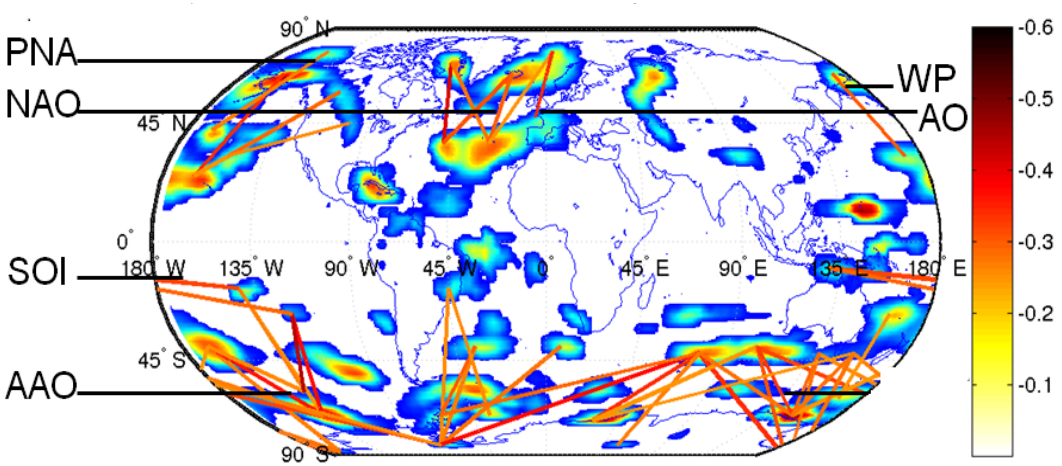


Potential Regions that can serve as dipole ends are identified as sets of locations that share reciprocal negative and positive neighbors.

Steinbach et al., 2003
Tsonis et al., 2004, 2006
Donges et al., 2009a,b
Kawale et al., 2011

Benefits of Automatic Dipole Discovery

- Detection of Global Dipole Structure
  - Most known dipoles discovered
  - New dipoles may represent previously unknown phenomenon.
  - Enables analysis of relationships between different dipoles
- Location based definition possible for some known indices that are defined using EOF analysis.
- Dynamic versions are often better than static
- Dipole structure provides an alternate method to analyze GCM performance



CIDU'11: Best Student Paper Award

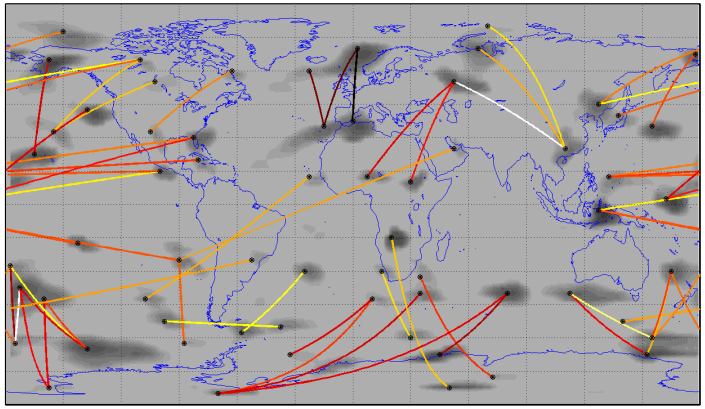
SC'11: Explorations in Science through Computation Award

Grace Hopper'12: Best Poster Award (Winner of the ACM Student Research Competition)

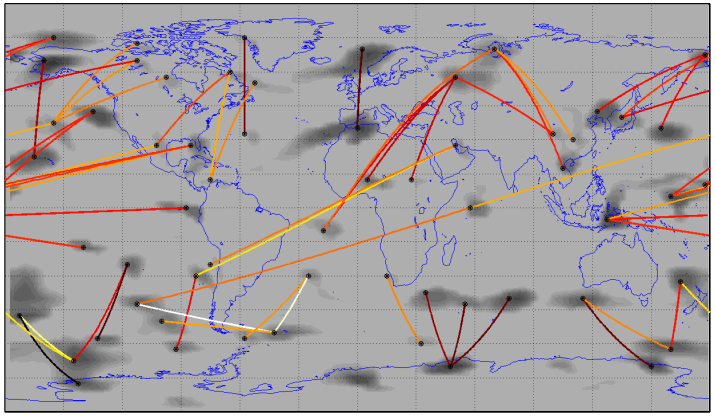
Kawale et al., 2011a,b, 2012

Comparing Dipole Structure in Historical (Reanalysis) Data

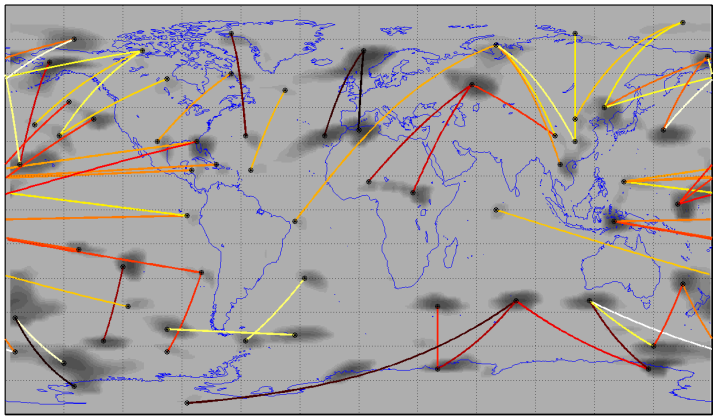
NCEP 1979-2000



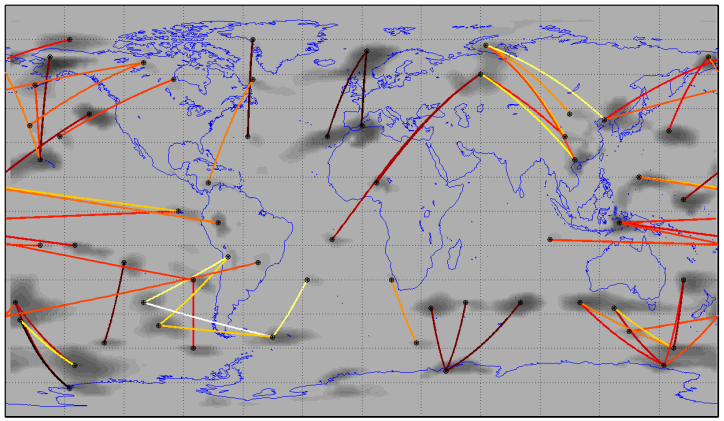
ERA-Interim 1979-2000

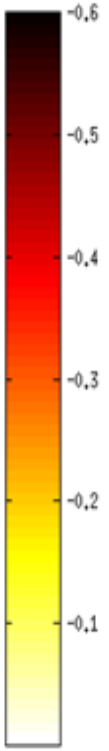


JRA-25 1979-2000

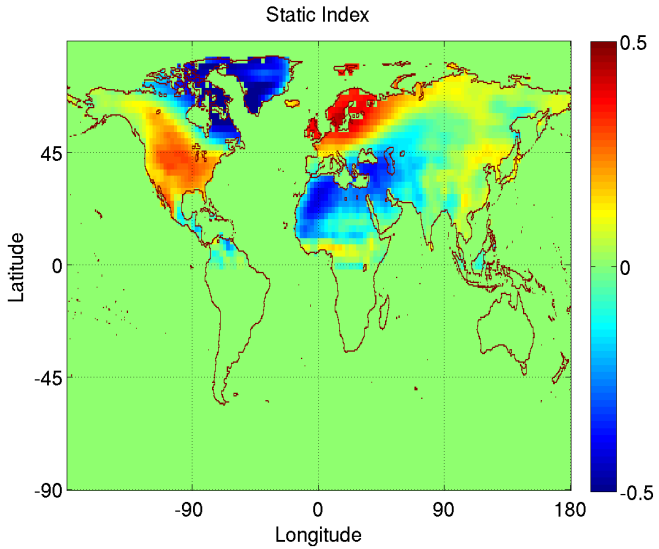


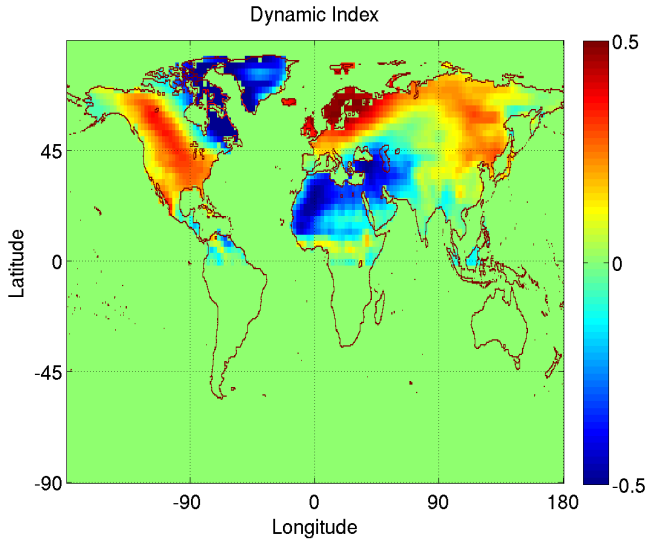
MERRA 1979-2000





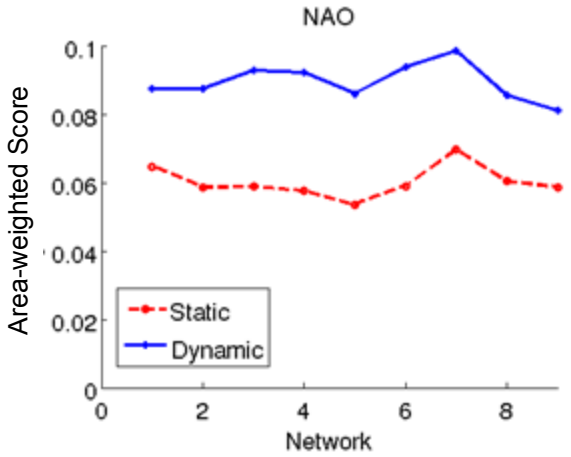
Static vs Dynamic NAO Index - Impact on land temperature



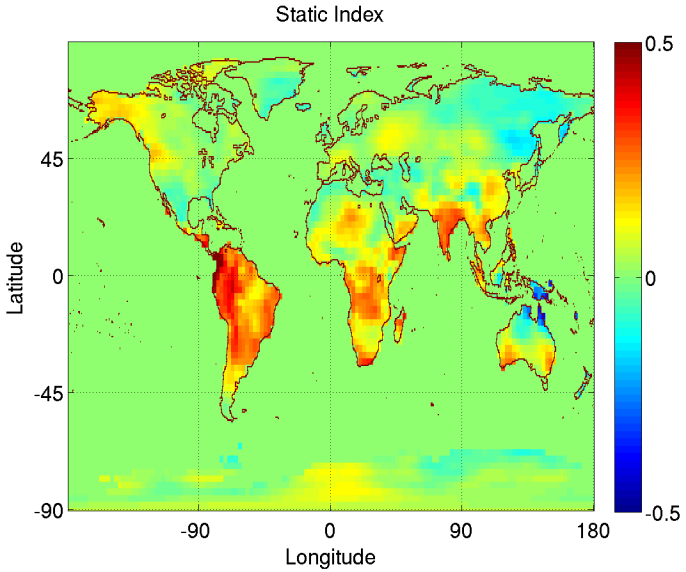


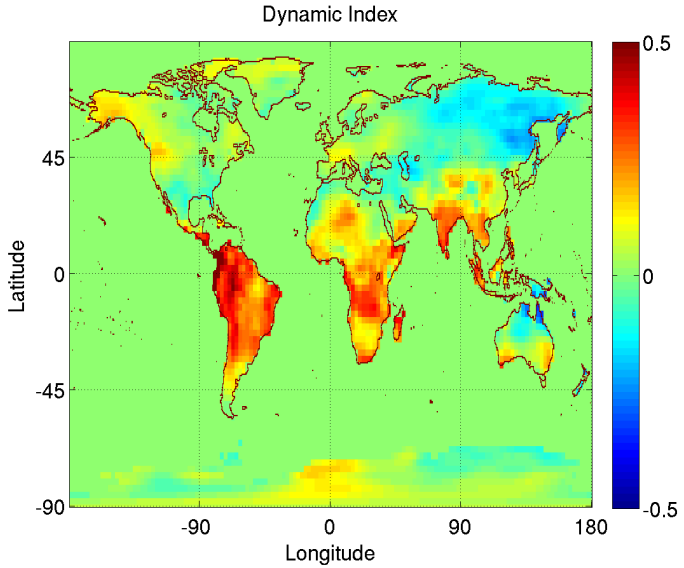
The dynamic index generates a stronger impact on land temperature anomalies as compared to the static index.

Figure to the right shows the aggregate area weighted correlation for networks computed for different 20 year periods during 1948-2008.



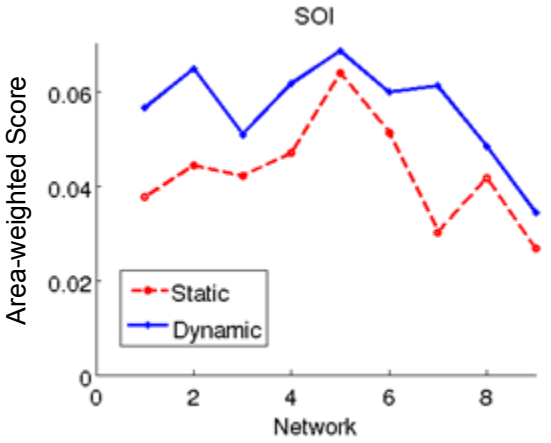
Static vs Dynamic NAO Index - Impact on land temperature



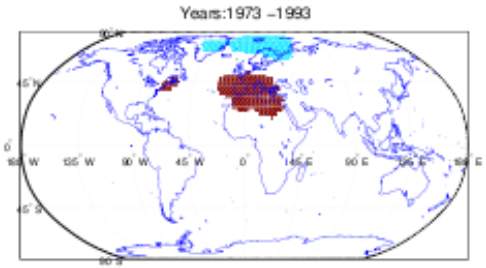


The dynamic index generates a stronger impact on land temperature anomalies as compared to the static index.

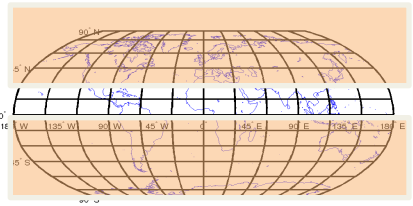
Figure to the right shows the aggregate area weighted correlation for networks computed for different 20 year periods during 1948-2008.



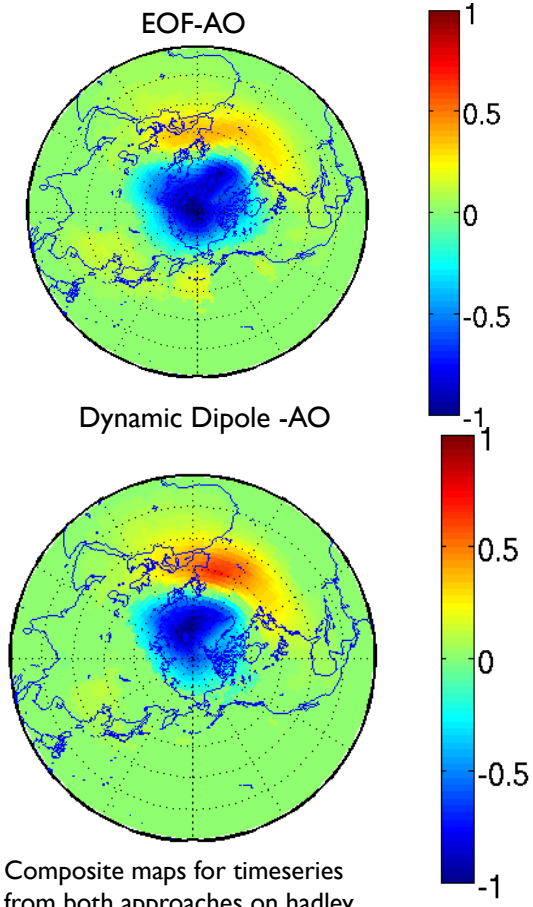
Location Based definition of AO



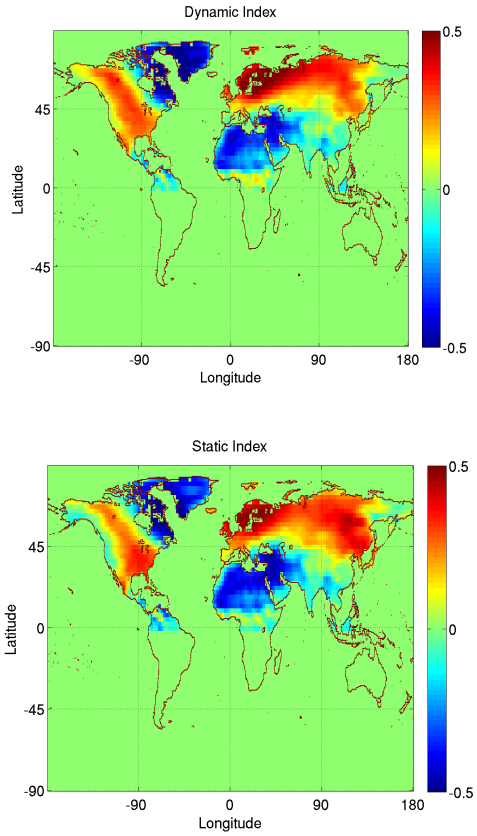
Static AO: EOF Analysis of 20N-90N Latitude



- Mean Correlation between static and dynamic index: 0.84
- Impact on land temperature anomalies comparatively same using static and dynamic index

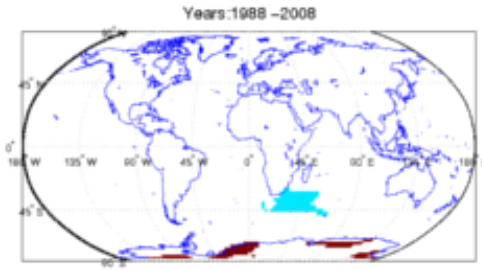


Composite maps for timeseries from both approaches on hadley center SLP data (1979-2011).

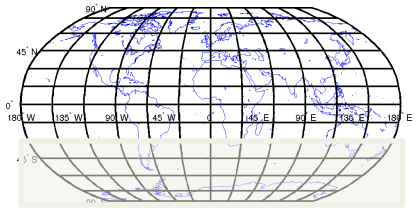


Impact on Land temperature Anomalies using Static and Dynamic AO

Location Based definition of AAO

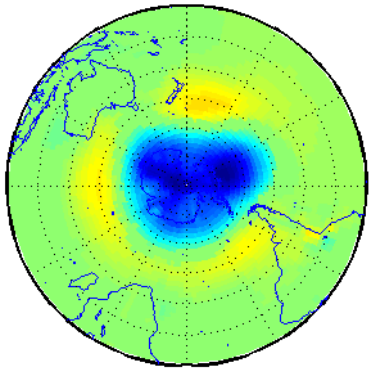


Static AAO: EOF Analysis of 20S-90S Latitude

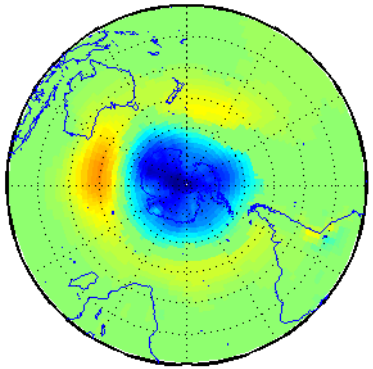


- Mean Correlation between Static and Dynamic index = 0.88
- Impact on land temperature anomalies comparatively same using static and dynamic index

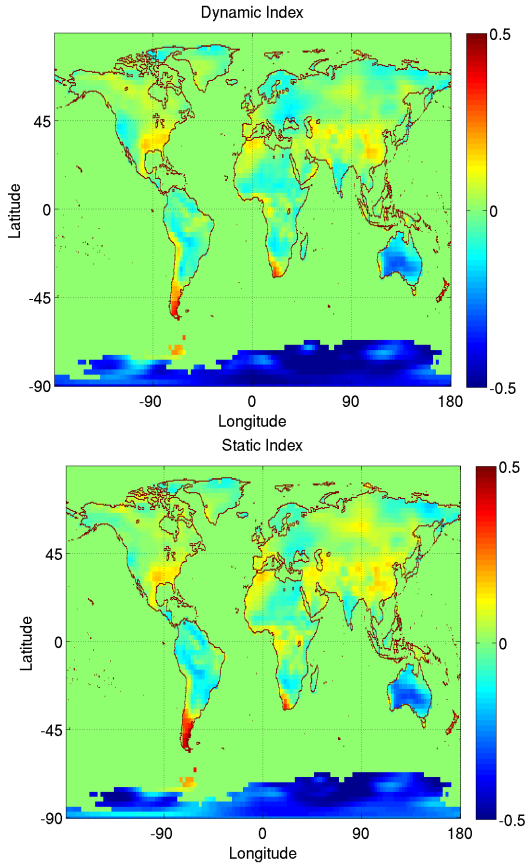
EOF-AAO



Dynamic Dipole -AAO



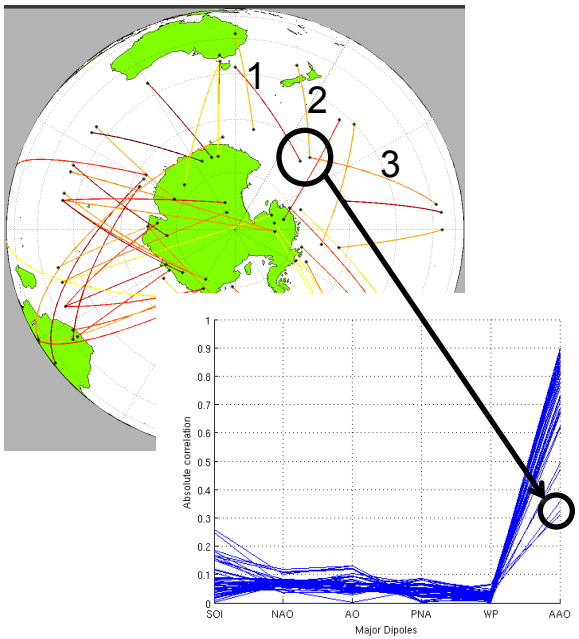
Composite maps for timeseries from both approaches on hadley center SLP data (1979-2011).



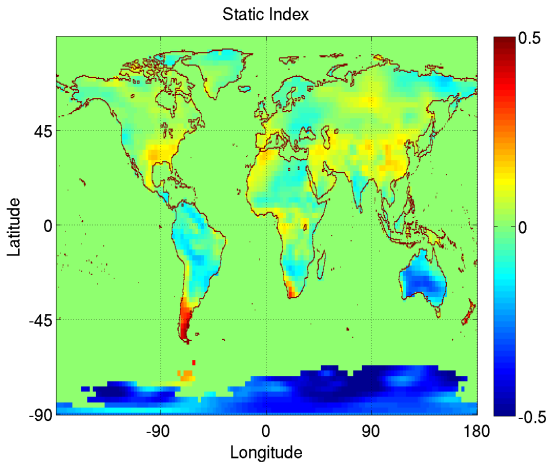
Impact on Land temperature Anomalies using Static and Dynamic AAO

A New Dipole near Australia?

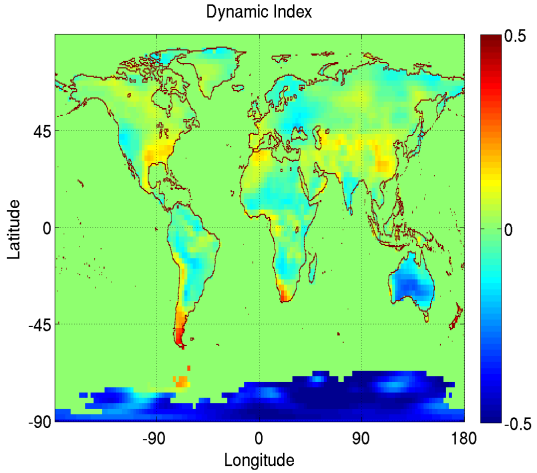
- Comparison of dipoles by looking at land temperature impact.
- Significant difference between the AAO impact and that due to dipoles 1,2,3 which are similar.



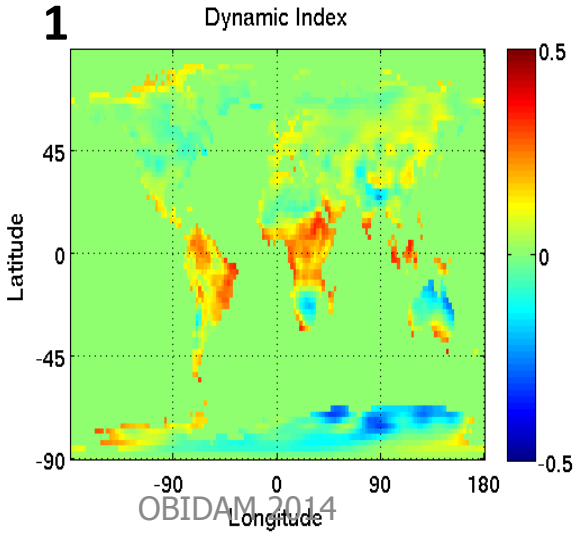
AAO



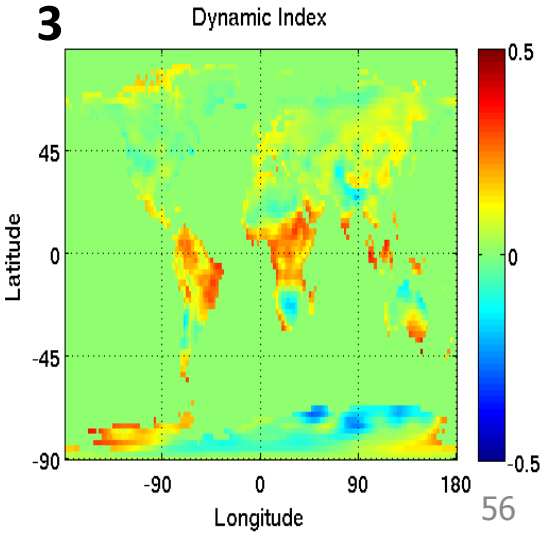
AAO



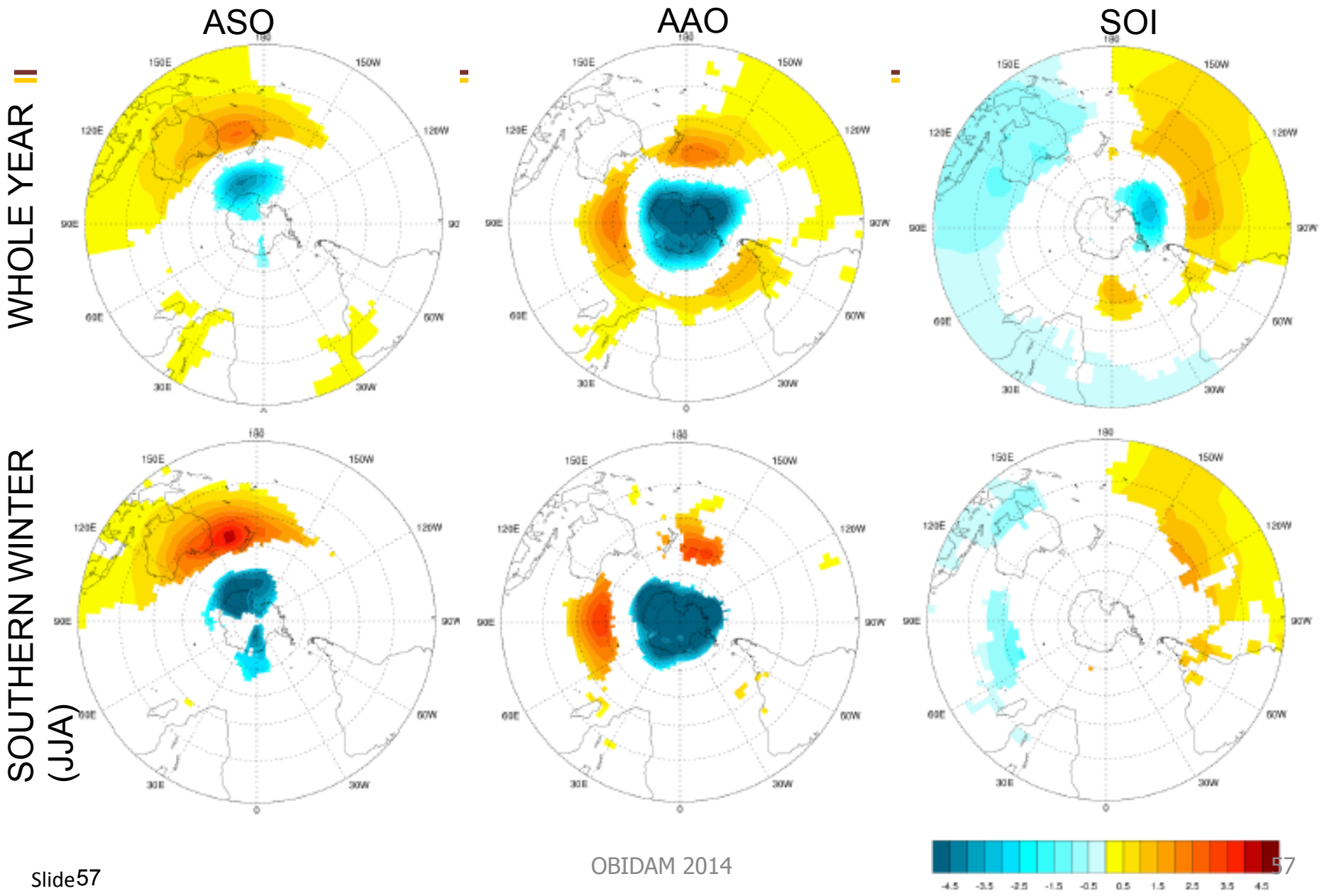
1



3



Composites of ASO dipole from Hadley center SLP data on Hadley center SLP at 95% confidence

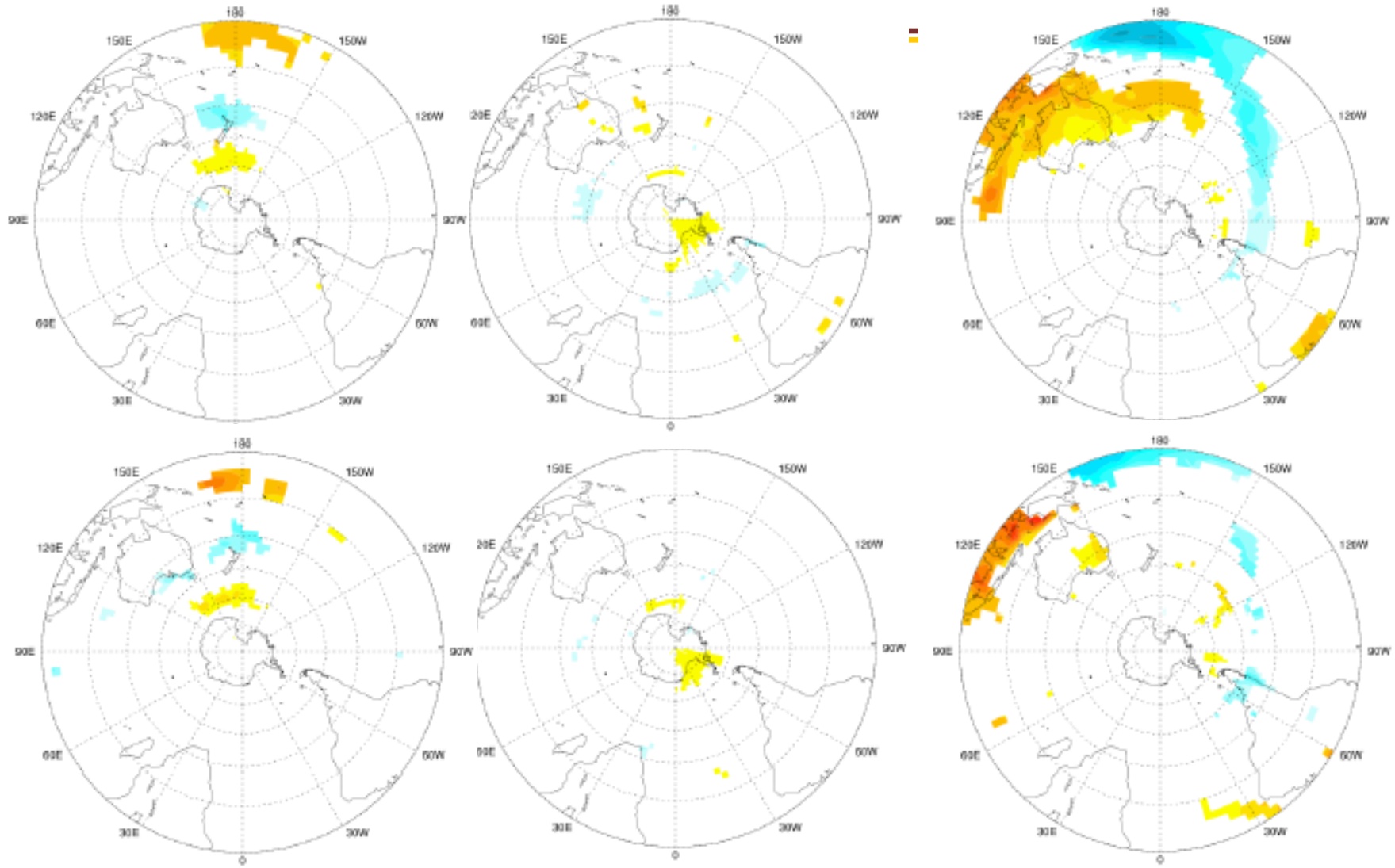


Composites of ASO dipole from Hadley center SLP data on GPCP precipitation data at 95% confidence

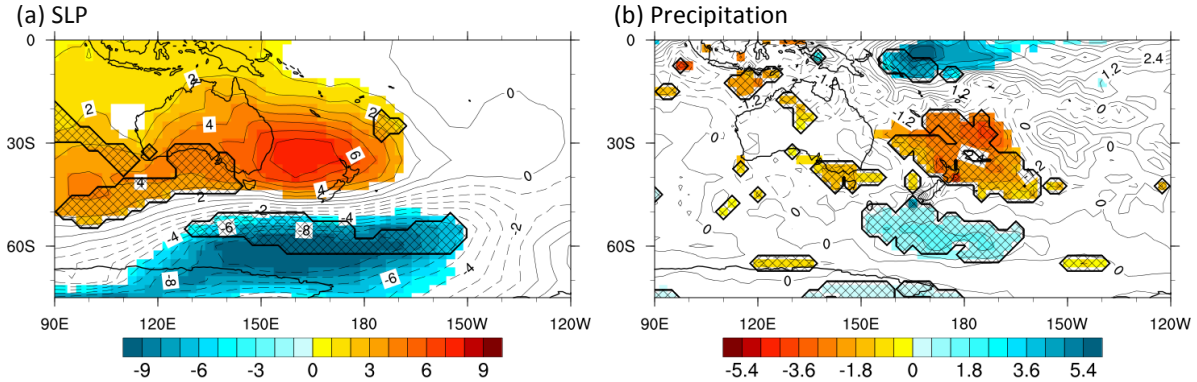
ASO | AAO | SOI

WHOLE YEAR

SOUTHERN WINTER (JJA)



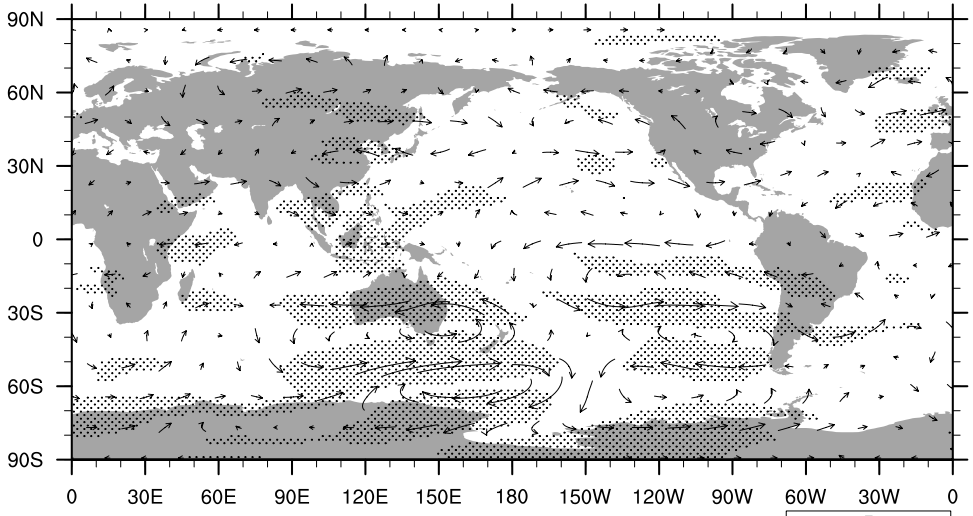
Different modes of variability over the Tasman Sea: Implications for Regional Climate (Liess et al. 2014 , J. Climate, accepted)



Annual (a) HadSLP2 [hPa] and (b) precipitation [mm d^{-1}] composites for thresholds of twice the standard deviation. Shading indicates 95% significance level. Hatching represents areas that are significant for the hybrid teleconnection, but not for SAM, ENSO, or IOD.

A positive index above twice the standard deviation is an indicator for large blocking situations and droughts over Australia.

Strong positive indices have occurred before and during the World War II Drought (1937-1945) and the Millennium Drought (1995-2012).



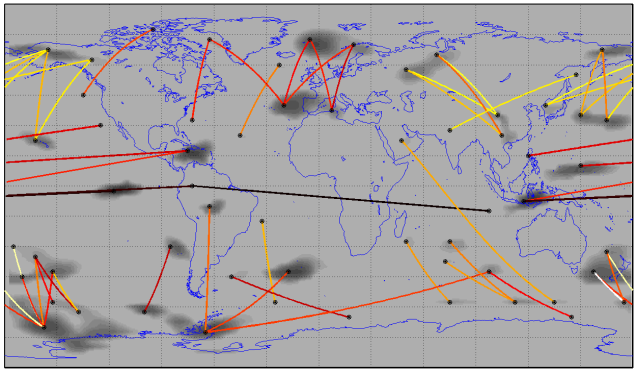
300-hPa wind vector anomalies [m s^{-1}] for positive minus negative index phases. Stippling indicates statistically significant differences in the u-wind at the 95% confidence level using a t-test.

Poleward propagating atmospheric Rossby waves can decrease the subtropical jet stream and enhance the polar jet stream, creating a counter-clockwise rotating blocking anomaly over the Tasman Sea.

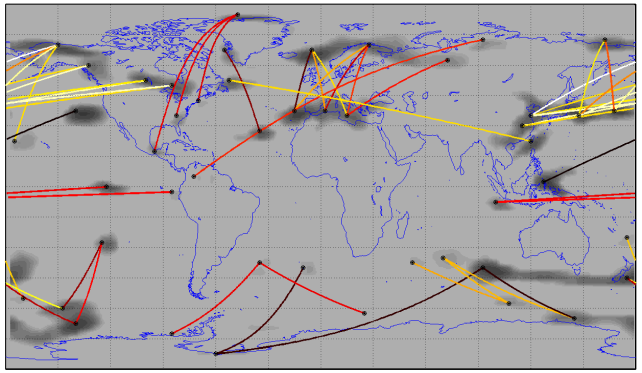
Model Analysis : Dipole Structure in CCSM and GFDL

- The dipole structure of the top 2 models from CMIP3 to CMIP5

GFDL (CMIP3)



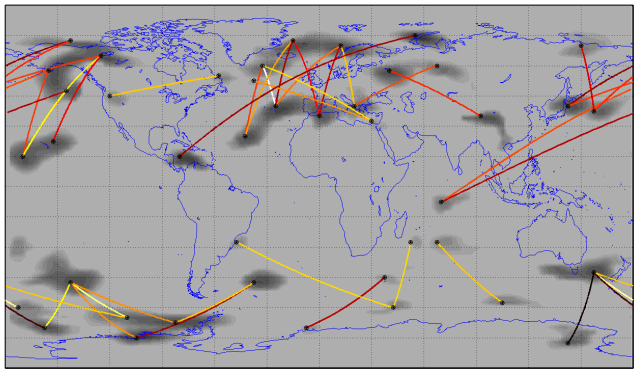
GFDL (CMIP5)



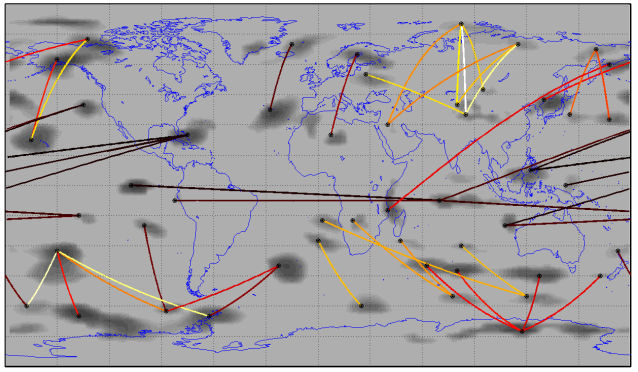
SOI present

SOI present

CCSM(CMIP3)



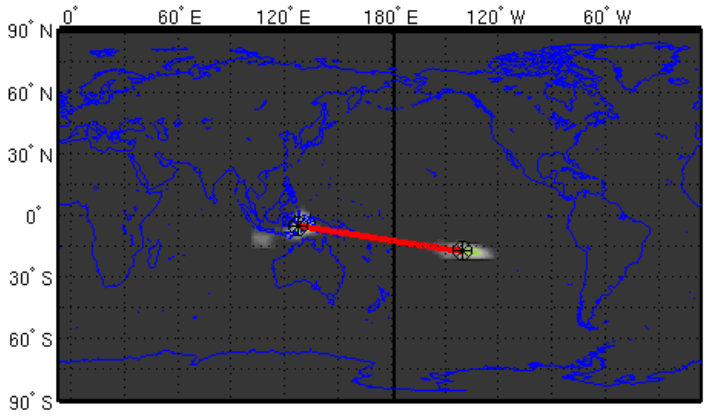
CCSM (CMIP5)

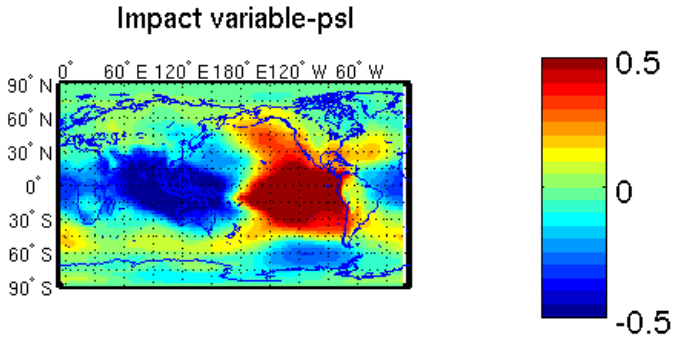


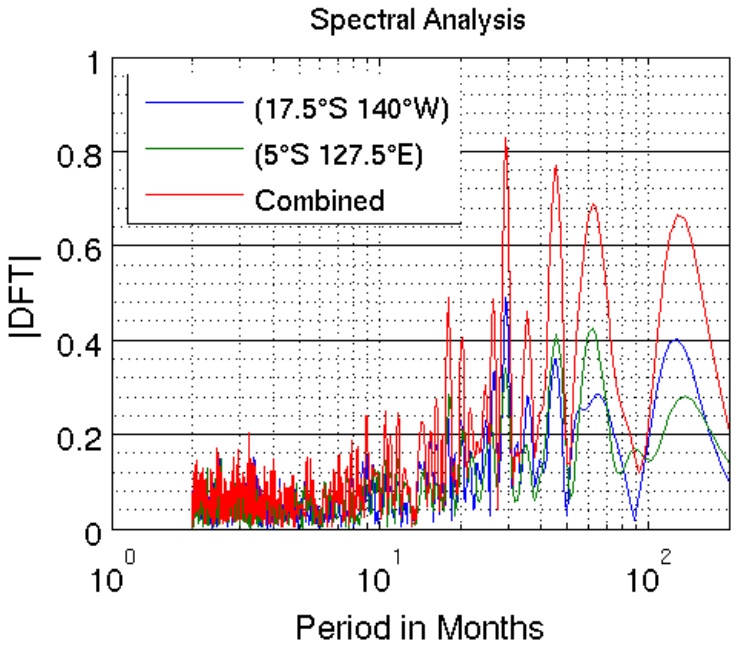
SOI absent

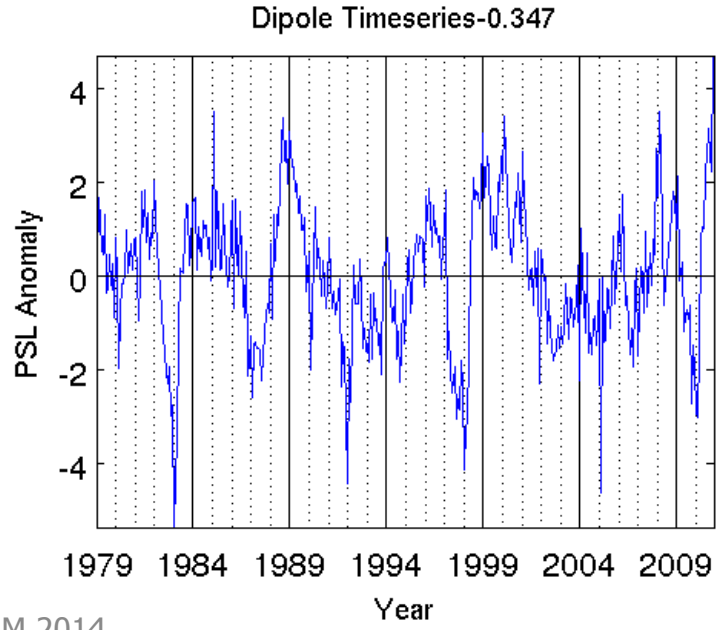
SOI present

NCEP2-73x144_SOI_1979_2011_psl__historical_detrend_mean
Parameters_0_25_0_0_0.85_0.8_2000_10_15_

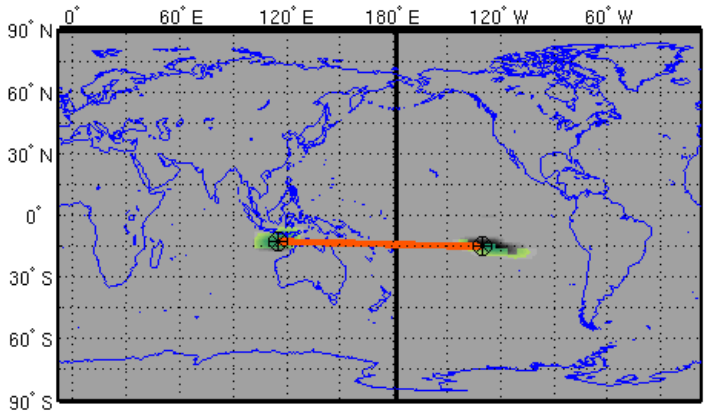


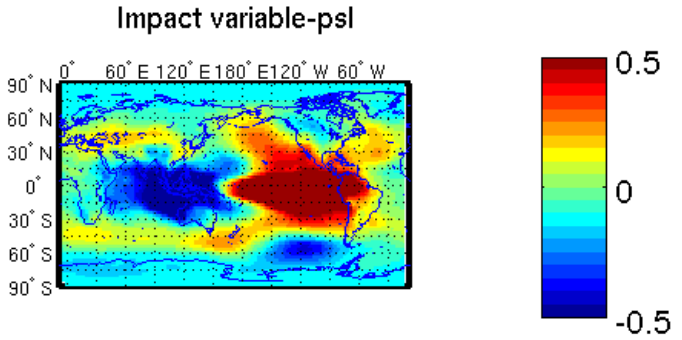


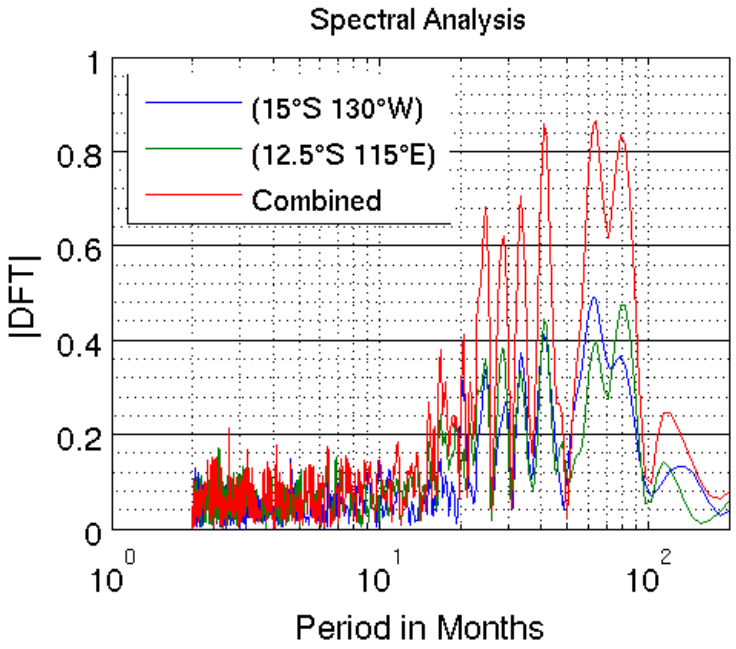


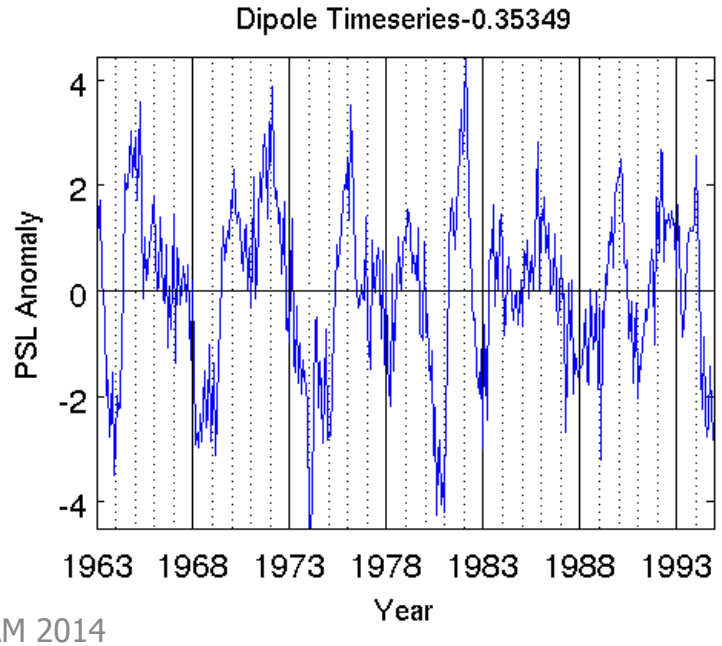


cmip3_gfdl_cm2_0-73x144_SOI_1963_1995_psl__historical_detrend_mean
Parameters_0_25_0_0_0.85_0.8_2000_10_15_

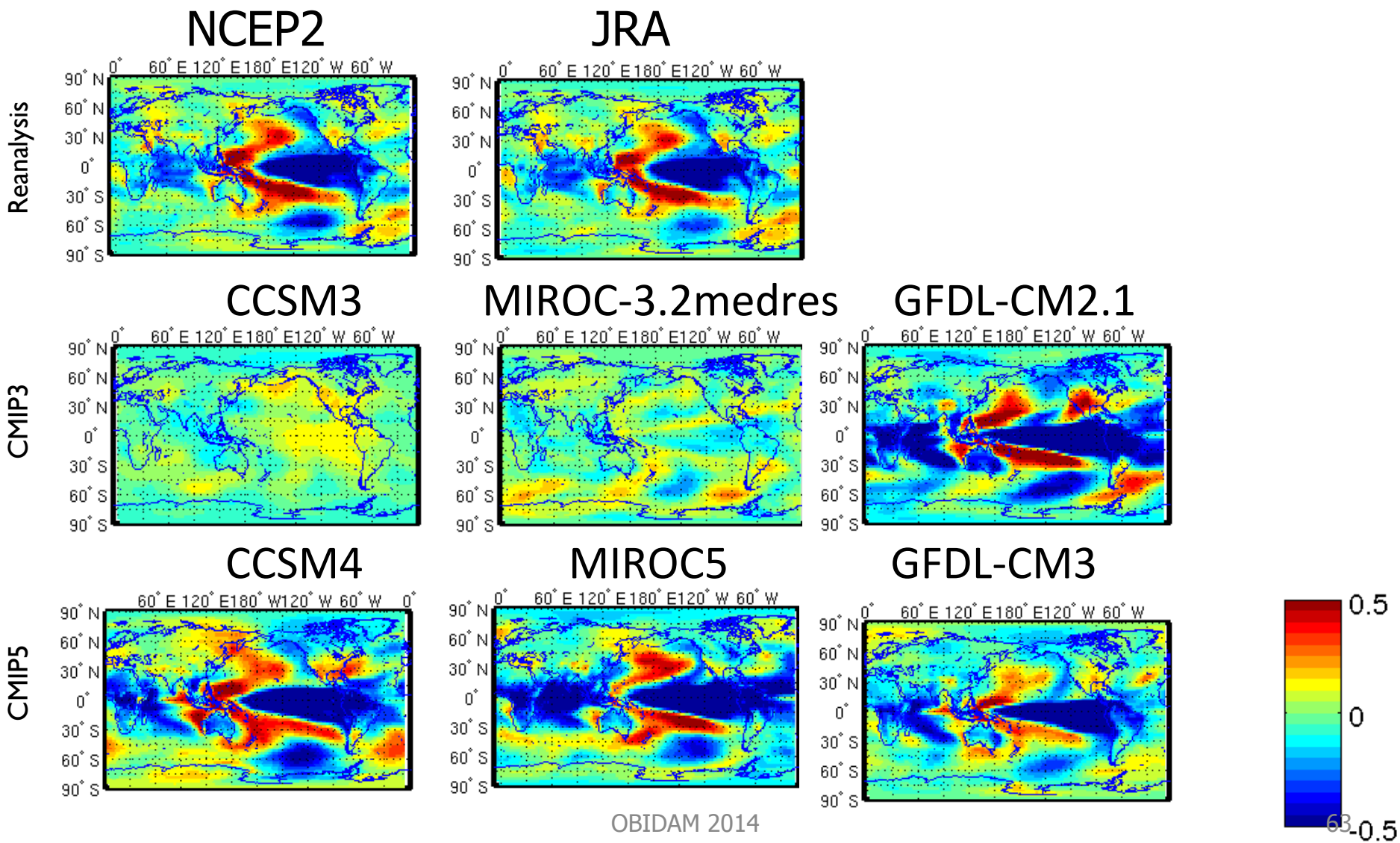








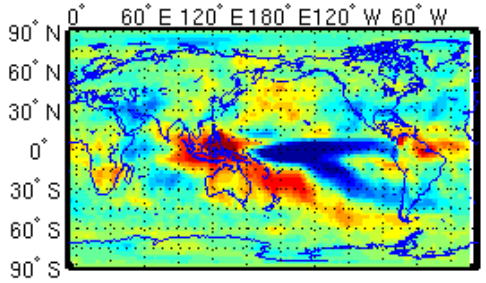
Surface Temperature Correlated with SOI



Precipitation Correlated with SOI

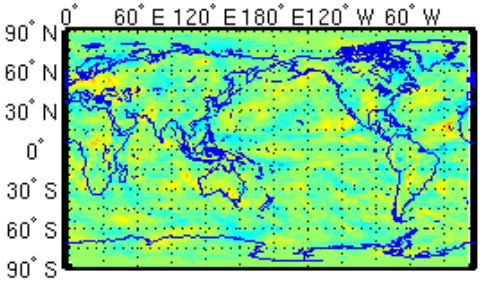
Observation

GPCP

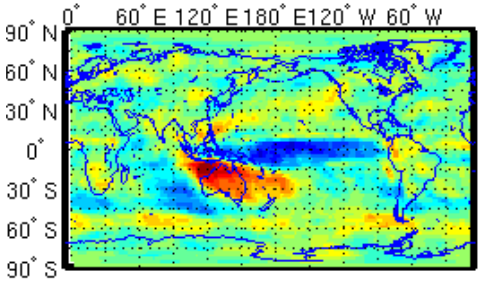


CMIP3

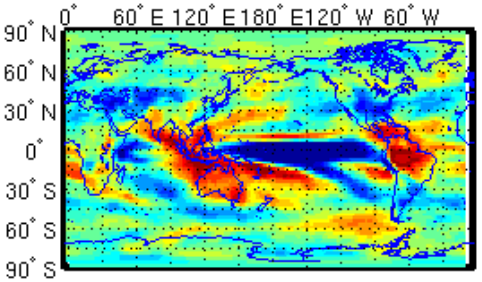
CCSM3



MIROC-3.2medres

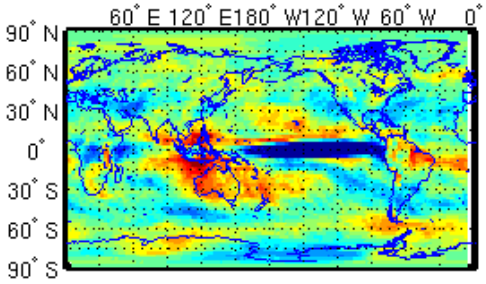


GFDL-CM2.1

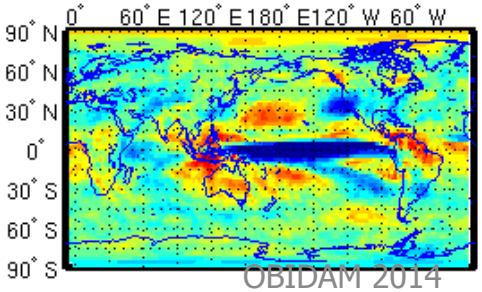


CMIP5

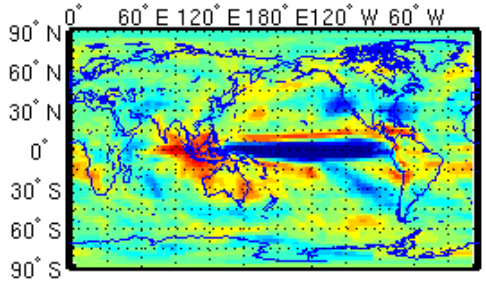
CCSM4

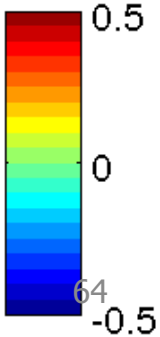


MIROC5

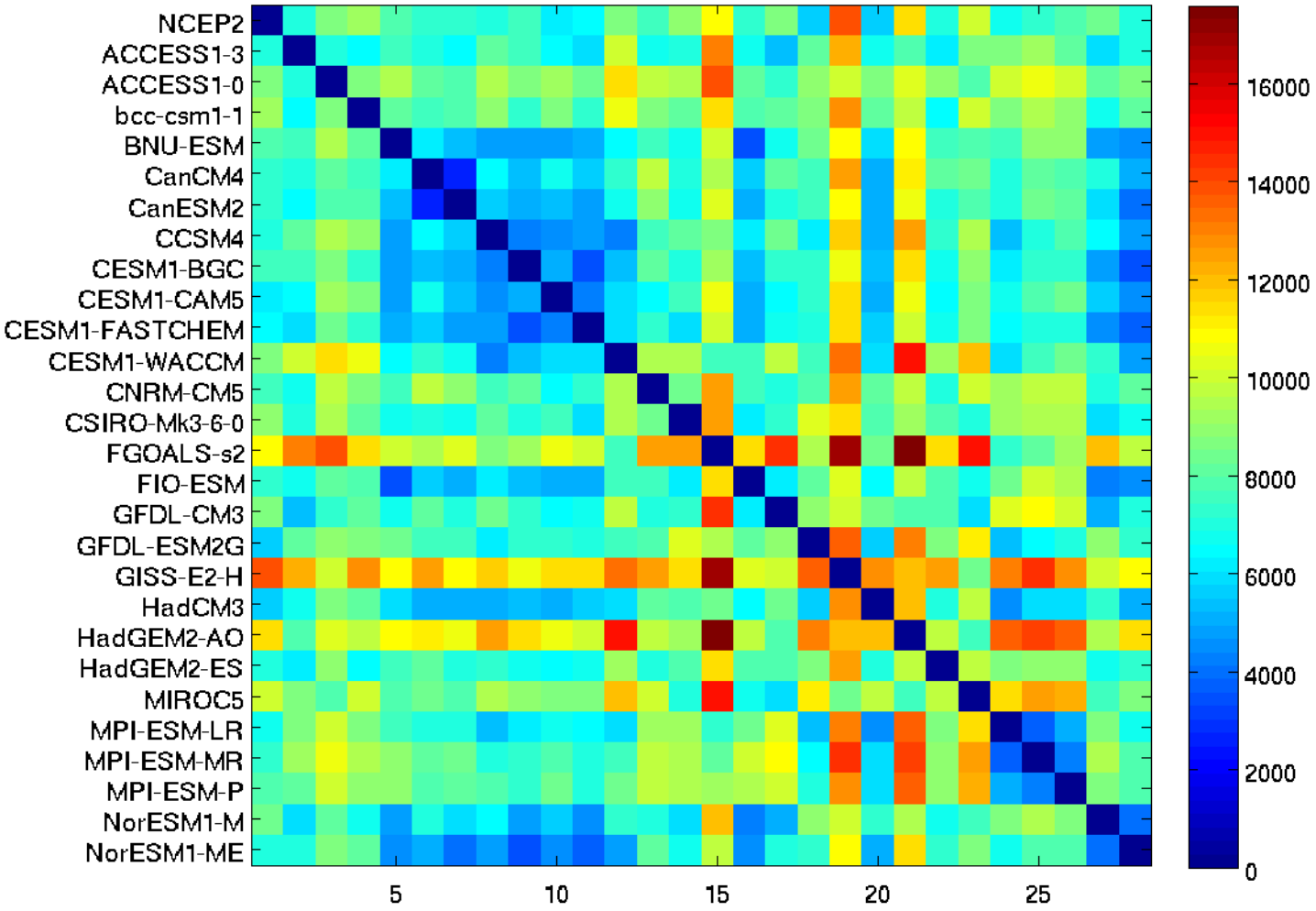


GFDL-CM3



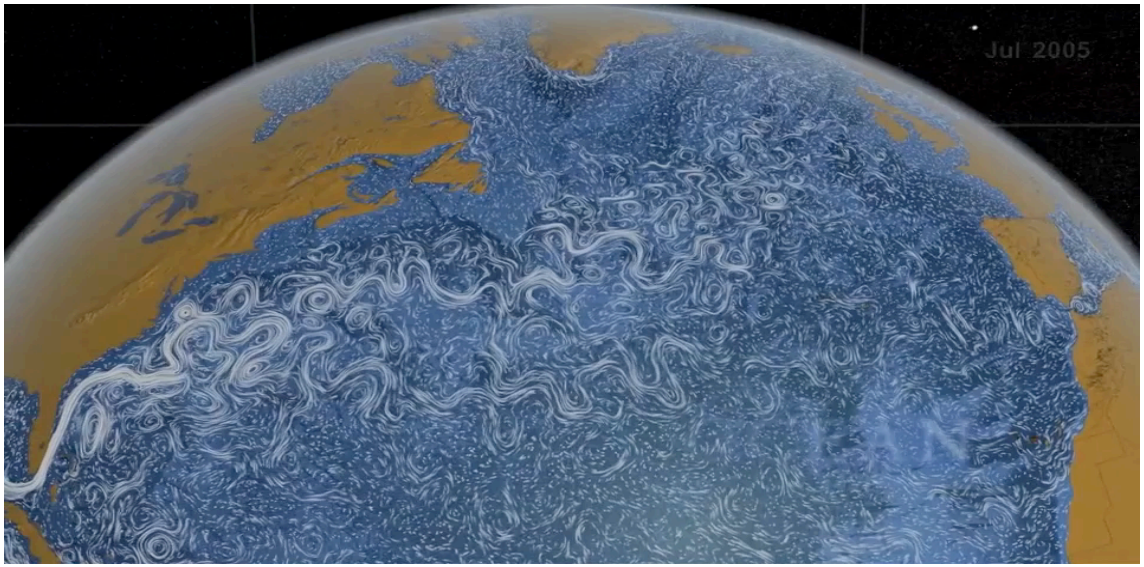


Intercomparison of CMIP5 models and NCEP2:
SOI impact on precipitation



Case 3:
Monitoring Ocean Eddies

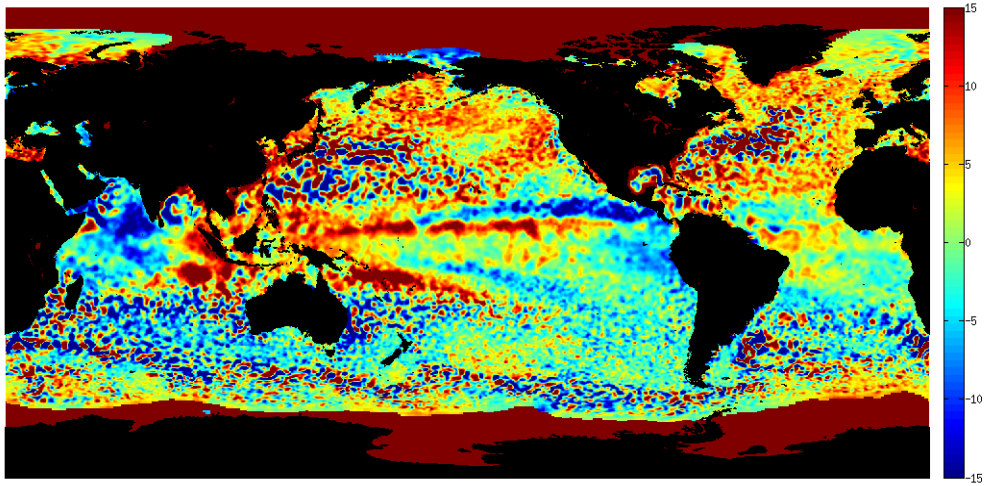
Why Are Eddies Important?



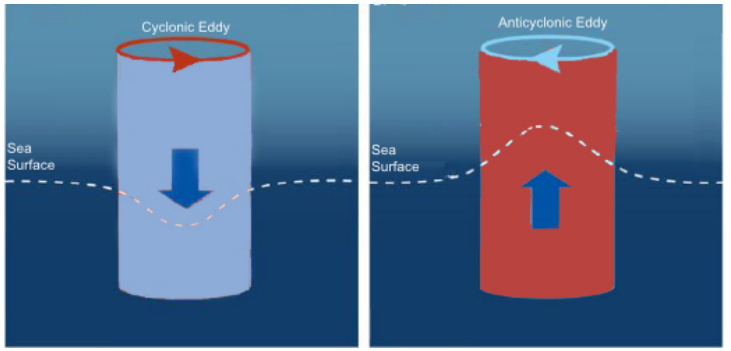
- Impact local near-surface wind, cloud properties and rainfall (Frenger et al. Nature Geo. (2013))
- Impact local near-surface chlorophyll distribution (Chelton et al. Science (2011))
- Impact larger atmospheric low-pressure systems (Yablonsky and Ginis, Mon. Wea. Rev. (2013))
- Warm "Agulhas Rings" play a potential moderating factor in global climate change (Beal et al. Nature (2011))

Global Eddy Monitoring

- Methods for autonomous eddy identification on global scale are of great interest:
  - Sea surface temperatures (e.g. Dong et al. 2011)
  - Chlorophyll (e.g. Pegau et al. 2002)
  - Sea surface height anomalies (e.g. Chelton et al. 2007, 2011)



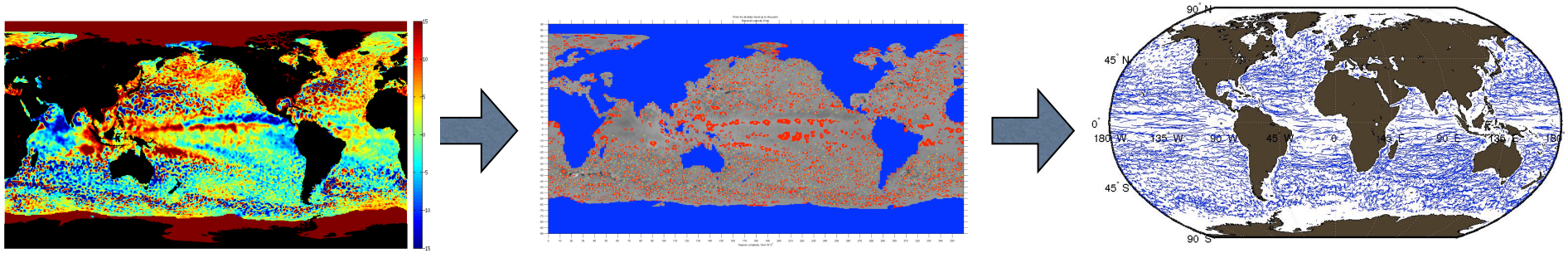
Filtered Sea Surface Height Anomaly Data



Eddies can be seen as close-contour positive (red) or negative (blue) anomalies.

Global Eddy Monitoring

Goal:
Track mesoscale eddies (50-200km) globally from 1993 - 2012 in 0.25˚ gridded daily SSH anomalies from the AVISO (Archiving, Validation, and Interpretation of Satellite Oceanographic) data



Continuous field | Discrete features | Coherent tracks

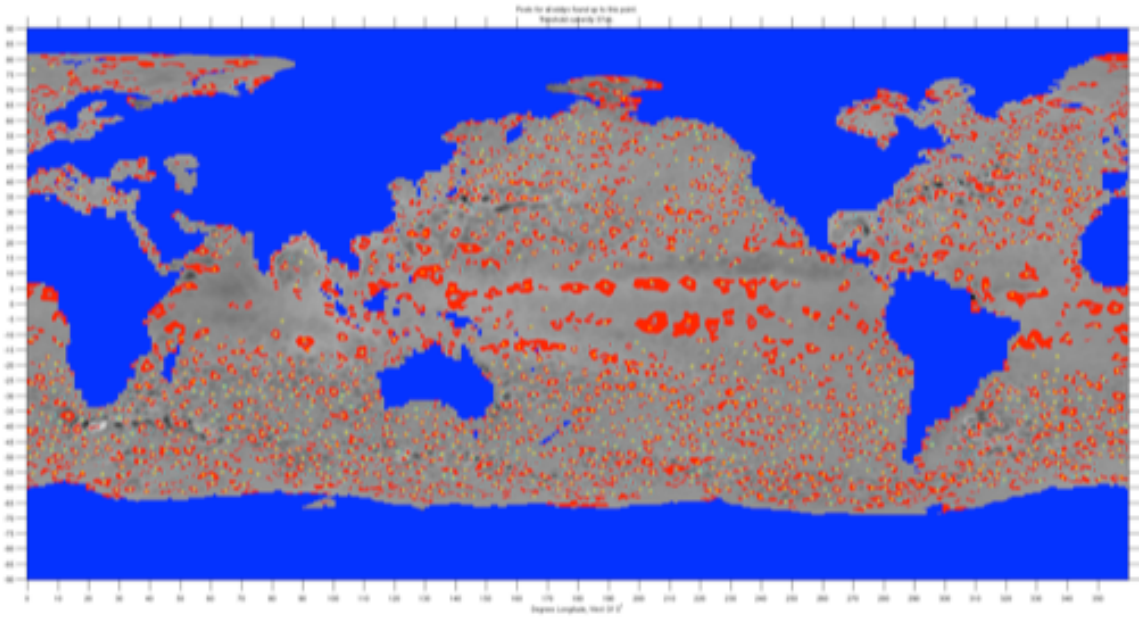
Data science challenges:

- Dynamic patterns in space and time
- Noisy and uncertain data
- Large natural variability
- No global verification dataset

Top Down Iterative Thresholding

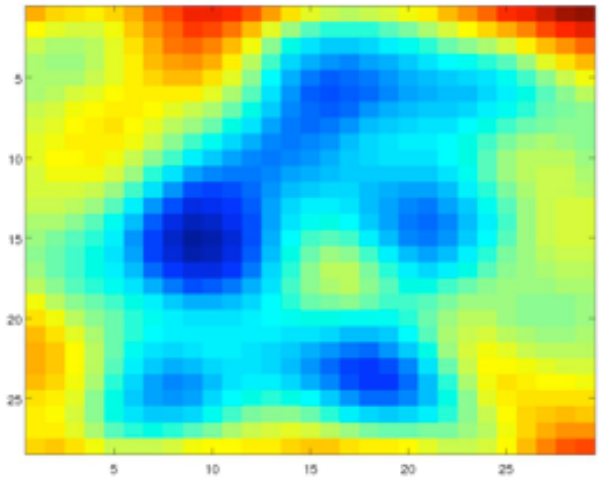
Chelton et al. 2007, 2011

- For each global snapshot of SSH:
  - Assign 1 or 0 to values if they are below or above a given threshold
  - Filter connected components by criteria such as minimum size and amplitude

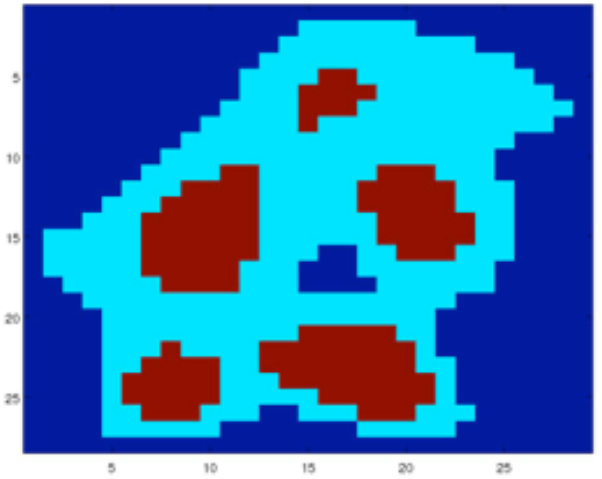


Limitations of Top Down Approach:
Merges nearby features (CSS'11)

An image of five distinct eddies that were merged together by the TD approach.



SSH anomaly.

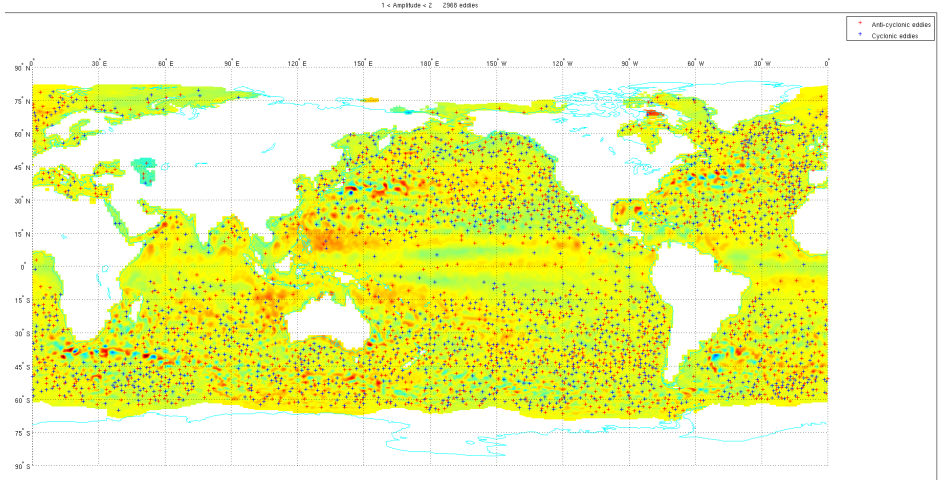


Contour of the eddy as identified by TD (light blue)

Ability to avoid artificial merges has significant benefits for describing eddy dynamics.

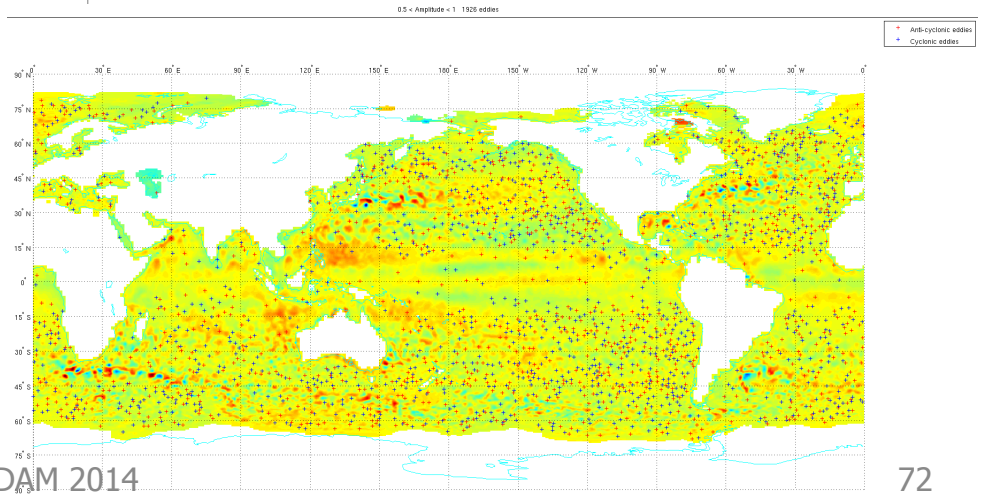
Limitations of Top Down Approach:
Filtering criteria are arbitrary

Eddy frames from July 1 – July 7, 2009



1 cm < Amplitude < 2 cm

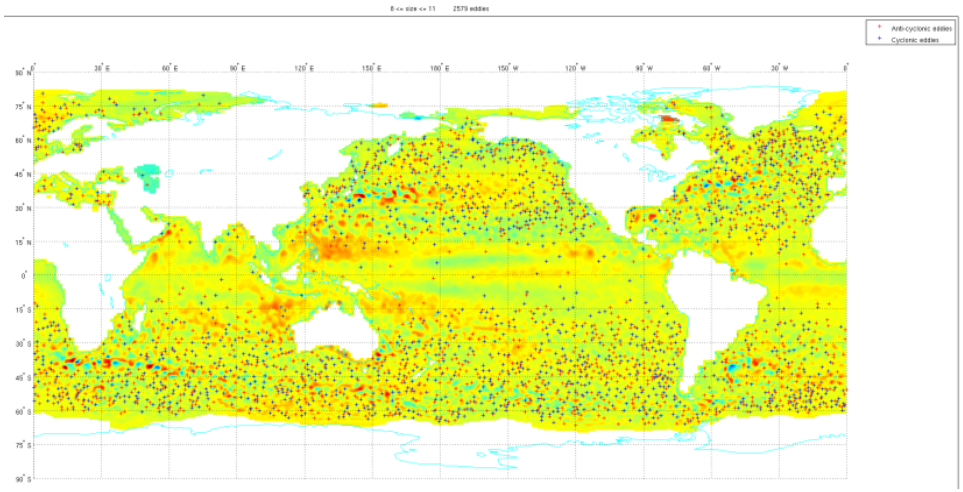
1 cm < Amplitude < 2 cm



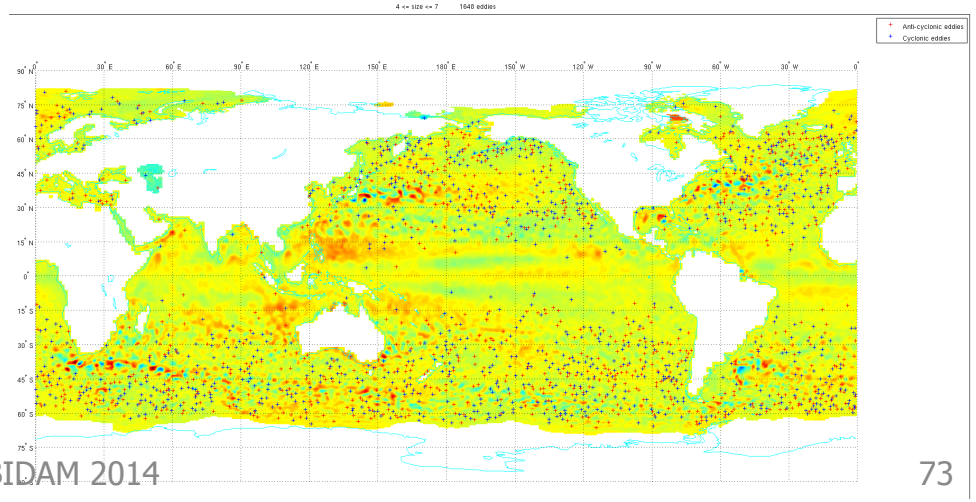
0.5 cm < Amplitude < 1 cm

Limitations of Top Down Approach:
Filtering criteria are arbitrary

Eddy frames from July 1 – July 7, 2009



8 pixels ≤ Size ≤ 11 pixels



4 pixels ≤ Size ≤ 7 pixels

Our approach

Physics-guided data mining to identify dynamic close-contour anomalies

Starts from most certain part of eddy: its extremum

Leverages the continuous field for intuitive stop criteria

Results in improved quality of features by seperating merged eddies

Faghmous et al. 2012 a,b,
Faghmous et al. 2013 a,b

Output

Physics-guided data mining to identify dynamic close-contour anomalies

~39 million features

~1 million eddy tracks

First publicly available comprehensive eddies dataset

Used by other centers

http://ucc.umn.edu/eddies

Eddy-TC Interactions

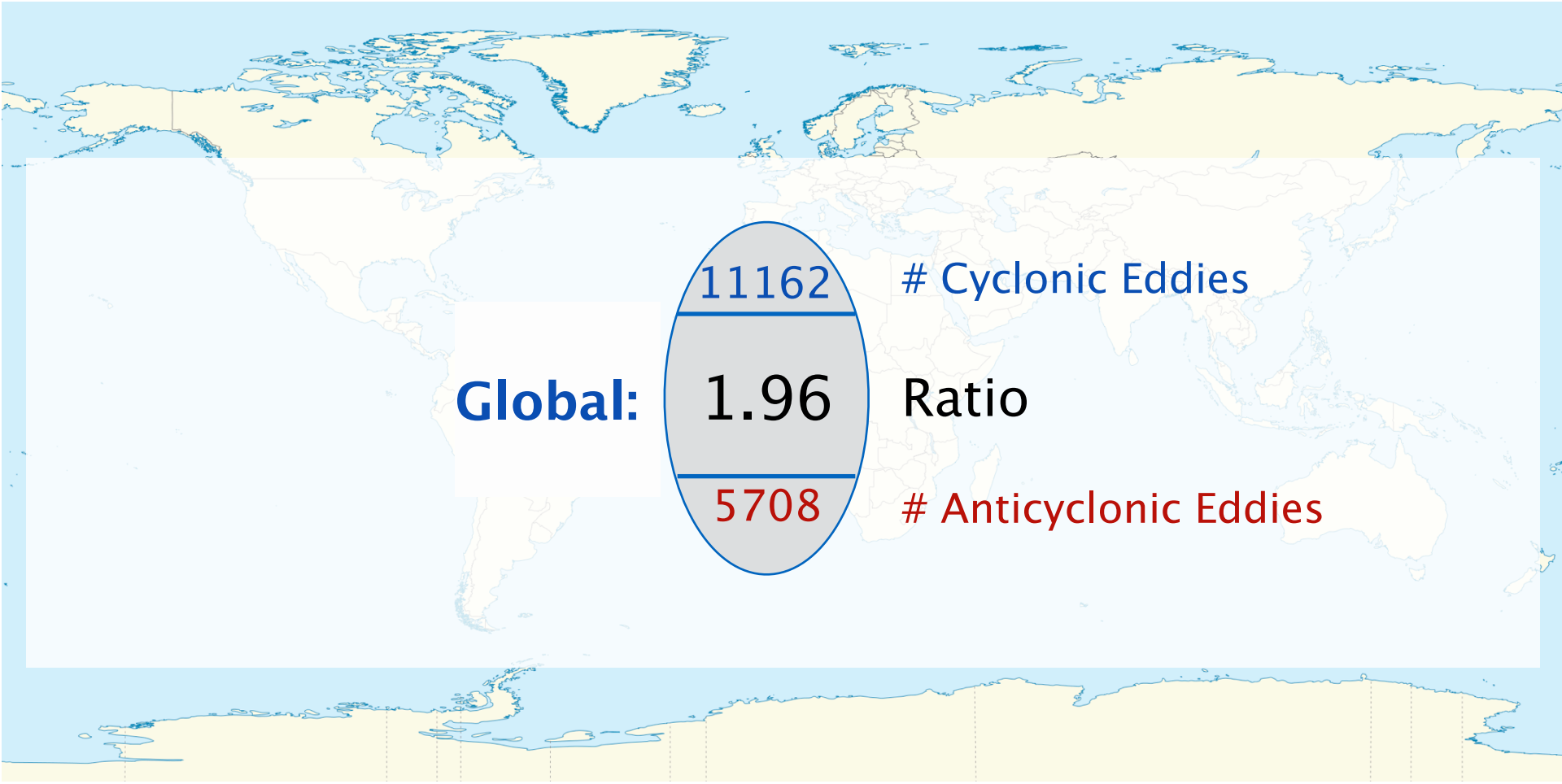
Experiment Design

1. 6-hourly Best Track TC data (IBTrACS)φ + daily ocean eddies from sea surface height anomalies data (1993-2012)
    - 2069 TC Tracks
    - 71652 TC Steps
        - 50k after removing TC steps with missing wind-speed data.
    - 19.5 Million Cyclonic Eddy Frames
    - 19.1 Million Anticyclonic Eddy Frames

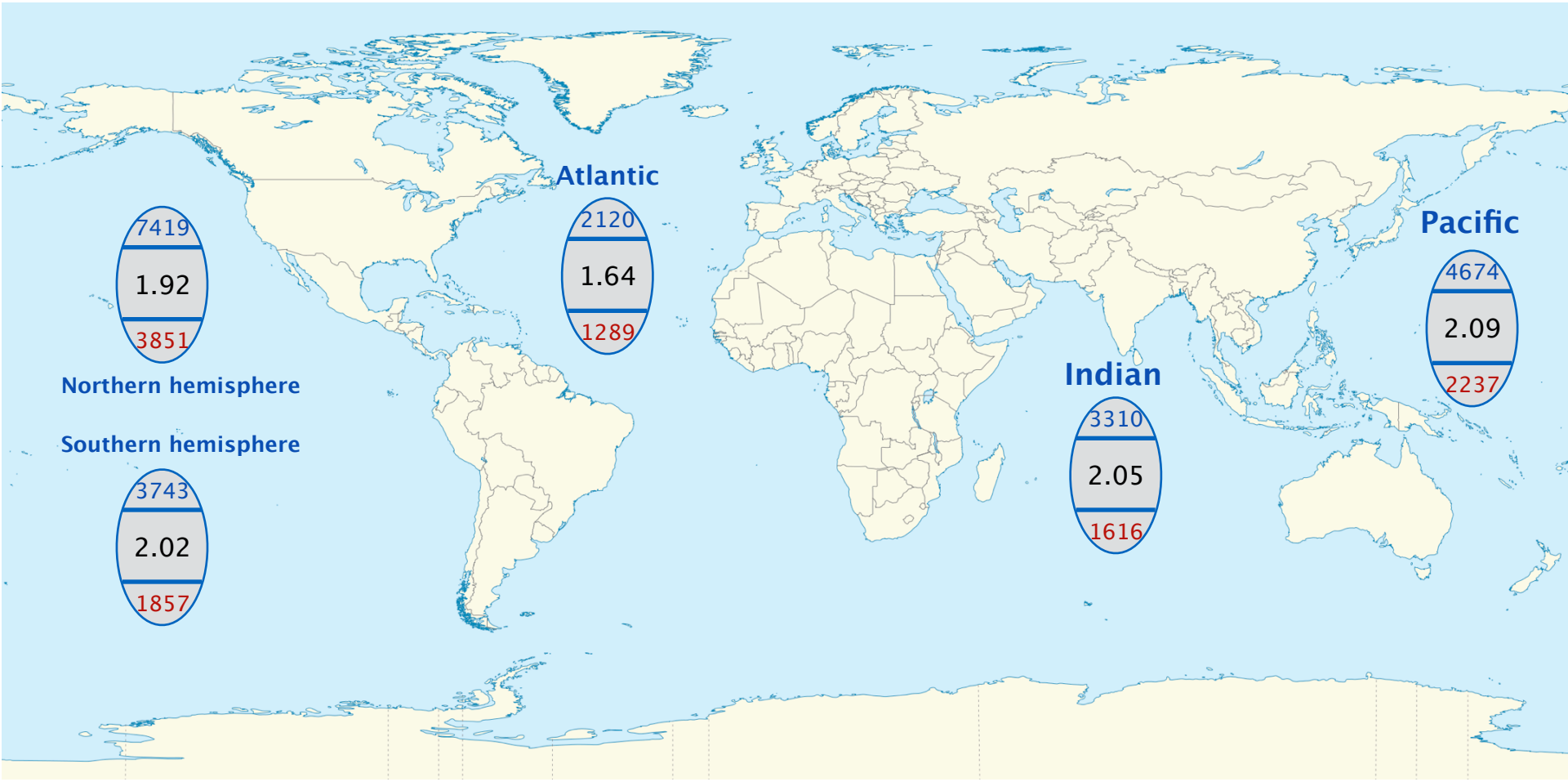
2. Associate a 6-hourly TC observation to an eddy if it passes within 20km of its contour
    - 11162 Cyclonic Eddy Interactions
    - 5708 Anticyclonic Eddy Interactions

φ http://www.ncdc.noaa.gov/ibtracs/

Interaction Statistics

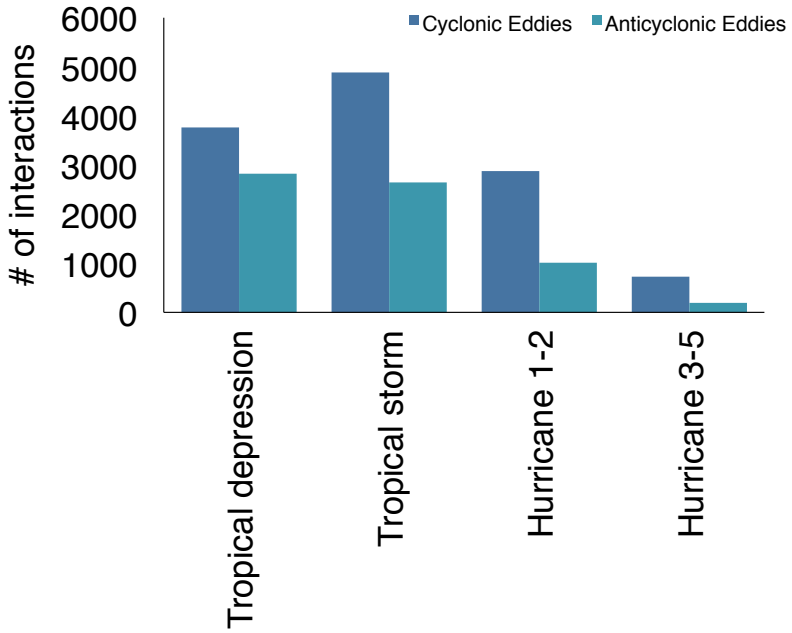


Interaction Statistics

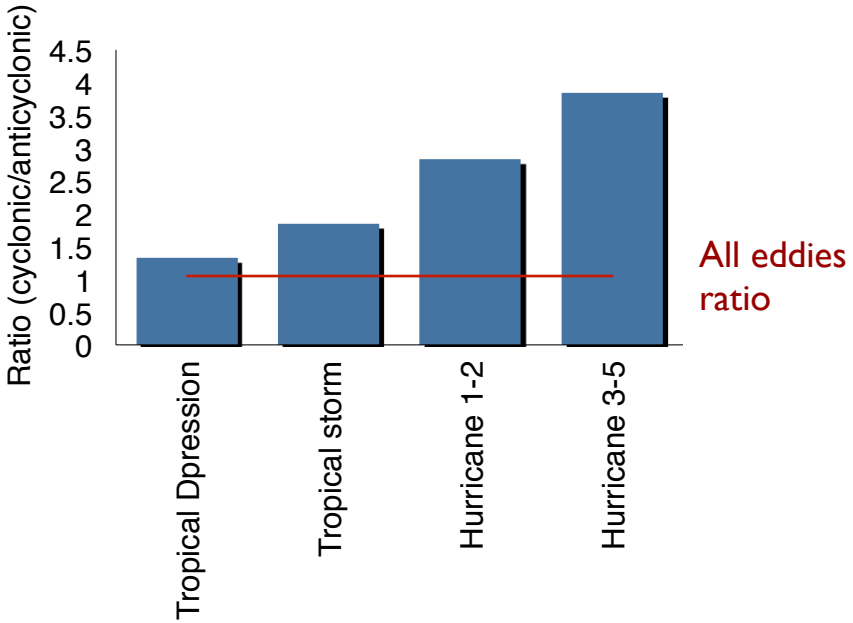


Eddy-TC Interactions by intensity

of interactions



Ratio (cyclonic/anticyclonic)



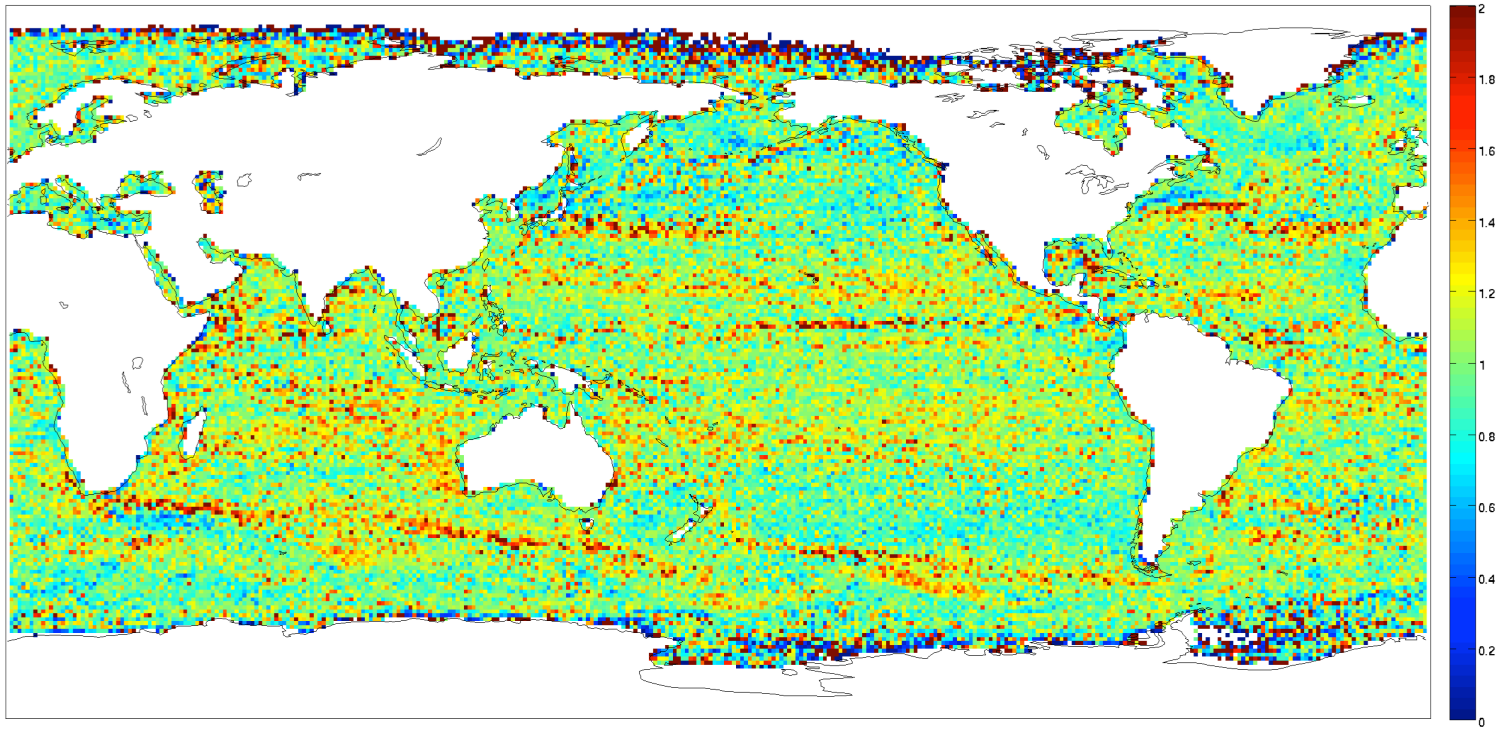
Some Hypotheses

1. There are more cyclonic than anticyclonic eddies
2. Cyclonic eddies are larger
3. TCs are spawning cyclonic eddies
4. Atmospheric conditions favor both cyclonic eddies and TCs

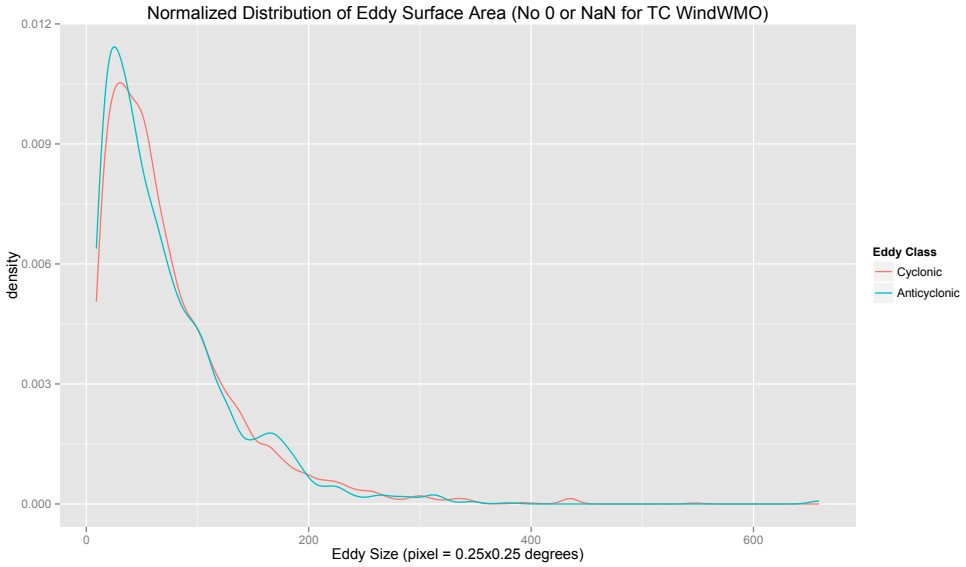
Hypothesis #1
More cyclonic than anticyclonic eddies

- Global cyclonic to anticyclonic ratio: 1.05
- Along TC tracks (3 degree radius) all times: 1.02

Global Eddy Ratio Map (1°x1°):Years 1993-2012

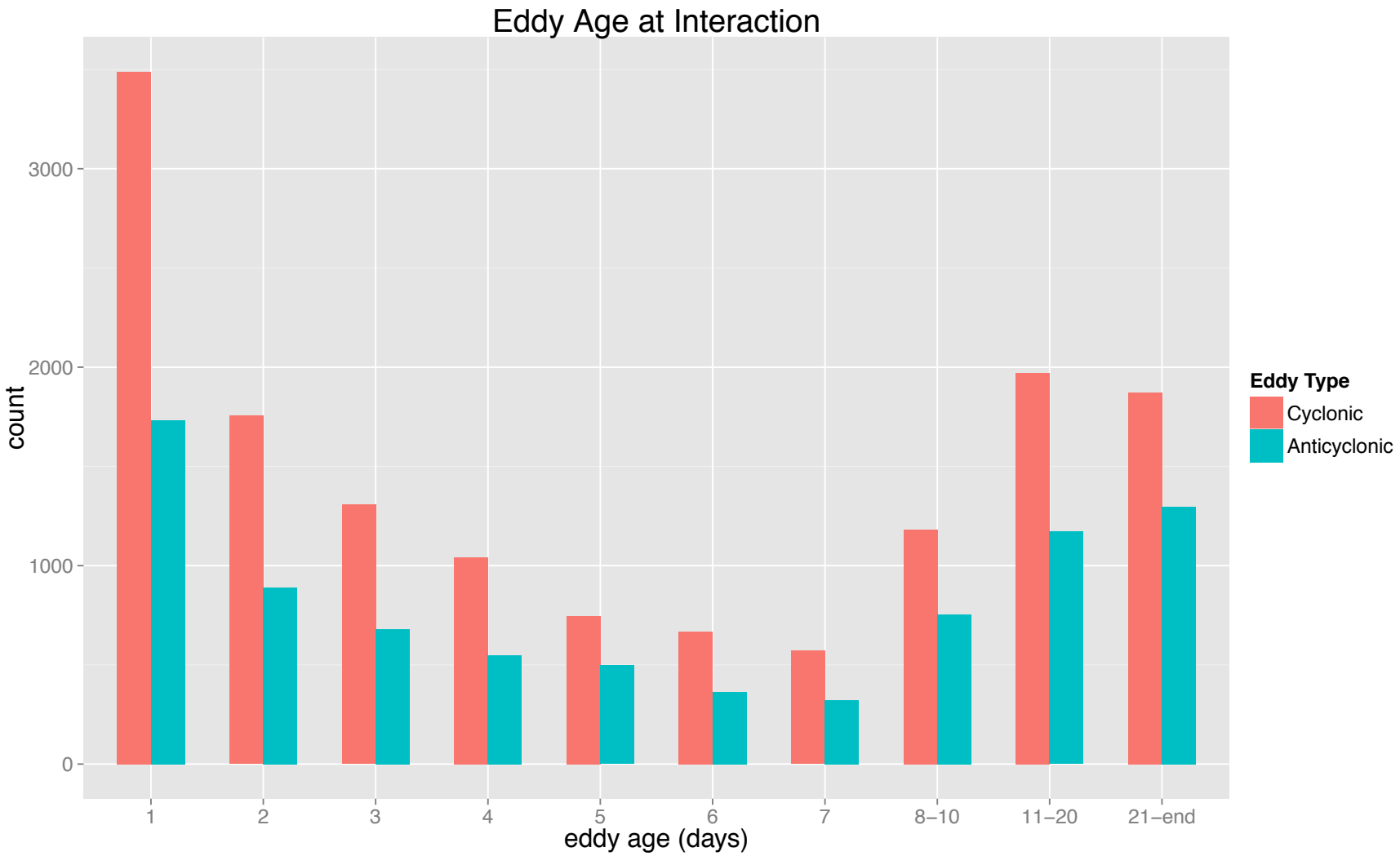


Hypothesis #2
Larger cyclonic than anticyclonic eddies

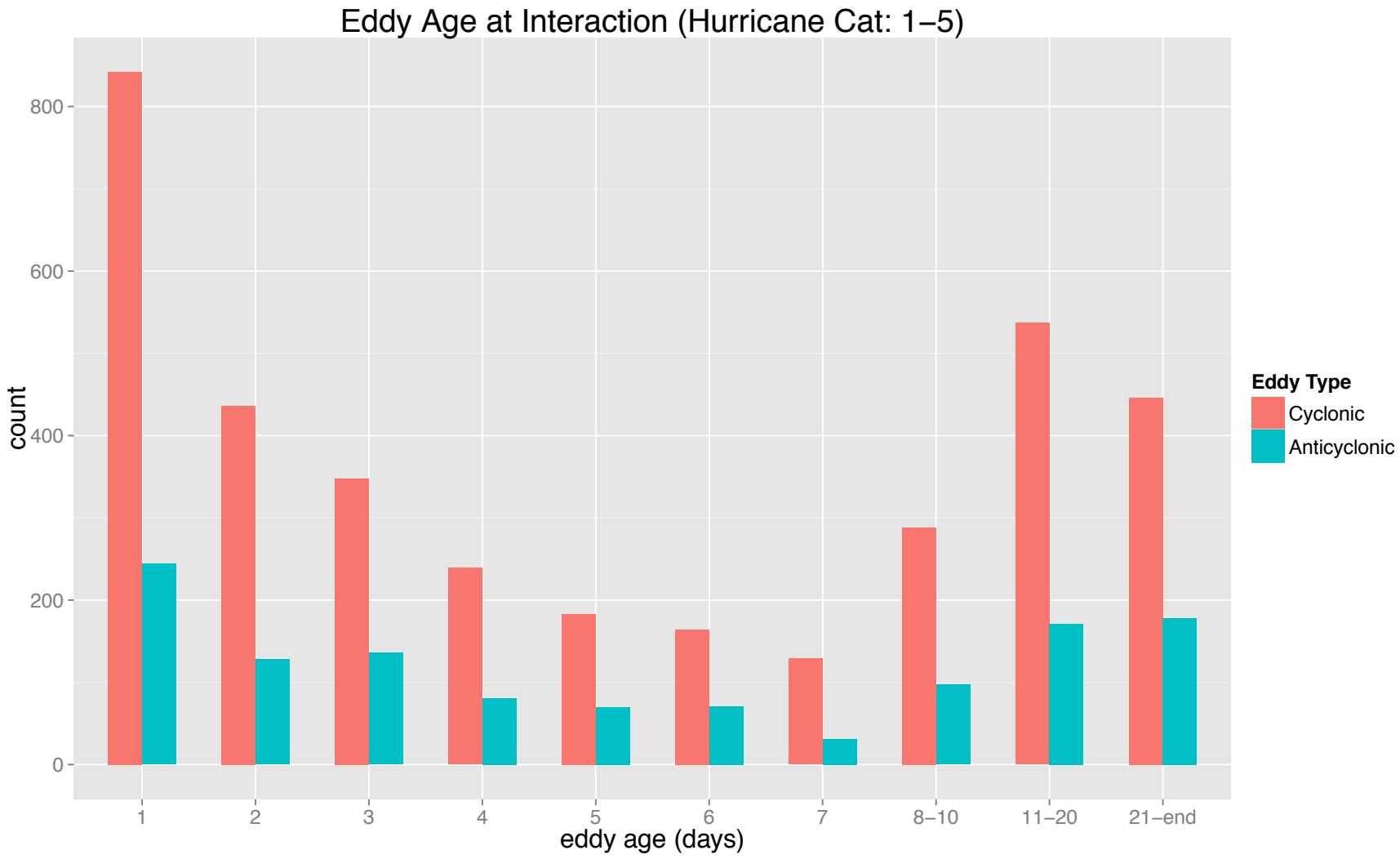


Global size ratio
1.05

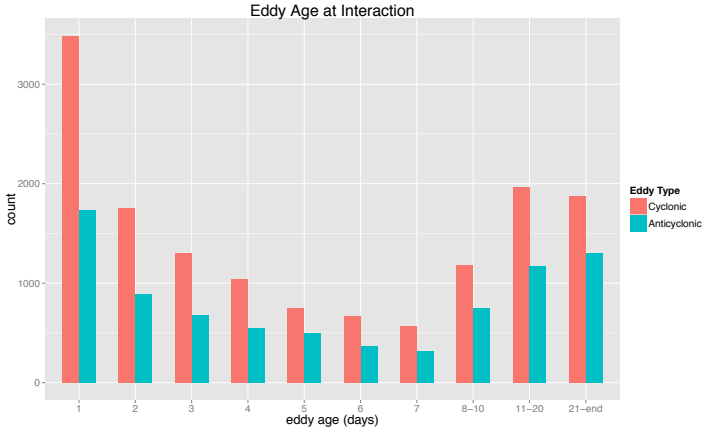
Hypothesis #3
TCs spawn eddies



Hypothesis #3
TCs spawn eddies



Hypothesis #3
TCs spawn eddies



Recent related work

Sun, L., et. al (2014), Effects of super typhoons on cyclonic ocean eddies in the western North Pacific: A satellite data-based evaluation between 2000 and 2008, J. Geophys. Res. Oceans, 119, doi: 10.1002/2013JC009575.

Some Hypotheses

1. ~~There are more cyclonic than anticyclonic eddies~~
2. ~~Cyclonic eddies are larger~~

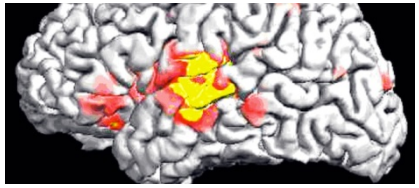
✔ 3. TCs are spawning cyclonic eddies

? 4. Atmospheric conditions favor both cyclonic eddies and TCs

Concluding Remarks

- Earth science problems provide transformative opportunities for data-driven research
  - Complex dependence and noise structures
  - Nonlinear dynamical spatiotemporal systems
  - Data size from few petabytes 350 petabytes by 2030
  - Motivates the development of "physics-guided data mining"
    - Patterns should be spatially and temporally consistent

- Novel spatiotemporal methods can generalize to multiple domains
  - Brain science
  - Ecology and biodiversity
  - Social networks
  - Geospatial Intelligence



– Help establish the field of "climate informatics" over the next 5-10 years