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

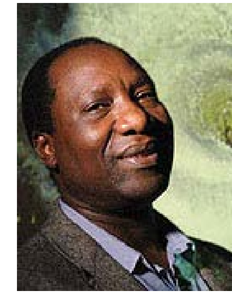
Expeditions Team


Vipin Kumar, UM


Auroop Ganguly, NEU

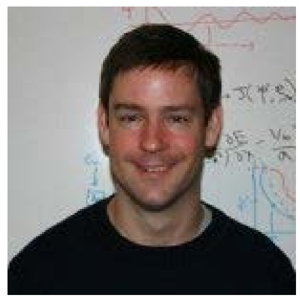

Nagiza Samatova, NCSU


Arindam Banerjee, UM


Fred Semazzi, NCSU


Joe Knight, UM


Shashi Shekhar, UM


Peter Snyder, UM


Jon Foley, UM

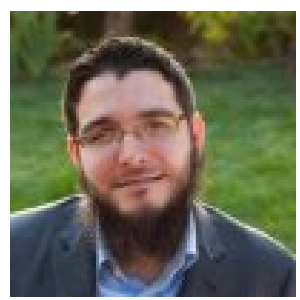

Alok Choudhary, NW


Ankit Agrawal, NW


Abdollah Homiafar
NCA&T


Michael Steinbach
UM


Singdhansu Chatterjee
UM

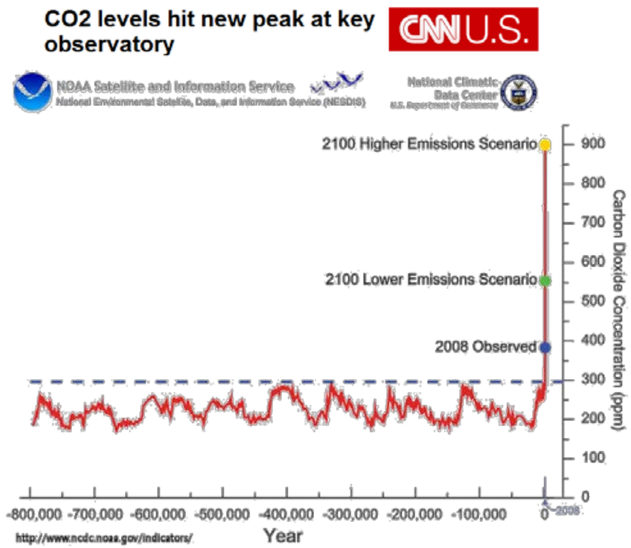

James Faghmous
UM

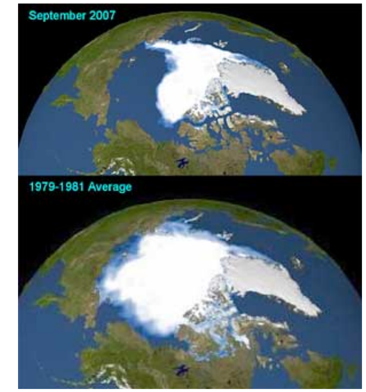

Stefan Liess
UM


Shyam Boriah
UM

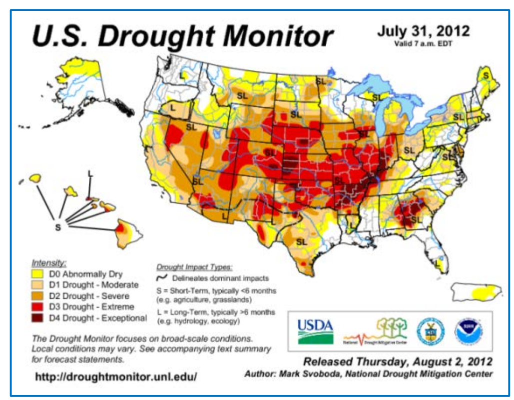
Understanding Climate Change - Motivation

- **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



Russia Burns, Moscow Chokes
NATIONAL GEOGRAPHIC, 2010

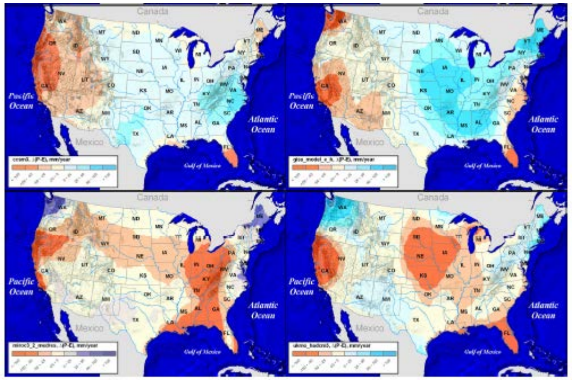
<u>Key outstanding science challenge:</u>

Actionable predictive insights to credibly inform policy

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

Disagreement between IPCC models



Regional hydrology exhibits large variations among major IPCC model projections

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

(Nature, 2010)

Physics based models

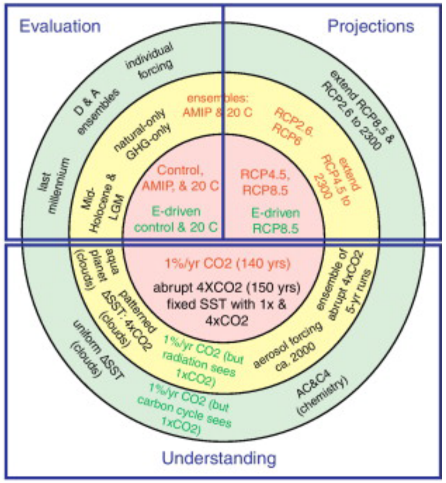
Low uncertainty (High confidence)	High uncertainty (Low confidence)	Out of scope
Temperature	Precipitation	Forest fires
Pressure	Hurricanes	Malaria outbreaks
Large-scale wind	Extremes	Landslides

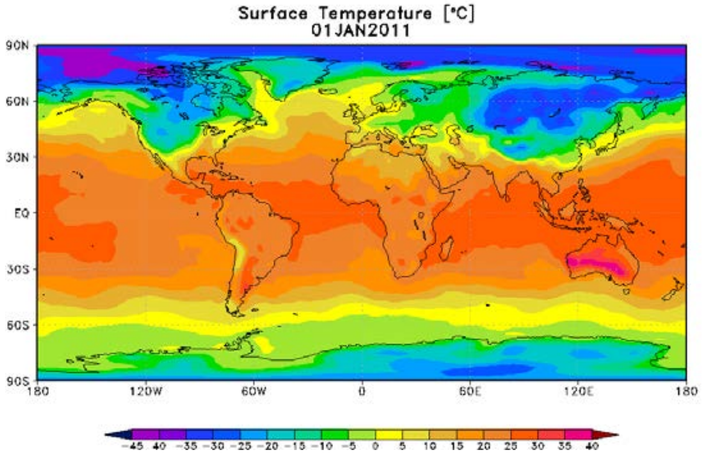
Big Data in Climate Science

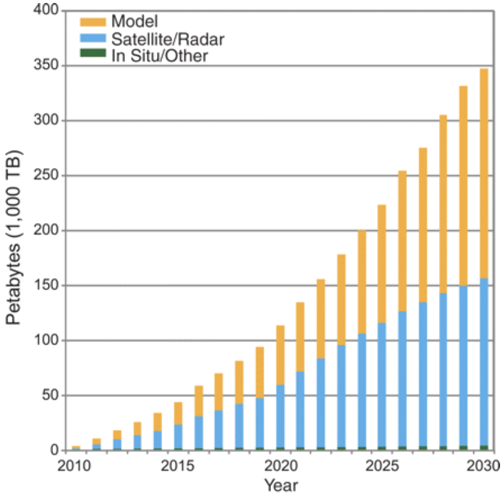
Transformation from Data-Poor to Data-Rich

- Sensor Observations
- Reanalysis Data
- Model Simulations

Source: NASA





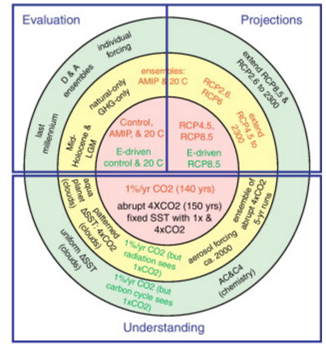


Source: Overpeck et al., *Science*, (2011)

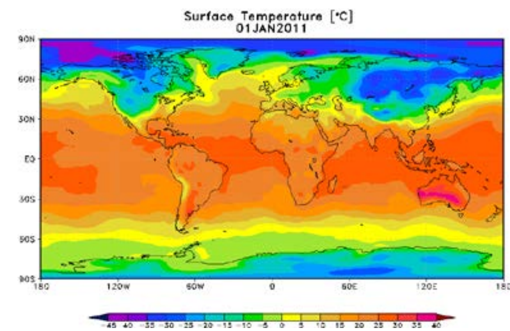
Big Data in Climate Science

Transformation from Data-Poor to Data-Rich

- Sensor Observations
- Reanalysis Data
- Model Simulations



"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)



White House Brings Together Big Data & Climate Change

CLIMATE CENTRAL
March 19, 2014

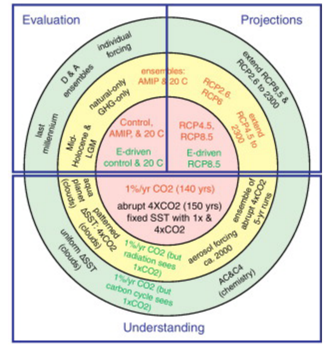
SCIENTIFIC AMERICAN™

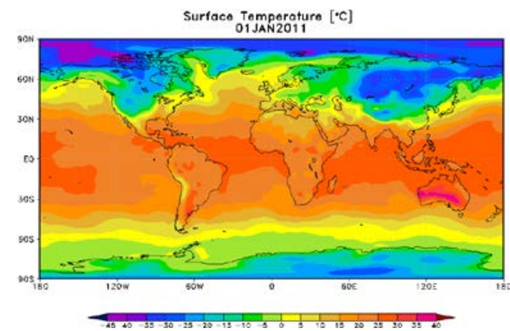
Can Big Data Help U.S. Cities Adapt to Climate Change?
March 20, 2014

Big Data in Climate Science

Transformation from Data-Poor to Data-Rich

- Sensor Observations
- Reanalysis Data
- Model Simulations





This Expedition aims to develop a new and transformative data-driven approach that:

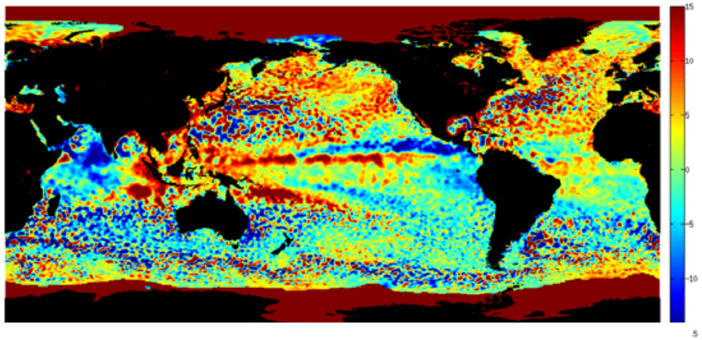
- Makes use of wealth of observational and simulation data
- Advances understanding of climate processes
- Informs climate change impacts and adaptation

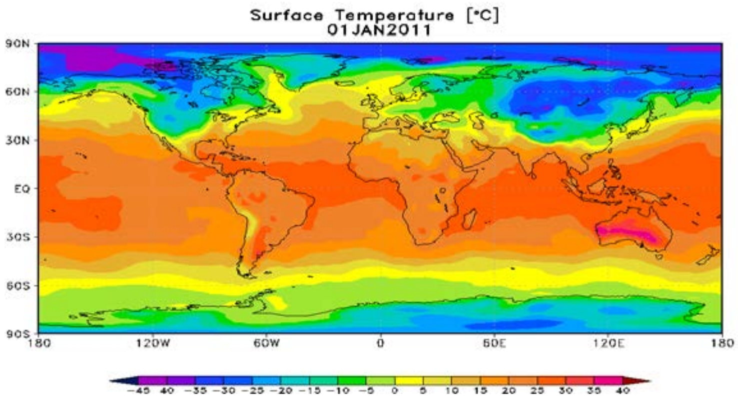
Project vision and scope

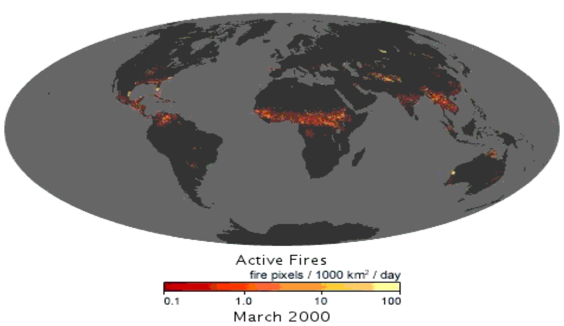
Transformative Computer Science Research
Advancing Climate Change Science

<table>
<tr>
<td>Process Understanding</td>
<td>Extreme Events
- Heat Waves
- Rainfall Extremes
- Droughts
- Hurricanes
Model Evaluation
Downscaling
- Statistical
- Dynamical
Ocean-Atm.-Land Interactions</td>
<td>Change Detection
- Abrupt vs. Gradual
- Point vs. Regions/Intervals
- Change in Extremes
Spatio-Temporal Classification
Sparse/High-Dim. Methods
Causal Relationships
Networks/Graphs
HPC</td>
<td>Computational Innovations</td>
</tr>
<tr>
<td colspan="4">Understanding Climate Change</td>
</tr>
</table>

Challenges in data-driven analysis of climate data







- 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)

Computer

Cross-cutting Theme: Theory Guided Data Mining

Computer

Theory-Guided Data Science for Climate Change

James H. Faghmous , Arindam Banerjee , Shashi Shekhar , Michael Steinbach Vipin Kumar, Auroop R. Ganguly Nagiza Samatova

Nonlinear Processes in Geophysics

An interactive open-access journal of the European Geosciences Union

Physics-driven data mining in climate change and weather extremes

Editor(s): A. Ganguly, V. Mishra, D. Wang, W. Hsieh, F. Hoffman, V. Kumar, and J. Kurths

A BIG DATA GUIDE TO UNDERSTANDING CLIMATE CHANGE:

The Case for Theory-Guided Data Science

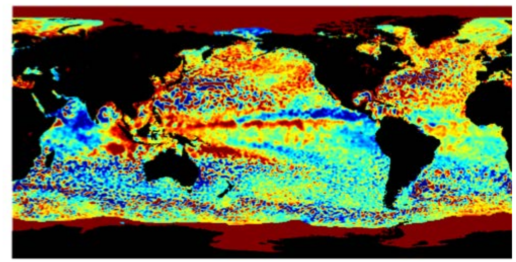
James H. Faghmous and Vipin Kumar

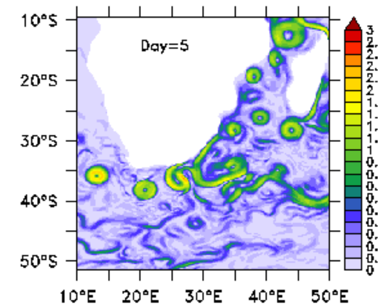
Highlights

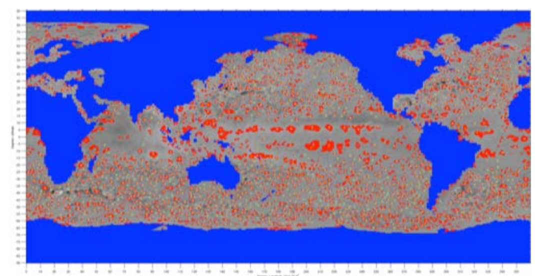
- Highly inter-disciplinary research:
  - Computer science, hydrology, earth sciences, statistics, civil engineering
- ~ 150 publications (journals, conferences, and workshops) with authors from multiple disciplines
- A number of best paper and outstanding dissertation awards
- Public release of software & 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
- Interdisciplinary community engagement: Computer science, engineering, physical sciences, and social or economic sciences

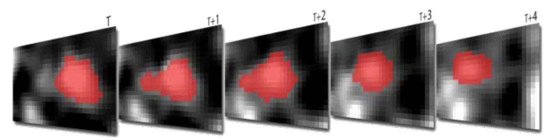
Pattern Mining: Ocean Eddies Monitoring

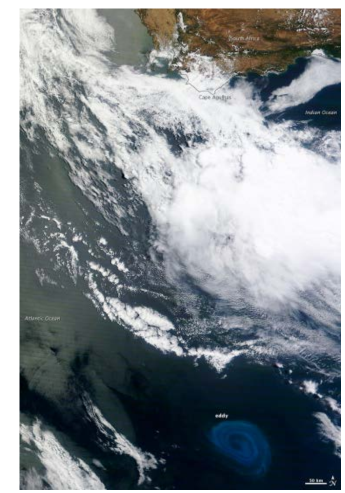
- Rotating coherent structures that are sources of intense physical and biological activity
- Identifying ocean eddies in satellite products is an active subject of research
- Used spatio-temporal context of the data to extract statistically significant features
- Open source data base of 20+ years of eddies and eddy tracks available for scientific applications
- Being used by oceanographers worldwide
- Enabled study of interactions between hurricanes and ocean eddies









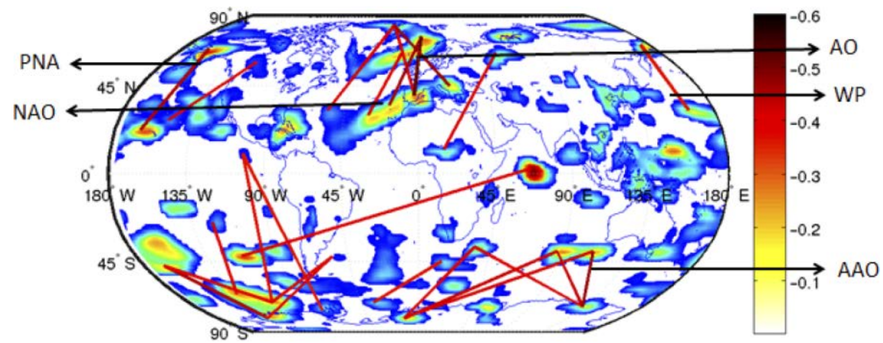


Faghmous et al. *AAAI* (2012a)
Faghmous et al. *CIDU* (2012b) **Best student paper award**
Faghmous et al. *AAAI* (2013)

Network analysis: Climate teleconnections

- Large-scale long-range relationships play a crucial role in the atmosphere
- Discovering such relationships is laborious and imprecise
- Developed automated procedures to discover teleconnections in large climate data sets
- Technique discovered more robust relationships, new undiscovered relationships, and is a tool to evaluate climate models
- Data-driven Inference of modulatory networks in climate science

Automatic Discovery of Dipoles from SLP data



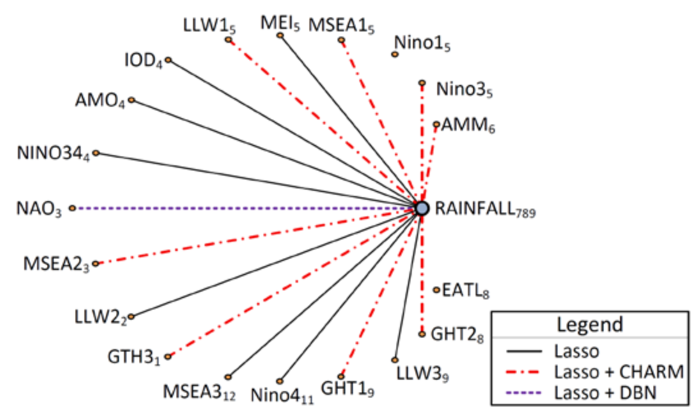
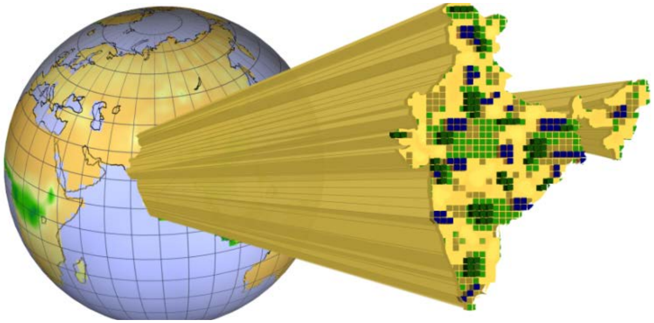


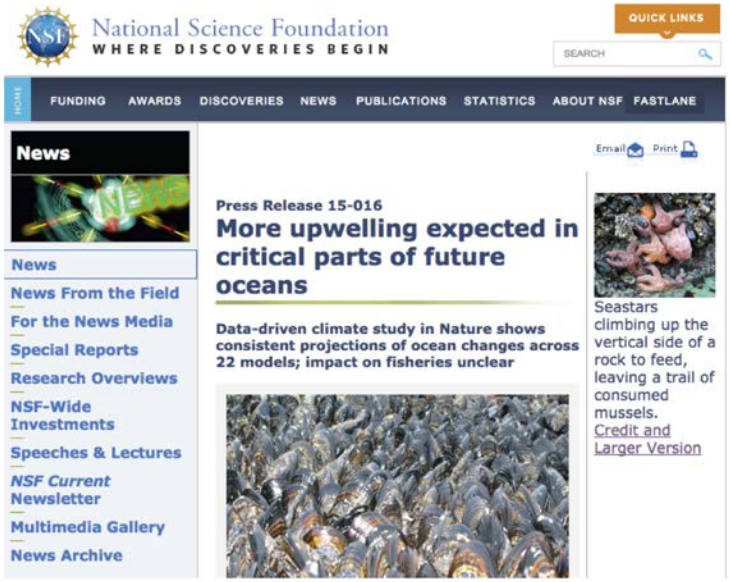
Figure 9. Relationships directly associated with rainfall for $\lambda = \text{EATL}_8$.

Kawale et al. *SDM*(2011a), *CIDU* (2011b) **Best student paper award,** *ACM SIGKDD* (2012)
Liess et al: Journal of Climate (2014)
Lu et al: under review Journal of Climate)
González et al. Nonlinear Processes in Geophysics Discussions, 2015
Samatova et al. SDM 2012, ICDM 2013, PKDD 2015

Extremes and uncertainty: Heat waves, heavy rainfall, …

- Climate change has been called "*global weirding*" owing to possible exacerbation of hydro-meteorological extremes and changes in regional weather patterns
- Understanding relevant dominant processes and discovering space-time dependence requires *physics-guided data mining* and *uncertainty quantification*
- Computational data and geospatial sciences can help translate insights from models and observations to metrics relevant for *adaptation and policy*
- Suite of methods developed new insights on climate extremes with uncertainty, and consequences for *critical infrastructures and key resources*





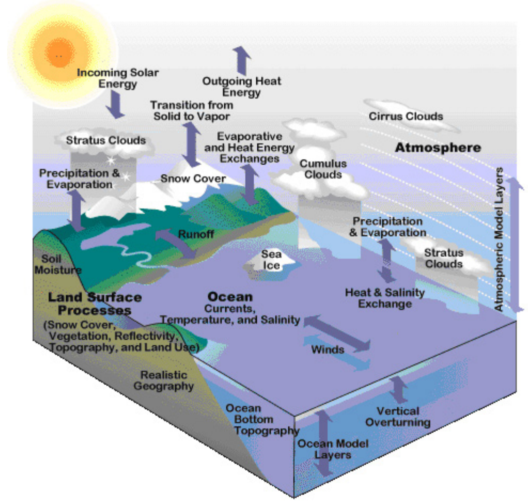
Ghosh *et al.* Nature Climate Change (2012)
Parish *et al.* Computers & Geosciences (2012)
Kodra *et al.* Environmental Research Letters (2012)
Ganguly *et al.* Climate Extremes & UQ: Book Chapter (2013)
Kodra *et al* Scientific Reports, Nature (2014)
Kumar *et al.* Climate Dynamics (2014)

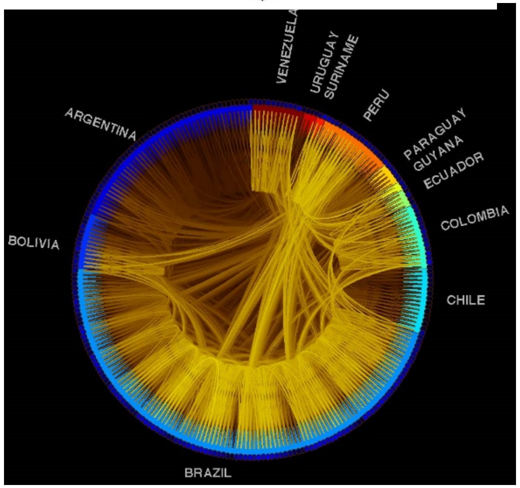
Predictive Modeling for Climate Data

- Regression for High dimensions, low sample setting
  - Statistical consistency using sparse hierarchical regularization
  - Examples: Sparse group lasso, Gaussian Markov random fields
- Multi-task learning with spatial smoothing
  - Combining outputs of multiple GCMs
  - Local regression with spatial smoothing
  - More accurate than local or global regression
- Inference using discrete graphical models (MRFs)
  - Mega-drought detection, trends over past 100-1000 years
  - Fast parallel algorithm: 10 mins on 20 cores, 30 secs on 500 cores
  - Detects all major droughts, model evaluation in progress
- Rare class prediction in the absence of ground truth
  - Used for mapping of forest fires globally
- Spatial Decision Tree (SDT) concept and SDT learning algorithms
  - Incorporates explicit Physics (e.g., continuity constraint)
  - Uses focal-feature based test and spatial-information-gain based objective functions.

SDM 2012, Best Student paper award
NIPS 2012
SDM 2013, Best Application paper award
ICML 2012
SDM 2012, 2013, 2014, 2015
UAI 2013
IJCAI 2015

Global Climate Models (GCMs)



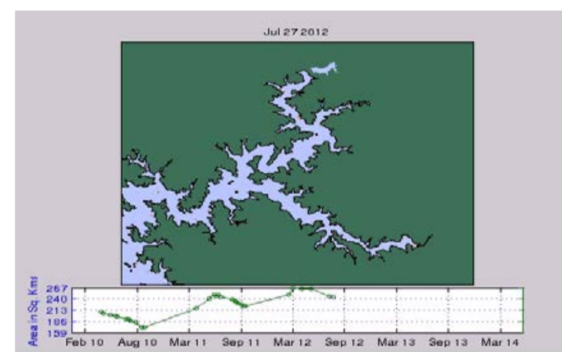


Temperature dependencies between Regions in South America

Change Detection

- The automatic detection of changes in large spatio-temporal datasets is important to monitoring and understanding the behavior of the global climate system.
- Abrupt Change Interval Miner (ACIM) and Sub-path Enumeration and Pruning (SEP)
  - Discovers interesting spatio-temporal sub-paths/intervals, such as those with abrupt changes.
  - These algorithms leverage explicit Physics (e.g., continuity violations)
- Detecting change points in non-stationary time series based on genetic algorithms
  - Can search for a global solution in the large search space of a non-convex constrained optimization.
  - Using information criteria methods, the optimal number of change points and clusters can be determined
- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Physics-guided approaches for global monitoring of changes in surface water bodies

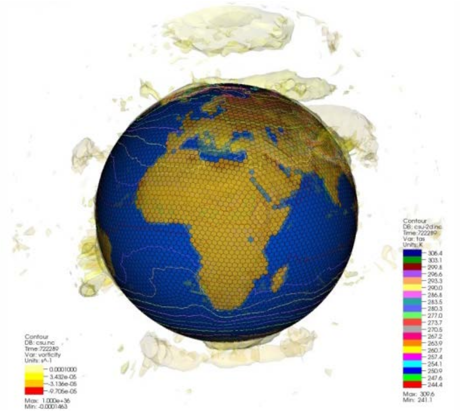
Land cover change detection

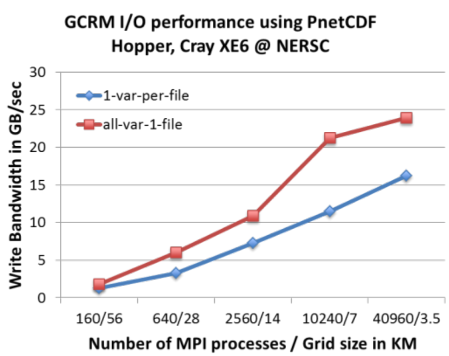


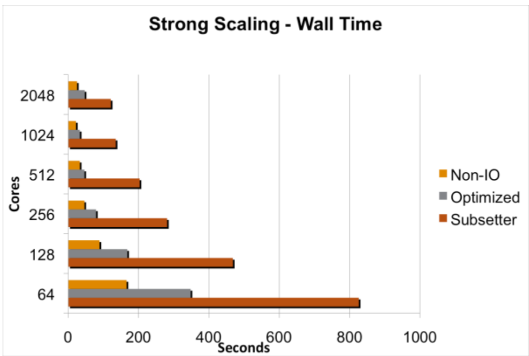
Monitoring the dynamics of surface water bodies

Scaling I/O and analytics: Global cloud resolving model

- Global Cloud Resolving Model (GCRM): simulates circulation associated with large convective clouds
- I/O was previously a major *bottleneck* for GCRM: **1.4 PB** data per simulation and **1.5 TB** per checkpoint
  - Improved I/O throughput Using *Parallel NetCDF I/O library* optimizations, massive scalability
  - Optimized memory utilization and process communication







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 at journal (2013)
Jin *et al.* in review at journal (2013)

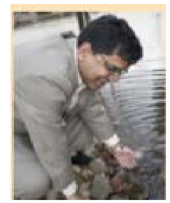
Sample of Education and Outreach Activities

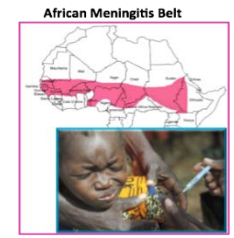
Education

- Undergraduate and graduate courses/programs at the intersection of climate and data sciences
  - Graph Mining and Real-Time Data Stream Analytics
  - Climate Statistics
  - Coursera MOOC, "From GPS and Google Maps to Spatial Computing", reached 21,844 across
  - 182 countries
- Cross disciplinary training environment

Professional Science Master's degree program in Climate Change & Society

Engagement with UNEP, IPCC and World Economic Forum and wider climate science and impact community





Application to Climate:
Meningitis Problem over West Africa

science360 News Service
Breaking science news that shapes your world

Breaking Story
Researchers Devise More Accurate Method For Predicting Hurricane Activity

Annual Workshop
Attended by 70-100 researchers from multiple disciplines

Nurturing a *"Climate Informatics"* Community

"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)**

Workshops and sessions in climate & computer science venues

CIDU IEEE acm CI NCAR AGU EGU

Special Issues

Nonlinear Processes in Geophysics
Physics Driven Data Mining

IEEE Computing in Science & Engineering Magazine
Computing in Climate: Challenges and Opportunities