NSF Expeditions in Computing

# Understanding Climate Change: A Data Driven Approach

**Vipin Kumar**

University of Minnesota

kumar@cs.umn.edu

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

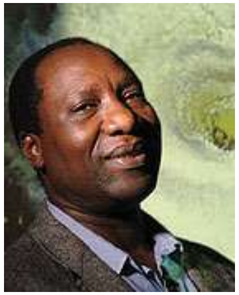
# Expeditions Team

Vipin Kumar, UM

Auroop Ganguly, NEU

Nagiza Samatova, NCSU

Arindam Banerjee, UM


Fred Semazzi, NCSU

Joe Knight, UM

Shashi Shekhar, UM

Peter Snyder, UM

Jon Foley, UM

Alok Choudhary, NW

Ankit Agrawal, NW

Abdollah Homiafar
NCA&T

Michael Steinbach
UM

Singdhansu Chatterjee
UM

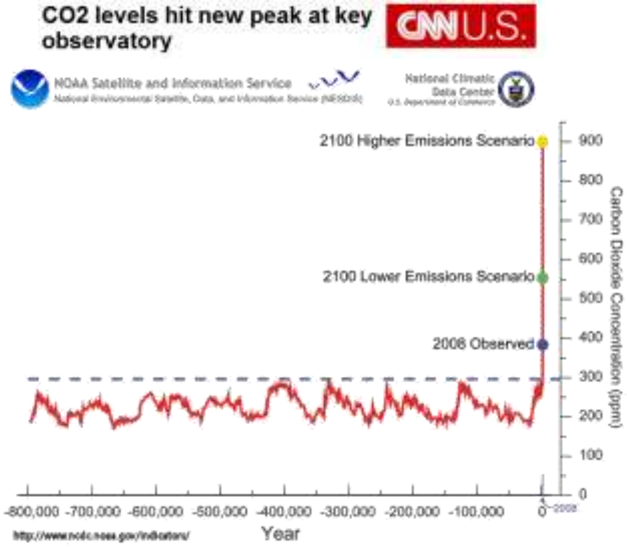
Karsten Steinhaeuser
UM

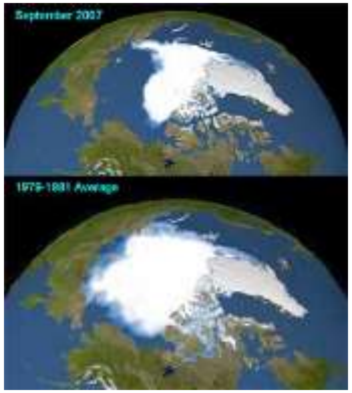
Stefan Liess
UM

Shyam Boriah
UM

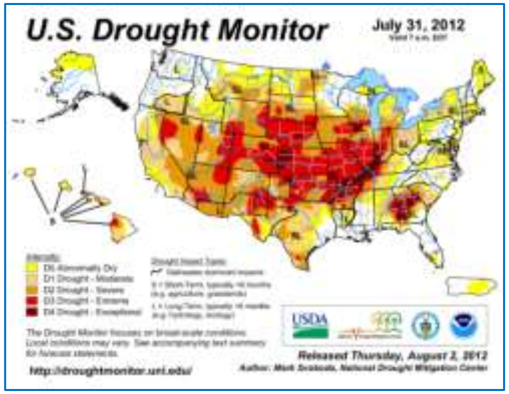
# Understanding Climate Change - Motivation

- **The planet is warming**
  - Multiple lines of evidence
  - Credible link to human GHG (green house gas) emissions
- **Consequences can be dire**
  - Extreme weather events
  - Regional climate and ecosystem shifts
- **There is an urgency to act**
  - Adaptation: “Manage the unavoidable”
  - Mitigation: “Avoid the unmanageable”
- **The societal cost of both action and inaction is large**





The Vanishing of the Arctic Ice cap
ecology.com, 2008


U.S. Drought Monitor
July 31, 2012
Intensity:
D0 Abnormally Dry
D1 Drought - Moderate
D2 Drought - Severe
D3 Drought - Extreme
D4 Drought - Exceptional
Drought Impact Types:
S = Short-Term, typically <6 months (e.g. agriculture, grasslands)
L = Long-Term, typically >6 months (e.g. hydrology, ecology)
The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.
http://droughtmonitor.unl.edu/
USDA
Released Thursday, August 2, 2012
Author: Mark Svoboda, National Drought Mitigation Center

Russia Burns, Moscow Chokes
NATIONAL GEOGRAPHIC, 2010

**Key outstanding science challenge:**

***Actionable predictive insights to credibly inform policy***

# Understanding Climate Change – Physics-Based Approach

**General Circulation Models:** Mathematical models with physical equations based on fluid dynamics

*Parameterization and non-linearity of differential equations are sources for uncertainty!*

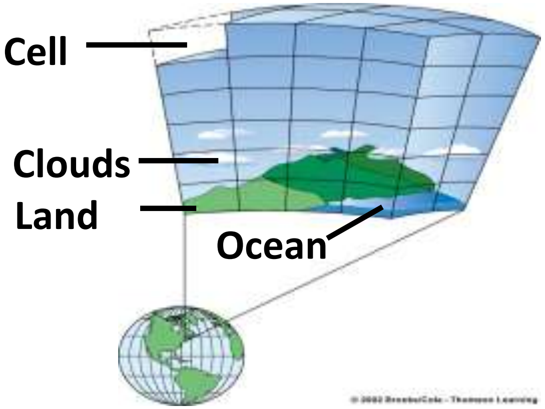
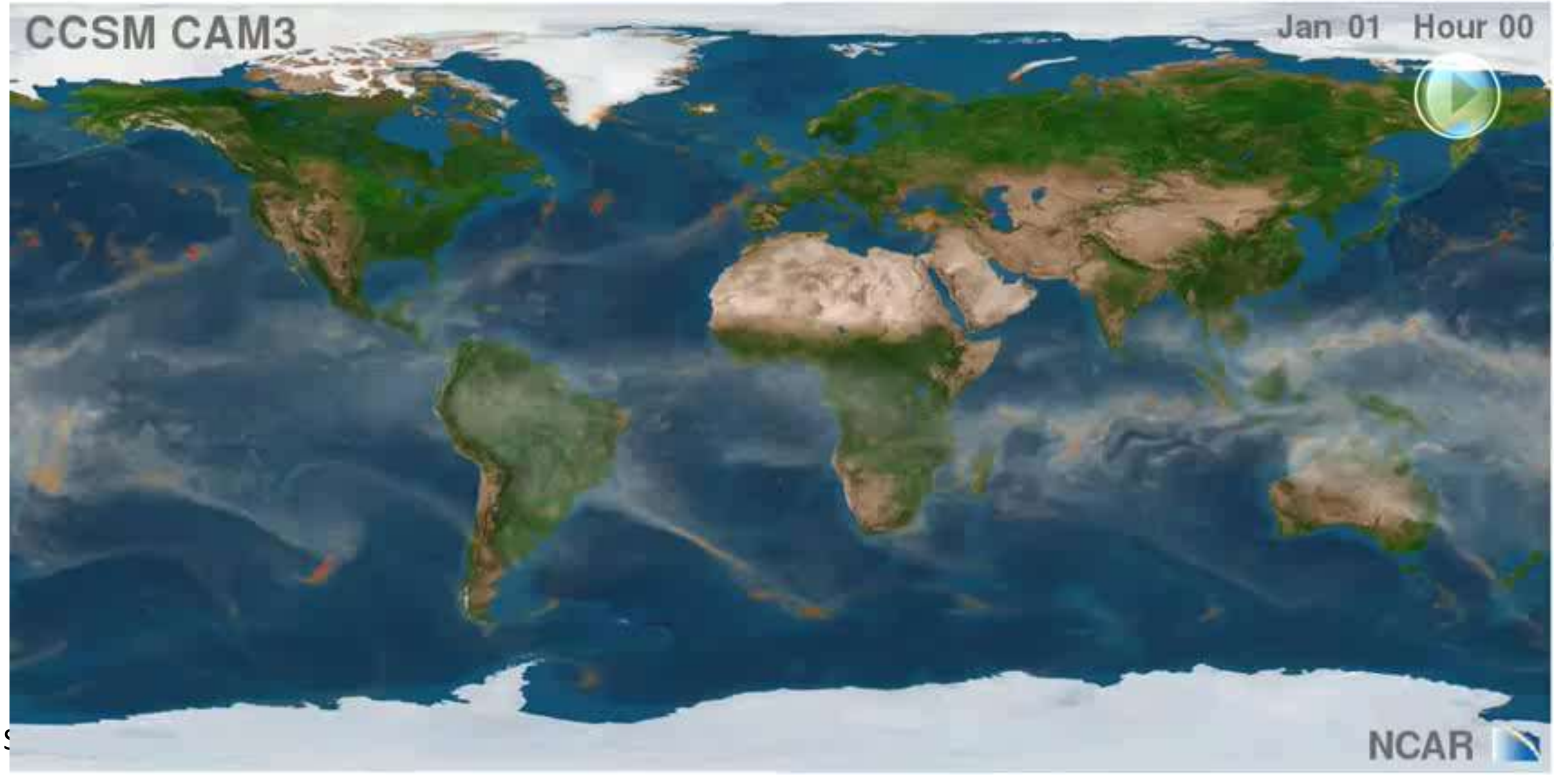
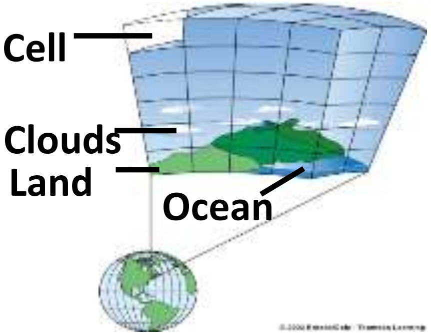


*Figure Courtesy: NCAR*



# Understanding Climate Change - Physics Based Approach

**General Circulation Models:** Mathematical models with physical equations based on fluid dynamics



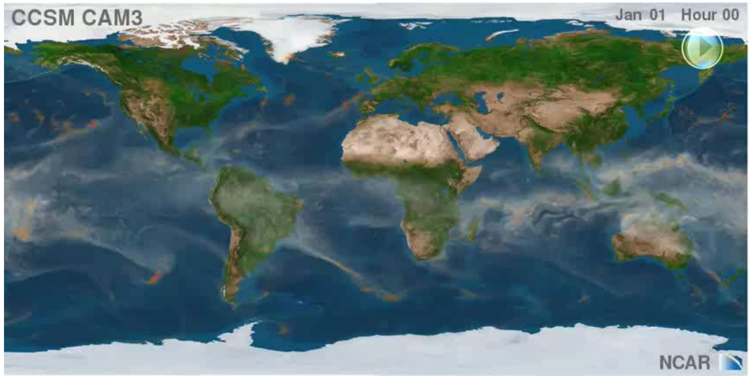
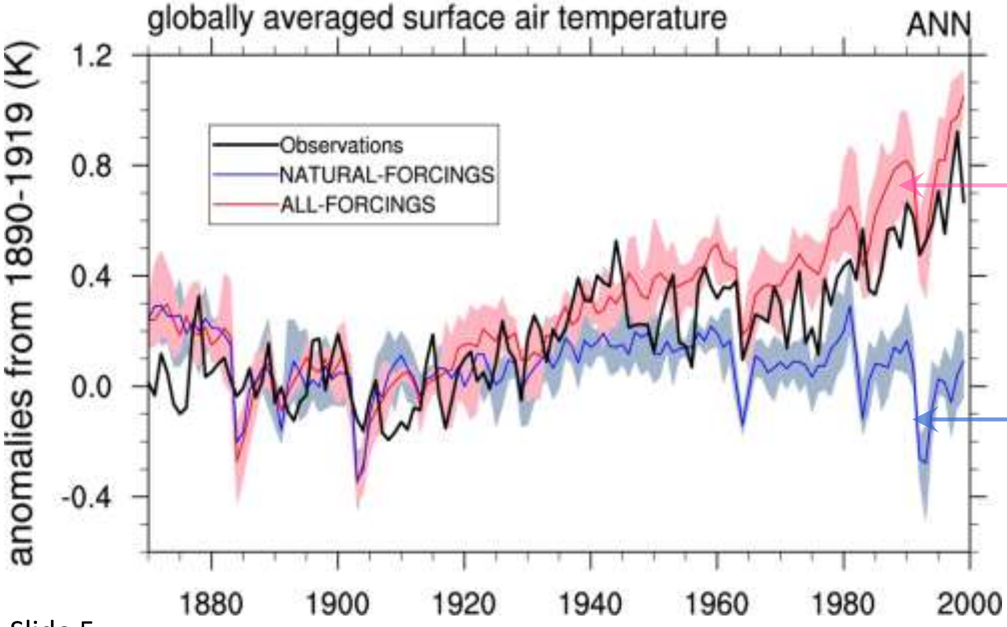


*Figure Courtesy: NCAR*

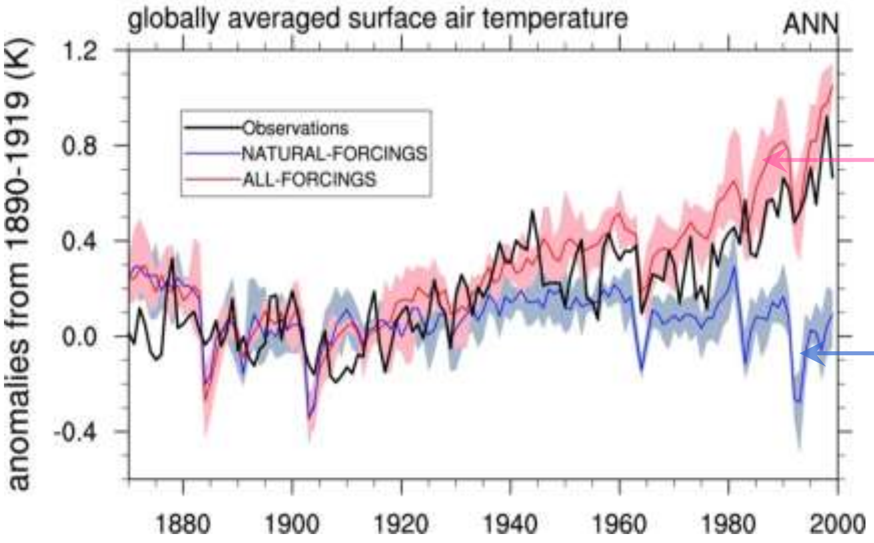


Ensemble average with observed greenhouse gas concentrations

Ensemble average with pre-industrial greenhouse gas concentrations

*Figure Courtesy: ORNL*

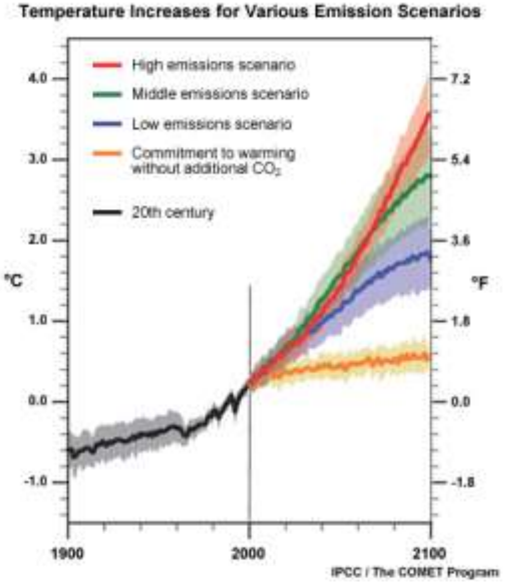
# Understanding Climate Change - Physics Based Approach



Ensemble average with observed greenhouse gas concentrations

Ensemble average with pre-industrial greenhouse gas concentrations

*Figure Courtesy: ORNL*



Projection of temperature increase under different **Special Report on Emissions Scenarios** (SRES) by 24 different GCM configurations from 16 research centers used in the **Intergovernmental Panel on Climate Change** (IPCC) 4th Assessment Report.

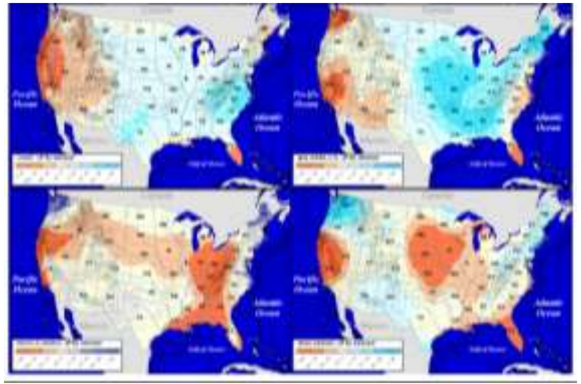
# Physics based models are essential but insufficient

- Relatively reliable predictions at global scale for ancillary variables such as temperature
- Least reliable predictions for variables that are crucial for impact assessment such as regional precipitation

***"The sad truth of climate science is that the most crucial information is the least reliable"***

(Nature, 2010)

**Disagreement between IPCC models**



Regional hydrology exhibits large variations among major IPCC model projections

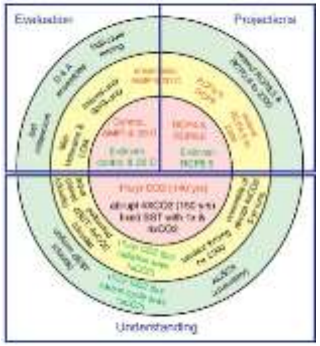
Physics based models

| Low uncertainty | High uncertainty | Out of scope |
|---|---|---|
| Temperature | Hurricanes | Fires |
| Pressure | Extremes | Malaria outbreaks |
| Large-scale wind | Precipitation | Landslides |

# Data-Driven Knowledge Discovery in Climate Science

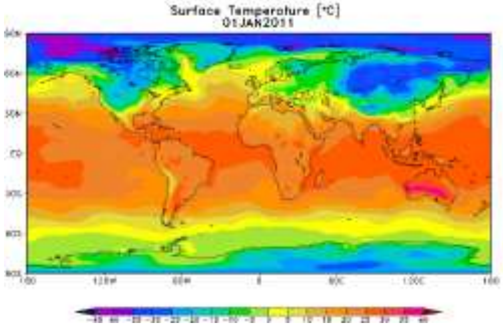
**Transformation from Data-Poor to Data-Rich**

- Sensor Observations
- Reanalysis Data
- Model Simulations



A new and transformative data-driven approach that:

- Makes use of wealth of observational and simulation data
- Advances understanding of climate processes
- Informs climate change impacts and adaptation



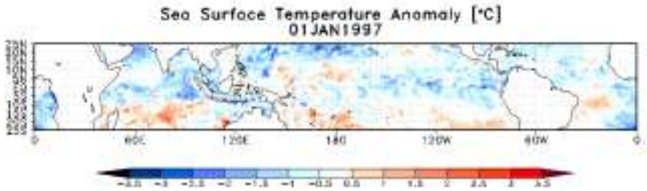
“Climate change research is now ‘big science,’ comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics.”

**(Nature Climate Change, Oct 2012)**

# Need for data driven analysis

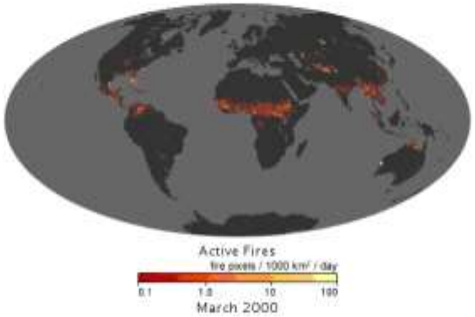
| Low uncertainty | High uncertainty | Out of scope |
|---|---|---|
| Temperature | Global hurricanes | Global fires |
| Pressure | Extremes | Malaria outbreaks |
| Large-scale wind | Precipitation | Landslides |

Global sea surface temperatures



Atlantic hurricanes

Global fires



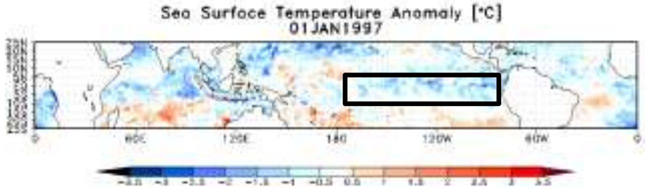
# Need for data driven analysis

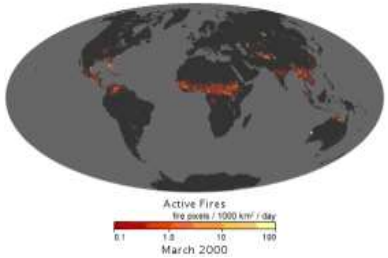
| Low uncertainty | High uncertainty | Out of scope |
|---|---|---|
| Temperature | Global hurricanes | Global fires |
| Pressure | Extremes | Malaria outbreaks |
| Large-scale wind | Precipitation | Landslides |

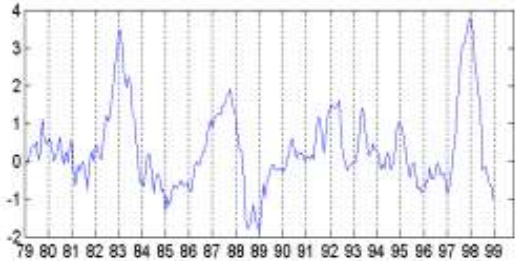
Global sea surface temperatures

Atlantic hurricanes

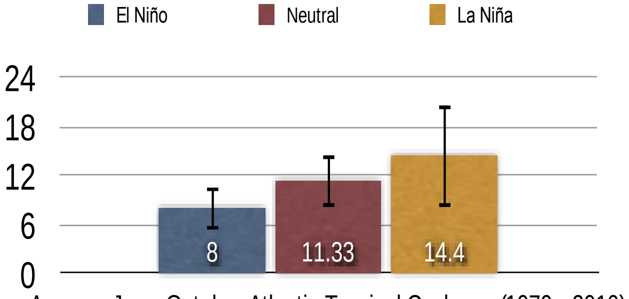
Global fires



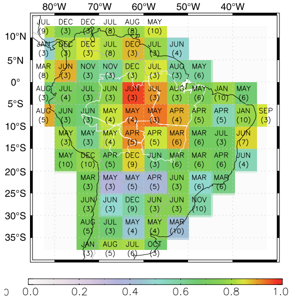




SST Anomaly Time Series in the ENSO region



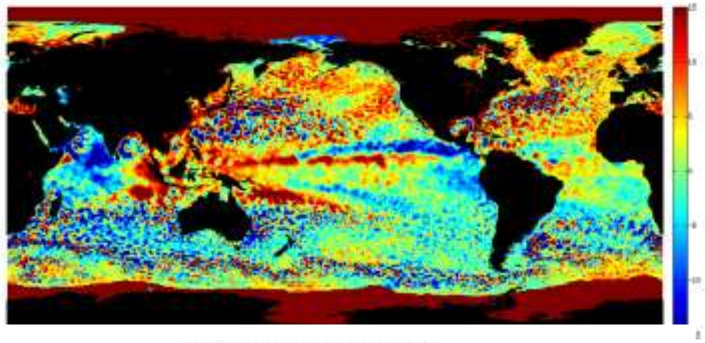
Average June-October Atlantic Tropical Cyclones (1979 - 2010)

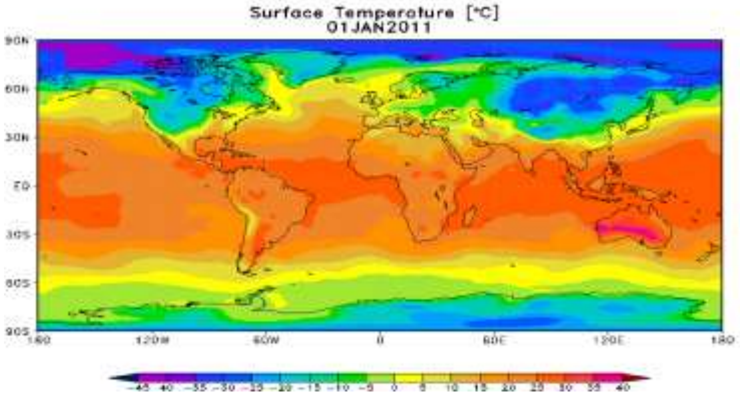


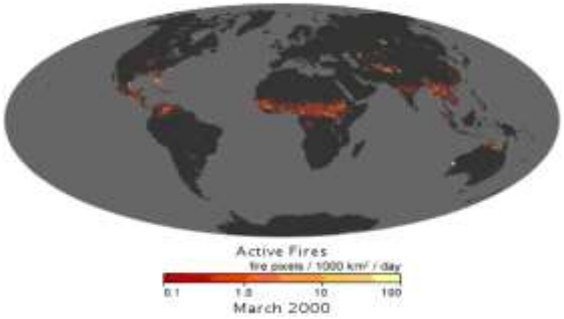
Correlation with fires in Amazon

*Chen et al., Science, 2011*

# Challenges in data driven analysis







- Spatio-temporal auto- and cross-correlation
- Noisy, heterogeneous, and uncertain
- Evolutionary processes
- Multiple spatio-temporal scales
- Unknown, non-linear, and long-range dependency structure
- Variability
- Class imbalance
- Multivariate non-stationary
- Large unlabeled datasets
- Significance testing

Faghmous and Kumar (2013)

# Guiding Theme

The discovery and characterization of *patterns and dependencies* have emerged as the primary research tasks because they…

1. Provide an empirical understanding of physical processes…
   - finding pressure dipole between Tahiti and Darwin led to the understanding of modulation of the Walker Circulation

2. Allow for prediction of unknown quantities…
   - where observations are sparse
   - for statistical downscaling
   - where physical models are inadequate (e.g., predicting the number of hurricanes using a large number of covariates)

3. Enable long-range projection of highly stochastic processes…
   - deriving climate extremes or hurricanes from low-resolution global model simulations

# Project vision and scope

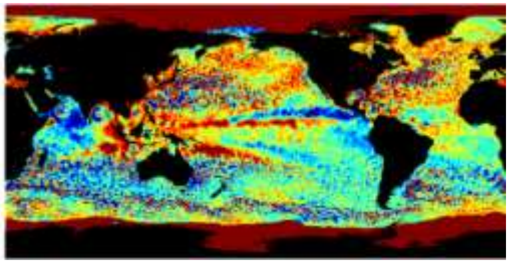
## Transformative Computer Science Research Advancing Climate Change Science

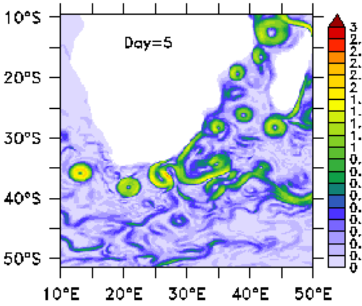
| | | | |
|---|---|---|---|
| Process Understanding | Extreme Events<br>- Heat Waves<br>- Rainfall Extremes<br>- Droughts<br>- Hurricanes<br>Model Evaluation<br>Downscaling<br>- Statistical<br>- Dynamical<br>Ocean-Atm.-Land Interactions | Change Detection<br>- Abrupt vs. Gradual<br>- Point vs. Regions/Intervals<br>- Change in Extremes<br>Spatio-Temporal Classification<br>Sparse/High-Dim. Methods<br>Causal Relationships<br>Networks/Graphs<br>HPC | Computational Innovations |
| **Understanding Climate Change** | | | |

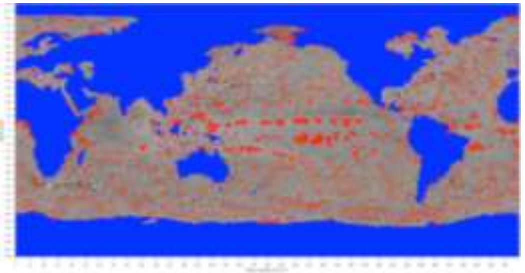
# Pattern Mining: Ocean Eddies Monitoring

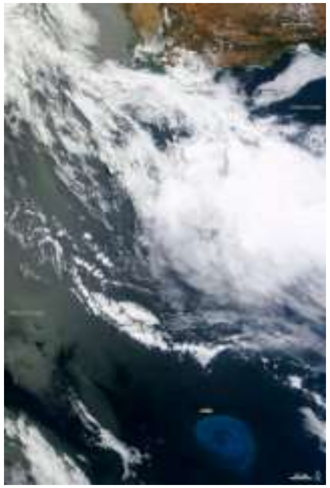
- Scalable spatio-temporal pattern mining algorithms for noisy and continuous data
- Novel multiple object tracking for uncertain features

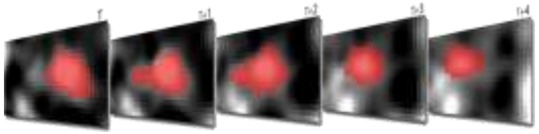
- Detect more accurate features and tracks for improved ocean dynamics monitoring
- Open source data base of 20+ years of eddies and eddy tracks available for scientific applications









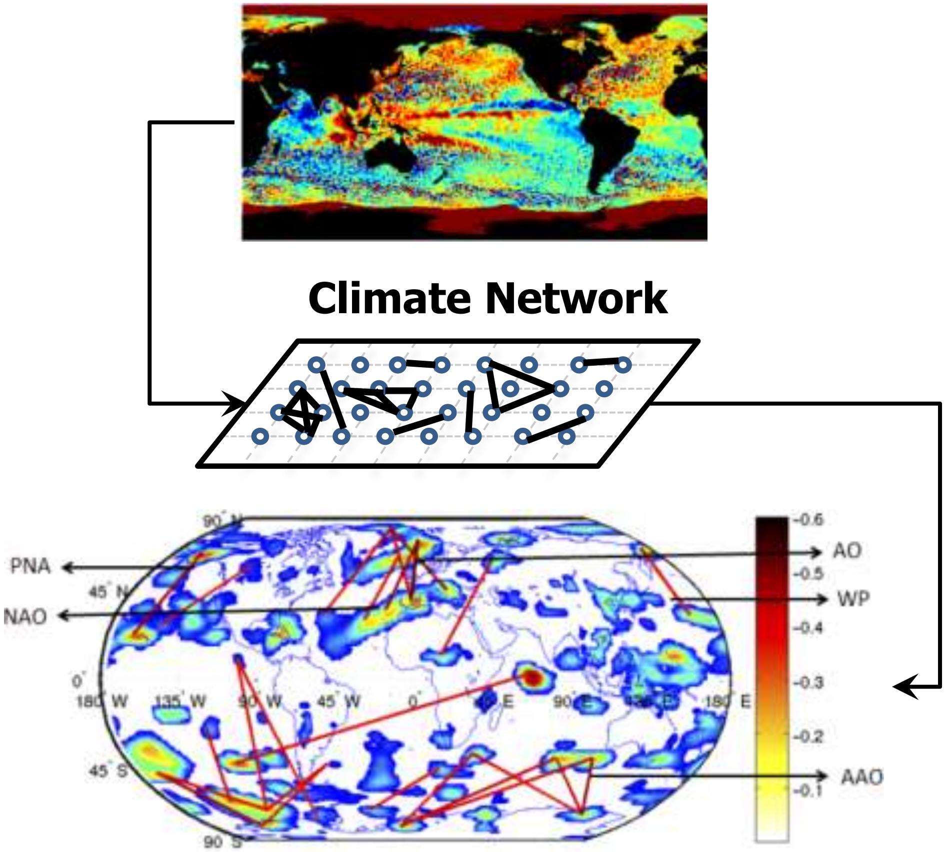


Faghmous et al. *AAAI* (2012a)
Faghmous et al. *CIDU* (2012b) **Best student paper award**
Faghmous et al. *AAAI* (2013)
NSF Nordic Research Opportunity Grant to conduct research at the Bjerknes Centre for Climate Research

# Network analysis: Climate Teleconnections

- Scalable method for discovering anti-correlated graph regions
- Novel dynamic graph clustering for dense directed graphs
- Significance testing for spatio-temporal patterns

- Discovered previously unknown climate teleconnection
- Analyzed climate network properties to better understand global climate dynamics
- Method used to compare climate models



Kawale et al. *SDM*(2011a)
Kawaleet al. *CIDU* (2011b) **Best student paper award**
Kawale et al. *ACM SIGKDD* (2012)
Steinhaeuser et al. *Climate Dynamics* (2012).
SC'11: Exploration in Science through Computation Award
Grace Hopper '12: Winner of the ACM Student Research Competition

# Predictive Modeling: Regression, Ensembles, Inference

- Hierarchical sparse regression: rates of convergence with low samples
- Multi-task learning with spatial smoothing
- Primal decomposition based LP solver for max-cut type problems (~10 million+ node graphs)

- Regional land-climate predictions from observations over oceans
- Combining multiple GCM outputs more accurately than state-of-art
- Mega-drought detection, trends over past 100-1000 years
- An objective method FASI was developed to estimate TC intensity

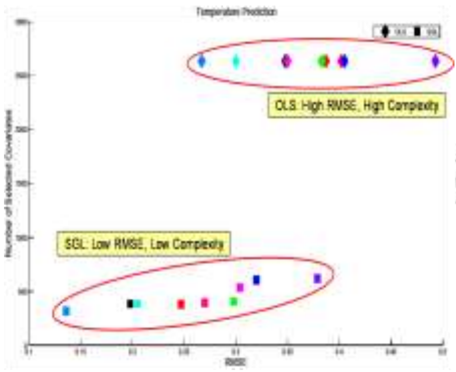
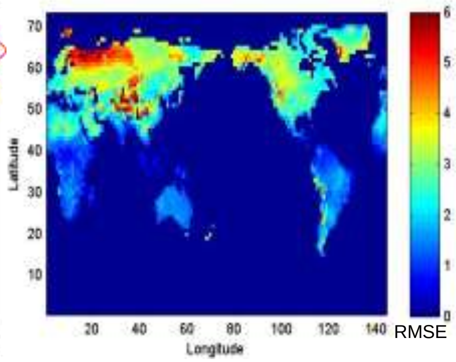
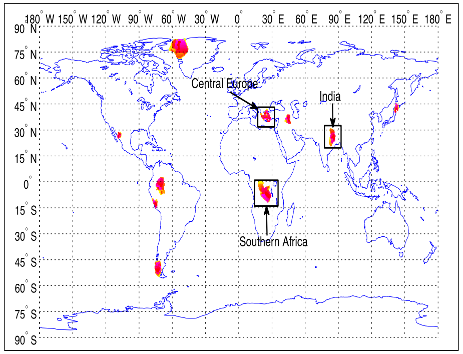


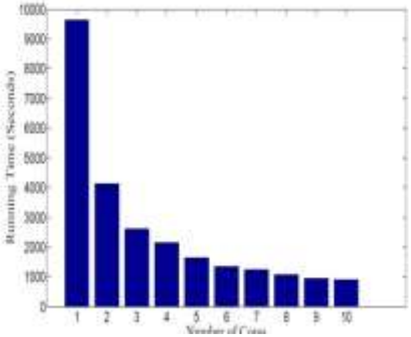
Fig. RMSE vs. Model Complexity of OLS and Sparse Regression Methods



Prediction RMSE from spatially smoothened Multi-model ensemble



Major droughts starting within the period 1981-1995.



Fu et al. *UAI(2013)*

Subbian et al. *SDM(2013)* **Best Application Paper Award**

Hsieh et al. *NIPS(2012)*

Wang et al. *ICML(2012)*

Chatterjee et al. *SDM(2012)* **Best Student Paper Award**

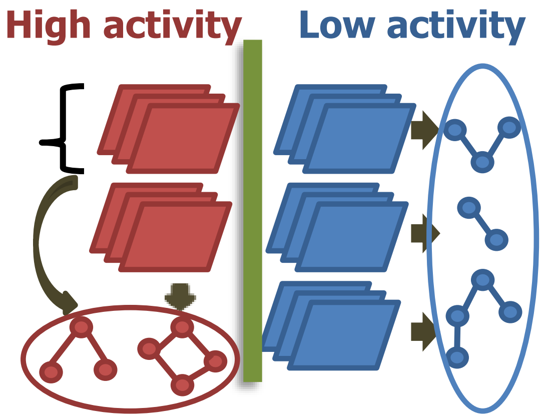
Fu et al. *SDM(2012)*

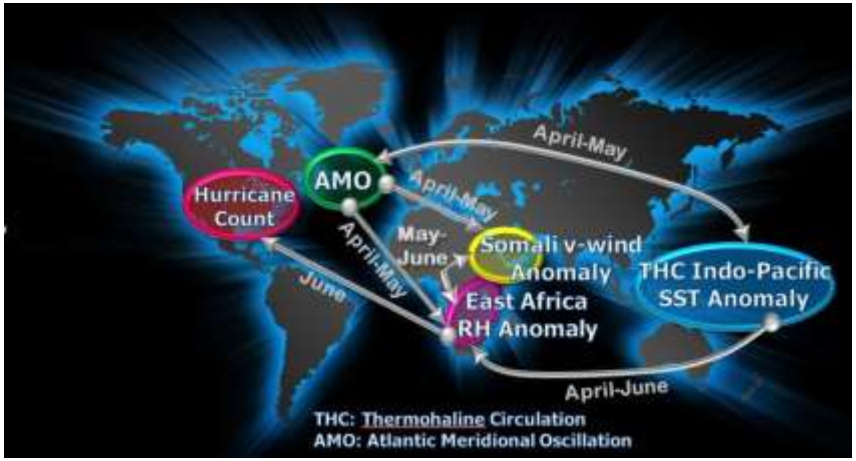
*Fetanat et. al Weather and Forecasting (2013)*

# Relationship mining: Seasonal hurricane activity

- Contrast-based network mining for discriminatory signatures
- Novel dynamic graph clustering for dense directed graphs
- Statistically robust methodology for automatic inference of modulating networks

- Improved forecast skill for seasonal hurricane activity
- Discovered key factors and mechanisms modulating NA hurricane variability
- Discovered novel climate index with much improved correlation with NA hurricane variability: 0.69 vs 0.49



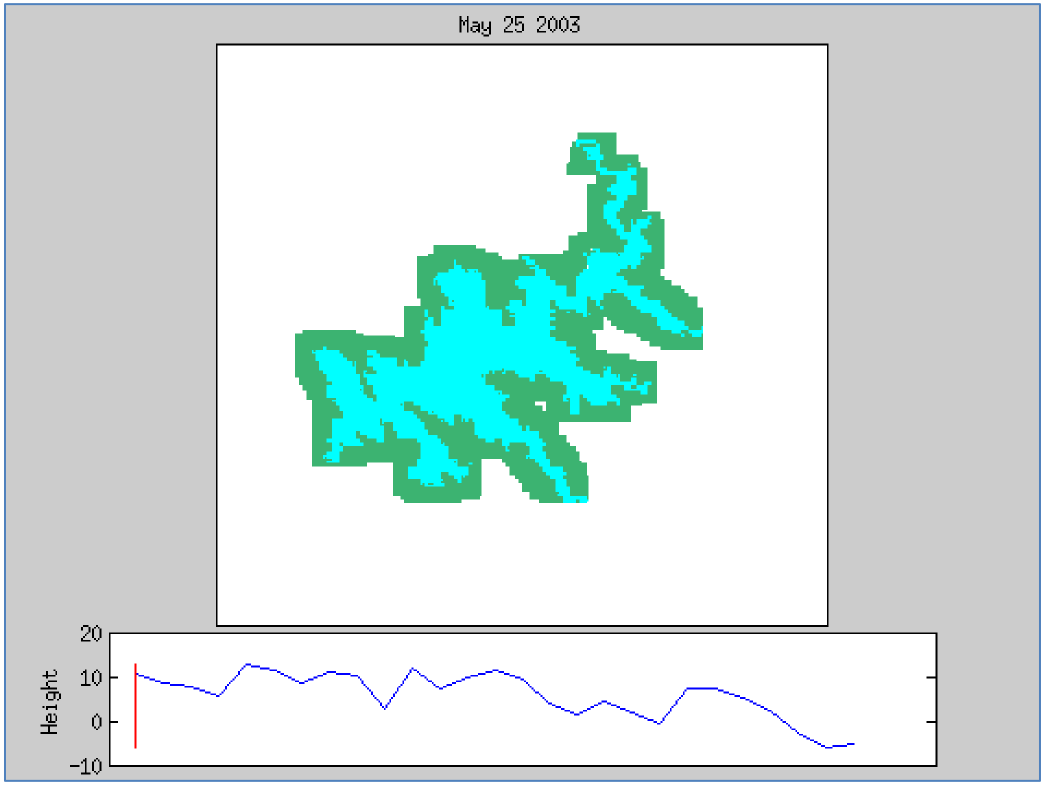


NSF News, DOE Research News, Science360

Sencan et al. *IJCAI* (2011)
Pendse et al. *SIAM SDM* (2012)
Chen et al. *Data Mining & Knowledge Discovery* (2012)
Chen *et al. SIAM SDM* (2013)
Chen *et al. IJCAI* (2013)
Semazzi *et al.* in review at journal (2013)

# Change Detection

- A Subpath Enumeration and Pruning technique to find intervals of abrupt change in collections of time series
- A supervised classification framework to monitor water body dynamics
- Common intervals of change in a collection of time series



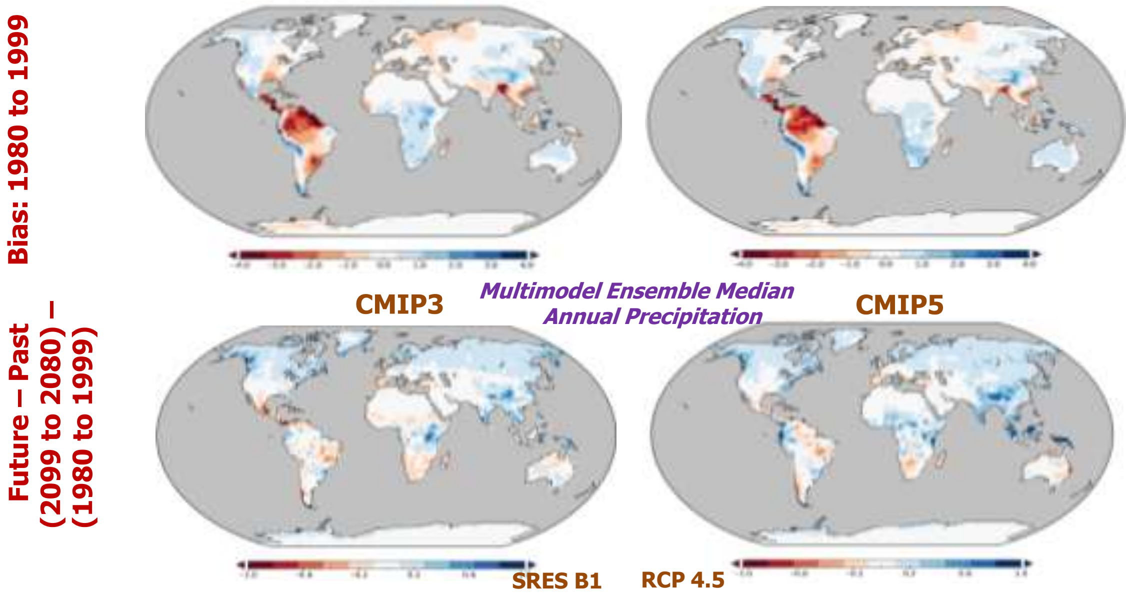
- Improved vegetation representation in climate simulations
- Monitoring of water bodies across the globe for more accurate hydrological modeling and effective water resource management.
- Improved multi temporal land cover classification products using HMM

Zhou, et al., Spatiotemporal (ST) 2013
Mithal, et al. SDM 2013
Mohan, et al., GIScience 2012
Zhou, et al., SIGSPATIAL GIS 2011
Zhou, et al., AGU 2011

# Multi-Model Ensembles

- Progress on the development of a hybrid physics-informed Bayesian scheme.
- Development of a Spatially Smooth Multi-Model Regression model
- Exploration of different techniques for model selection and averaging

- Insights into the strengths and limitations of multi-model ensembles
- Evaluation of CMIP3 vs. CMIP5
- Predict future changes using projected climate information (CMIP5)

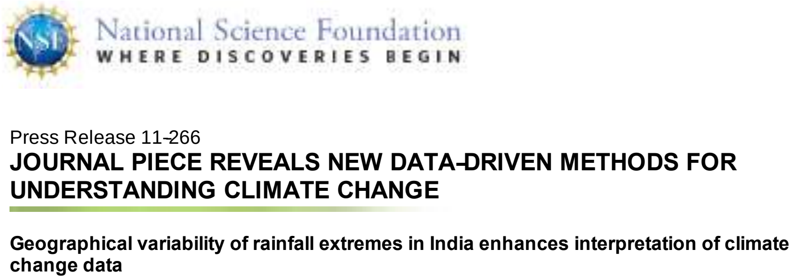

Bias: 1980 to 1999
Future – Past (2099 to 2080) – (1980 to 1999)
CMIP3
Multimodel Ensemble Median Annual Precipitation
CMIP5
SRES B1
RCP 4.5

CMIP5 vs. CMIP3: No significant improvement or change

Chatterjee et al., Second Int’l Workshop on Climate Informatics, 2012
Kodra et al., ICML Workshop on Grand Challenges, 2011
Subbian et al. *SDM, 2013* **Best Application Paper Award**

# Extremes and uncertainty: Heat waves, heavy rainfall, …

- Extreme value theory in space-time and dependence of extremes on covariates
- Mutual information and copula-methods for space-time extremes dependence
- Uncertainty quantification with Bayesian and resampling techniques
- Physics-guided data mining and quantification of uncertainty

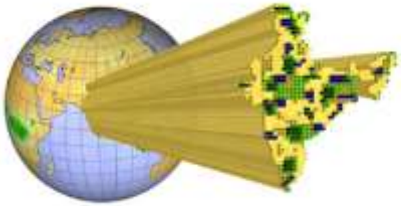


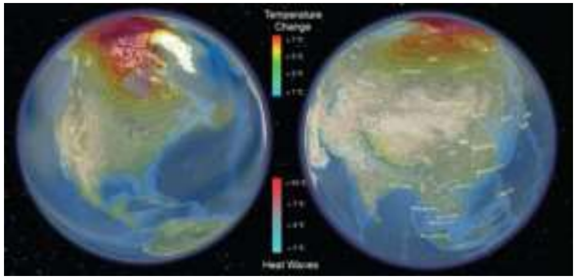
National Science Foundation
WHERE DISCOVERIES BEGIN

Press Release 11-266

**JOURNAL PIECE REVEALS NEW DATA-DRIVEN METHODS FOR UNDERSTANDING CLIMATE CHANGE**

**Geographical variability of rainfall extremes in India enhances interpretation of climate change data**





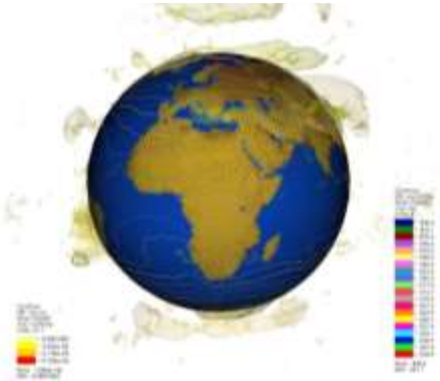
- Spatiotemporal trends in heat waves, cold snaps, and heavy rain with climate change
- Climate model evaluation and physics-guided uncertainty quantification
- Covariate-based improvement of extremes projections under climate change
- Translation to adaptation and stakeholder relevant metrics

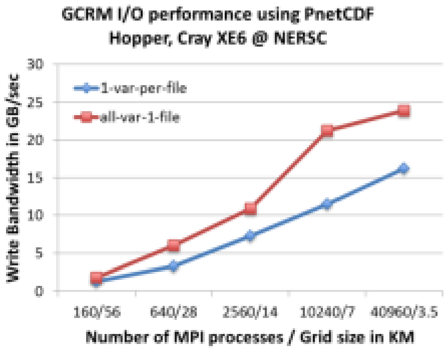
Ghosh *et al.* Nature Climate Change (2012)
Parish *et al.* Computers & Geosciences (2012)
Kodra *et al.* Environmental Research Letters (2012)
Ganguly *et al.* Climate Extremes & UQ: Book Ch. (2013)
Kodra *et al.* in revision at journal (2013)
Kumar *et al.* in review at journal (2013)

# High Performance Tools and Methods

- Created a library of common data mining / machine learning kernels for clustering, classification, PCA, etc.
  - Many algorithms have shown speedups of two to three orders of magnitude.
- Developed technologies for compressing and querying huge datasets, and for performing similarity searches with a more than 10-fold speed-up
- Devised an image indexing technique based on a new Locality Sensitive Hashing (LSH) scheme.
- Developing HPC solutions for our collaborators, including bootstrapping methods for extreme value prediction and Markov Random Field based abrupt change detection
- Developed a new scheme known as Fast Locality Sensitive Hashing(FLSH) algorithm, which outperforms the existing LSH scheme has been developed for satellite image-base retrieval.

Improving I/O for the Global Cloud Resolving Model




GCRM I/O performance using PnetCDF
Hopper, Cray XE6 @ NERSC
Write Bandwidth in GB/sec
30
25
20
15
10
5
0
1-var-per-file
all-var-1-file
160/56
640/28
2560/14
10240/7
40960/3.5
Number of MPI processes / Grid size in KM

Jin *et al*. EuroMPI (2011)
Patwary *et al*. SC (2012)
Hentrix *et al*. HPC (2012)
Kumar *et al*. IPDPS (2011)
Rangel *et al.* in review (2013)
Jin *et al.* in review (2013)
*Buaba et. al* IJCA (2013)

# Recent Best Papers

- Sriram Lakshminarasimhan, et al. won the Best Paper Award at the 22nd International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC13) for their paper, titled: ***Scalable In Situ Scientific Data Encoding for Analytical Query Processing***.
- Karthik Subbian and Arindam Banerjee won the Best Application Paper Award at the 2013 SIAM International Conference on Data Mining (SDM13) for their paper, titled: **Climate Multi-Model Regression Using Spatial Smoothing**.
- James Faghmous, et al. won the Best Student Paper Award at the 2012 IEEE Conference on Intelligent Data Understanding for their paper ***EddyScan: A Physically Consistent Global Ocean Eddy Monitoring Application***.
- C. Jin, et al. won the Best Paper Award at the BigMine'13 workshop (KDD13) for their paper, titled: ***Solving Combinatorial Optimization Problems using Relaxed Linear Programming: A High Performance Computing Perspective***

# Other Notable Activities

- Shashi Shekhar appointed a member of the NRC Computer Science and Telecommunications Board committee on Geotargeted Disaster Alerts and Warnings.
- Auroop Ganguly and his Sustainability and Data Sciences (SDS) Lab were highlighted as Faces of NSF Research in the NSF Current monthly newsletter.
- Research on more accurate predictions of hurricane activity, led by Nagiza Samatova and Frederick Semazzi, was highlighted by the NSF Science360 News Service.
- Discussions are underway between NOAA and Abdollah Homiafar for the “operational transition” of techniques he and his team developed for hurricane intensity estimation.
- Vipin Kumar invited to become a member of the World Economic Forum Global Agenda Council on Measuring Sustainability.

# Education/Outreach Activities

- Undergraduate and graduate courses/programs  at the intersection of climate and data sciences
- Cross disciplinary training environment
- Extensive research opportunities for students from historically underrepresented groups
- Fostering interdisciplinary collaborations by organizing workshops and sessions at climate and computer science venues
- Engagement with UNEP, IPCC, World Economic Forum

Annual workshop

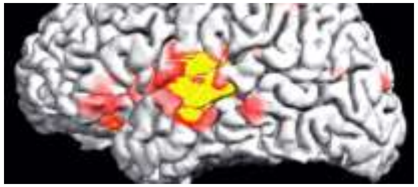
Researchers Devise More Accurate Method For Predicting Hurricane Activity

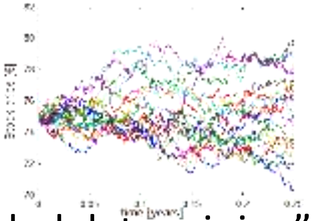
Journal Piece Reveals New Data-driv Understanding Climate Change

Climate Prediction Community Interface

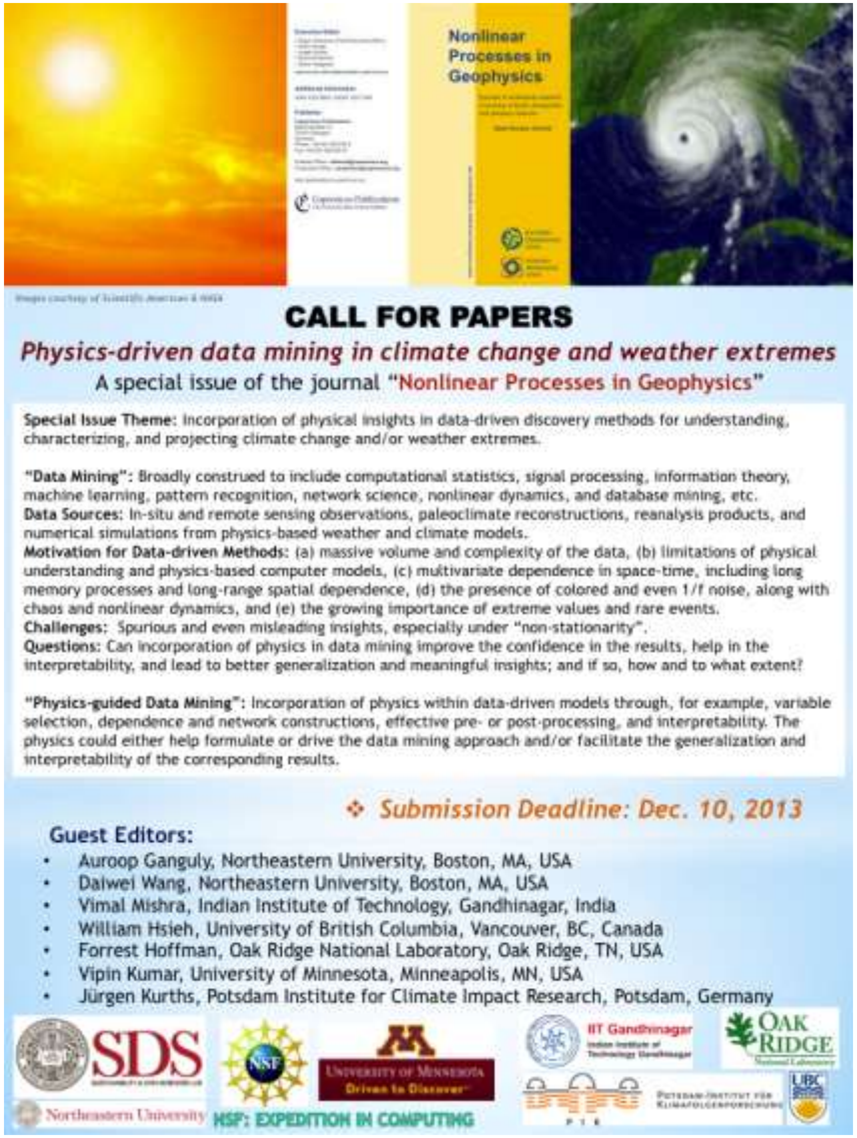
# Future Directions and Goals

- **Computer science**
  - Transformative research opportunities in data mining and machine learning
    - Complex dependence and noise structures within nonlinear dynamical spatiotemporal system
    - Focus of climate community on extremes and change
    - Data size from few petabytes 350 petabytes by 2030
  - Transformative spatio-temporal  methods can generalize to multiple domains
    - fMRI
    - Computational finance
    - Insurance





  - Develop a new suite of approaches in “physics-guided data mining”
- **Climate inroads**
  - Embed new methods and insights into assessment processes used by climate organizations, for example, IPCC, US GCRP, WMO, NOAA, etc.
  - Help establish the field of “climate informatics” over the next 5-10 years
    - Embed data mining or machine learning in suite of methods for climate scientists
    - Help climate scientists understand when to use what tools developed in statistics, information theory, data mining, machine learning, and high performance computing
    - Develop exemplary papers in high-impact climate and interdisciplinary journals with computational data-driven approaches



Nonlinear
Processes in
Geophysics

Images courtesy of Scientific American & NASA

# CALL FOR PAPERS

## *Physics-driven data mining in climate change and weather extremes*

A special issue of the journal "Nonlinear Processes in Geophysics"

**Special Issue Theme:** Incorporation of physical insights in data-driven discovery methods for understanding, characterizing, and projecting climate change and/or weather extremes.

**"Data Mining":** Broadly construed to include computational statistics, signal processing, information theory, machine learning, pattern recognition, network science, nonlinear dynamics, and database mining, etc.
**Data Sources:** In-situ and remote sensing observations, paleoclimate reconstructions, reanalysis products, and numerical simulations from physics-based weather and climate models.
**Motivation for Data-driven Methods:** (a) massive volume and complexity of the data, (b) limitations of physical understanding and physics-based computer models, (c) multivariate dependence in space-time, including long memory processes and long-range spatial dependence, (d) the presence of colored and even 1/f noise, along with chaos and nonlinear dynamics, and (e) the growing importance of extreme values and rare events.
**Challenges:** Spurious and even misleading insights, especially under "non-stationarity".
**Questions:** Can incorporation of physics in data mining improve the confidence in the results, help in the interpretability, and lead to better generalization and meaningful insights; and if so, how and to what extent?

**"Physics-guided Data Mining":** Incorporation of physics within data-driven models through, for example, variable selection, dependence and network constructions, effective pre- or post-processing, and interpretability. The physics could either help formulate or drive the data mining approach and/or facilitate the generalization and interpretability of the corresponding results.

❖ *Submission Deadline: Dec. 10, 2013*

**Guest Editors:**

- Auroop Ganguly, Northeastern University, Boston, MA, USA
- Daiwei Wang, Northeastern University, Boston, MA, USA
- Vimal Mishra, Indian Institute of Technology, Gandhinagar, India
- William Hsieh, University of British Columbia, Vancouver, BC, Canada
- Forrest Hoffman, Oak Ridge National Laboratory, Oak Ridge, TN, USA
- Vipin Kumar, University of Minnesota, Minneapolis, MN, USA
- Jürgen Kurths, Potsdam Institute for Climate Impact Research, Potsdam, Germany