Understanding Climate Change: 
A Data-Driven Approach

Alok Choudhary, Northwestern University
Nagiza F. Samatova, NC State and ORNL

choudhar@eecs.northwestern.edu
samatova@cs.ncsu.edu
Example Use Cases: Extreme Events Prediction

NH Tropical Cyclone (TC) Activity

Climate-Meningitis Outlook

Forecasting NA Hurricane Tracks
Climate systems are complex because of non-linear coupling of its subsystems (e.g., the ocean and the atmosphere).

**Challenge:**
How to “connect the dots”, that is, to construct *predictive phenomenological* models explaining *structure-dynamics-function relationships* in the complex climate system.

From Simplicity to Complexity
*Science 3 September 2010: 1125.*
Modeling a Climate System as a Network

Climate Data

Anomaly time series at each node

Climate Network

Edge weights: significant correlations
Nodes in the graph: grid points on the globe

Correlation between two anomaly time series

Stat. significant correlations

Extreme Phase

Normal Phase

Multivariate Networks

Multiphase Networks

Slide 5
Subgraphs Common to Extreme Event Climate Networks

Networks for Climate Systems during Extreme Events

Networks for Climate Systems during Normal Events
Identifying patterns in the evolution of the climate system – Example: Analysis of Decadal Trends in Climate

1. Data processing to reduce seasonality
2. Division of data into overlapping decadal time windows
3. Analysis of dependencies
4. Analysis of climate network evolution using stable clusters
5. Characterization of the climate networks through clustering
6. Construction of decadal climate networks by applying correlation threshold
Enabling large-scale data-driven science for complex, multivariate, spatio-temporal, non-linear, and dynamic systems:

End-to-end demonstration of this major paradigm for future knowledge discovery process.

**Complex Networks**
- Study collective behavior of interacting climate subsystems

**Relationship Mining**
- Discovery of complex dependence structures such as non-linear relationships

**Predictive Modeling**
- Model typical and extreme behavior from multivariate spatio-temporal data

**High Performance Computing**
- Efficient analytics on future generation exascale HPC platforms with complex memory hierarchies

Crucial
A Complementary Interplay of R&D Portfolios

Applications

Parallel R

Parallel netCDF

Basic Research

Hypotheses & Discoveries

Power-aware Analytics

Climate Extremes

Complex Networks

Prototype

HPC Production
Illustrative Case for HPC: CMIP3 → CMIP5

- Coupled Model Inter comparison Project
- Spatial resolution: 1 – 0.25 degrees
- Temporal resolution: 6 hours – 3 hours
- Models: 24 - 37
- Simulation experiments: 10s - 100s
  - Control runs & hindcast
  - Decadal & centennial-scale forecasts
- Covers 1000s of simulation years
- 100+ variables
- 10s of TBs to 10s of PBs

Summary of CMIP5 model experiments, grouped into three tiers
Scaling I/O and Analytics

- **Global Cloud Resolving Model (GCRM)**
  - Simulate circulation associated with large convective clouds
  - Developed by David Randell (Colorado State U) & Karen Schuchardt (PNNL)

- **Geodesic grid model**

- **1.4 PB data per simulation**
  - 4 km resolution, 3 hourly, 1 simulated year
  - 1.5 TB per checkpoint

- **Parallel NetCDF I/O library outreaches climate community under NSF Expeditions in Computing project**

I/O was previously a major bottleneck: The only reason execution at this scale became possible was due to I/O scaling.
Illustrative Results

- Improved I/O throughput
  - Using PnetCDF optimizations, massive scalability
  - For 3.5 km grid resolution, grid size is 41.9M cells with 256 vertical layers
  - Data analysis read and simulation checkpoint

![GCRM I/O performance using PnetCDF](image)

![Strong Scaling - Wall Time](image)
Taking Climate Science to the Next Level with HPC-Illustration

- **Our HPC goals are enabling data analysis at:**
  - Higher spatial or temporal resolution
    - Precipitation extremes analysis
    - Network-based hurricane prediction
    - Estimation of spatiotemporal dependence
  - **Higher data dimensionality**
    - Bayesian analysis of multi-model ensembles
    - Sampling-based statistical methods
    - Multivariate quantile analysis
  - **Greater complexity per data point**
    - Estimation of complex dependence structures
    - Handling non-stationarity
    - Multi-resolution analysis
  - **Shorter response time**
    - Interactive hypothesis testing

Significant correlations for hurricane prediction
(Sencan, Chen, Hendrix, Pansombut, Semazzi, Choudhary, Kumar, Melechko, and Samatova, 2011)

Prediction of land climate using ocean climate variables
(Chatterjee, Steinhaeuser, Banerjee, Chatterjee, and Ganguly, 2012)

Intensity of heaviest Indian storms
(Ghosh, Das, Kao, and Ganguly, 2011)
Enabling Large-scale Analytics: An HPC Library of Data Analysis Kernels

Performance typically dominated by a few computational kernels.

<table>
<thead>
<tr>
<th>Application</th>
<th>Top 3 Kernels</th>
<th>Σ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kernel 1 (%)</td>
<td>Kernel 2 (%)</td>
</tr>
<tr>
<td>K-means</td>
<td>Distance (68)</td>
<td>Center (21)</td>
</tr>
<tr>
<td>Fuzzy K-means</td>
<td>Center (58)</td>
<td>Distance (39)</td>
</tr>
<tr>
<td>BIRCH</td>
<td>Distance (54)</td>
<td>Variance (22)</td>
</tr>
<tr>
<td>HOP</td>
<td>Density (39)</td>
<td>Search (30)</td>
</tr>
<tr>
<td>Naïve Bayesian</td>
<td>probCal (49)</td>
<td>Variance (38)</td>
</tr>
<tr>
<td>ScalParC</td>
<td>Classify (37)</td>
<td>giniCalc (36)</td>
</tr>
<tr>
<td>Apriori</td>
<td>Subset (58)</td>
<td>dataRead (14)</td>
</tr>
<tr>
<td>Eclat</td>
<td>Intersect (39)</td>
<td>addClass (23)</td>
</tr>
<tr>
<td>SVMlight</td>
<td>quotMatrix (57)</td>
<td>quadGrad (38)</td>
</tr>
</tbody>
</table>

Library of highly optimized, scalable kernels

- Flexibility to define custom analytics pipelines
- High scalability
- Integrate into a software framework (e.g. R)
- MPI, OpenMP, CUDA, Parallel I/O
Scalable & Power-aware Data Analytics
Representative Data Analytics Kernels

- Parallel hierarchical clustering
  - Speedup of 18,000 on 16k processors
  - I/O significant at large scale

Power-aware analytics
- Reduced bit fixed-point representations
- Pearson correlation
  - 2.5-3.5 times faster
  - 50-70% less energy
- K-means
  - ~44% less energy with an error of only 0.03% using 12-bit representation

Energy Consumption Correlations
Speedup Correlation
K-means: Error vs. Energy
### Data Mining and Analytics – Broader Impact

<table>
<thead>
<tr>
<th>Illustrative Applications</th>
<th>Feature, data reduction, or analytics task</th>
<th>Data analysis kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry, <strong>Climate</strong>, Combustion, Cosmology, Fusion, Materials science, Plasma</td>
<td>Clustering</td>
<td>k-means, fuzzy k-means, BIRCH, MAFIA, DBSCAN, HOP, SNN, Dynamic Time Warping, Random Walk</td>
</tr>
<tr>
<td>Biology, <strong>Climate</strong>, Combustion, Cosmology, Plasma, Renewable energy</td>
<td>Statistics</td>
<td>Extrema, mean, quantiles, standard deviation, copulas, value-based extraction, sampling</td>
</tr>
<tr>
<td>Biology, <strong>Climate</strong>, Fusion, Plasma</td>
<td>Feature selection</td>
<td>Data slicing, LVF, SFG, SBG, ABB, RELIEF</td>
</tr>
<tr>
<td>Chemistry, Materials science, Plasma, <strong>Climate</strong></td>
<td>Data transformations</td>
<td>Fourier transform, wavelet transform, PCA/SVD/EOF analysis, multidimensional scaling, differentiation, integration</td>
</tr>
<tr>
<td>Combustion, <strong>Earth science</strong></td>
<td>Topology</td>
<td>Morse-Smale complexes, Reeb graphs, level set decomposition</td>
</tr>
<tr>
<td><strong>Earth science</strong></td>
<td>Geometry</td>
<td>Fractal dimension, curvature, torsion</td>
</tr>
<tr>
<td>Biology, <strong>Climate</strong>, Cosmology, Fusion</td>
<td>Classification</td>
<td>ScalParC, decision trees, Naïve Bayes, SVMlight, RIPPER</td>
</tr>
<tr>
<td>Chemistry, <strong>Climate</strong>, Combustion, Cosmology, Fusion, Plasma</td>
<td>Data compression</td>
<td>PPM, LZW, JPEG, wavelet compression, PCA, Fixed-point representation</td>
</tr>
<tr>
<td><strong>Climate</strong></td>
<td>Anomaly detection</td>
<td>Entropy, LOF, GBAD</td>
</tr>
<tr>
<td><strong>Climate</strong>, Earth science</td>
<td>Similarity / distance</td>
<td>Cosine similarity, correlation (TAPER), mutual information, Student's t-test, Eulerian distance, Mahalanobis distance, Jaccard coefficient, Tanimoto coefficient, shortest paths</td>
</tr>
<tr>
<td>Cosmology</td>
<td>Halos and sub-halos</td>
<td>SUBFIND, AHF</td>
</tr>
</tbody>
</table>
Examples and Results
Climate System Complexity

The Complexity of Climate Systems Comes from Interconnections.

Climate systems are complex because of non-linear coupling of its subsystems (e.g., the ocean and the atmosphere).

**Challenge:**
How to “connect the dots”, that is, to construct *predictive phenomenological* models explaining *structure-dynamics-function relationships* in the complex climate system.

*From Simplicity to Complexity*

*Science* 3 September 2010: 1125.

Slide 18
What are Climate Indices?

Climate indices are defined to quantitfy climatic phenomena. Many of them are defined in terms of teleconnection patterns or dipoles.

North Atlantic Oscillation
Dipole - difference in sea level pressure between the azores and a region near Iceland

El Niño (Warm Phase)
Teleconnection pattern - above average Sea Surface Temperature across the tropical Pacific
leads to drought like conditions in the Sahel region

ENSO index family

Nino1+2  Nino3  Nino4  Nino3.4
Cold phase of the Atlantic Dipole is associated with weak increased low-level outflow from the south Atlantic ocean basin (cold SST anomalies) and, hence, positive rainfall anomalies in Sahel.
1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

Can data-driven approaches expedite such discoveries?

* North African Orographic Forcing

www.psdgraphics.com
Example Use Cases: Extreme Events Prediction

NH Tropical Cyclone (TC) Activity

Climate-Meningitis Outlook

Northern Indian | North Pacific | North Atlantic

West Sahel | Mali | Nig | East Sahel

Forecasting NA Hurricane Tracks
Modeling a Climate System as a Network

Climate Data

Correlation between two anomaly time series

Stat. significant correlations

Anomaly time series at each node

Climate Network

Edge weights: significant correlations
Nodes in the graph: grid points on the globe

Multivariate Networks

Extreme Phase

Normal Phase

Multiphase Networks

Jan 25, 2012
Subgraphs Common to Extreme Event Climate Networks

Networks for Climate Systems during Extreme Events

Networks for Climate Systems during Normal Events

Slide 24
Jan 25, 2012
Intuition: If an extreme event (e.g. hurricane track) is in one of its key phases (e.g. land-hitting), then there exist network motifs (recurrent patterns in climate networks) that are specific to that phase.
Robust & Accurate Seasonal Hurricane Forecasts through Comparative Climate Networks Analytics

Comparative analysis of climate networks leverages the DOE-funded network theory & scalable algorithms.

Expedition’s novel data-driven methods already promise to excel beyond the traditional methods in climate prediction tools
(Fred Semazzi, Nobel Prize team member)

Forecast Years

Hindcast Years

35 Training Years
Forecasting Hurricane Tracks

Improving but have mean error (>185km) beyond 48 h

Physics-based Models

What if the error gets interpolated to 10-15 day in advance forecast?

~500 km

HURDAT Historic Data

Hurricane End-game Track Forecast

Forecast **10-15 days in advance** the end-game of a North Atlantic since hurricane embryonic formation in Western Africa.

- Nearly **east-oriented SLP** edges suggest horizontal pressure gradient configuration in the same direction.
- Based on Buys Ballot’s law, this pressure gradient would be associated with **wind flow in the north-south direction**.
- Onshore wind anomaly flow would promote favorable conditions for landfall; opposite flow anomaly would be more favorable for hurricanes tracks in no-landfall.

SLP (yellow/dashed) and SST (red/solid) (+)correlated teleconnections;
L—biased toward land-hitting tracks;
O—biased toward offshore tracks.

<table>
<thead>
<tr>
<th>Performance of Land-hitting vs. Offshore</th>
<th>LOO</th>
<th>10-FOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLP</td>
<td>SLP</td>
<td>SST</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>F1-meas.</strong></td>
<td>0.90</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Hierarchical modularity is a known principle of complex system’s organization & function. These functionally associated modules often combine in a hierarchical manner into larger, functionally less cohesive subsystems.

**Divide Step:**

Divide all system features into modules that likely function together to define what state the system is in: modules with **stronger associations within the modules** than between them.

**Conquer Step:**

Conquers each of these modules in order to refine the **specificity of the inter-feature relationships within the module.**
Cross-talk between Regional & Global Systems

There is an inherent interplay (e.g., feedback) between regional scale subsystems and the global scale system. Ignoring these relationships by focusing on a specific region is a simplification.

We could use these relationships for detecting the prediction errors and/or possibly correcting them.
92% Accuracy w/ Leave One Out Cross Validation

Seasonal Hurricane Activity

LOOCV Accuracy for 100 Class Label Randomizations

Hurricane Counts by Classes
Hurricane Activity Class Forecast vs. State-of-art

FORECASTER Performance on North Atlantic Hurricane

<table>
<thead>
<tr>
<th>Metric</th>
<th>FORECASTER NC State</th>
<th>[1], 2009 Colorado</th>
<th>[2], 2010 GA Tech</th>
<th>Random Forest</th>
<th>Bagging</th>
<th>Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>93.3</td>
<td>64.0</td>
<td>65.5</td>
<td>76.7</td>
<td>73.3</td>
<td>75.0</td>
</tr>
<tr>
<td>HSS</td>
<td>0.90</td>
<td>0.45</td>
<td>0.49</td>
<td>0.66</td>
<td>0.60</td>
<td>0.62</td>
</tr>
<tr>
<td>PSS</td>
<td>0.92</td>
<td>0.44</td>
<td>0.50</td>
<td>0.65</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>GSS</td>
<td>0.96</td>
<td>0.50</td>
<td>0.68</td>
<td>0.65</td>
<td>0.67</td>
<td>0.66</td>
</tr>
</tbody>
</table>

ML-based Regression Hybrid


HSS: Heidke score, measures how well relative to a randomly selected forecast;
PSS: Peirce score, difference between the hit rate and the false alarm rate;
GSS: Gerrity score, occurrences substantially less frequent.
Forecasts

Model Ensemble Predictions

- Hurr. Count: 7
- Low [0;4]
- Normal [5;7]
- High [8;15]

Accuracy for 1-year Forecast

- Hit: 21%
- Normal: 34%
- High: 45%

Accuracy for 5-year Forecasts

Accuracy for 10-year Forecasts

- Miss: 1 out 5 yrs
- Accuracy: 0.8
Effectiveness of **DETECTOR + FORECASTER**
Regional subsystems and global system interplays

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>FORECASTER</th>
<th>DETECTOR + FORECASTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>STCP</td>
<td>NH</td>
<td>90.0</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>NA1</td>
<td>88.3</td>
<td>93.3</td>
</tr>
<tr>
<td>SHP</td>
<td>NA2</td>
<td>93.3</td>
<td>98.6</td>
</tr>
<tr>
<td></td>
<td>LNA</td>
<td>86.7</td>
<td>93.4</td>
</tr>
<tr>
<td>NARP</td>
<td>SH</td>
<td>88.9</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>WS</td>
<td>90.7</td>
<td>96.3</td>
</tr>
</tbody>
</table>

**Tropical cyclone activity (STCP):**
- NH: Northern Hemisphere
- NA1: North Atlantic

**Hurricane activity (SHP):**
- NA2: North Atlantic hurricane
- LNA: North Atlantic land-falling

North Africa **rainfall activity (NARP):**
- SH: Sahel area
- WS: West Sahel.

![Map of West Sahel and East Sahel](image-url)

![Map of global weather systems](image-url)
Predicted Network Motifs Agree with Climate Indices Related to Hurricane Activity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spatial location</th>
<th>Climate indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>(4N, 114W)</td>
<td>Nino 3</td>
</tr>
<tr>
<td></td>
<td>(2S, 168W)</td>
<td>ENSO</td>
</tr>
<tr>
<td></td>
<td>(42N, 30W)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(32S, 16W)</td>
<td></td>
</tr>
<tr>
<td>VWS</td>
<td>(27.5N, 65W)</td>
<td>MDR</td>
</tr>
<tr>
<td></td>
<td>(52.5N, 37.5W)</td>
<td>NAO</td>
</tr>
<tr>
<td></td>
<td>(7.5N, 122.5W)</td>
<td>Nino 3</td>
</tr>
<tr>
<td></td>
<td>(10S, 60W)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(27.5N, 55W)</td>
<td></td>
</tr>
<tr>
<td>PW</td>
<td>(52.5N, 135E)</td>
<td>PDO</td>
</tr>
<tr>
<td></td>
<td>(82.5N, 15W)</td>
<td>AO</td>
</tr>
<tr>
<td></td>
<td>(37.5N, 40E)</td>
<td></td>
</tr>
<tr>
<td>SLP</td>
<td>(57.5N, 22.5W)</td>
<td>NAO</td>
</tr>
<tr>
<td></td>
<td>(60N, 155E)</td>
<td>PDO</td>
</tr>
<tr>
<td></td>
<td>(37.5N, 162.5W)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(12.5N, 122.5E)</td>
<td></td>
</tr>
</tbody>
</table>

Published Facts
- Nino3 SSTs correlate with Atlantic hurricane activity
- ENSO modulates NA TCs
- SSTs in MDR contribute to hurricanes in MDR region
- NAO June correlates with NA hurricane tracks
- Shifts in the PDO phase can have significant implications for Atlantic hurricane activity

New Hypotheses
Atlantic multi-decadal Oscillation (AMO) and Arctic Oscillation (AO) indices might affect the North Atlantic tropical cyclone activities
0.67 Spearman Rank-order Correlation between Network-based Climate Index & Hurricane Activity

Comparison against 33 known climate indices
Best absolute correlation for January-June
Not all p-values are significant
Hypothesis: NAO modulates the climate drivers of the West African climate—the Atlantic Dipole & Atlantic ENSO—via the low-level westerly jet.

1986-2009 climate research on key factors affecting the West African Climate are being advanced by data-driven phenomenological modeling.

Data-driven inference of active phase causality for the NAO-driven hypothesis

Expedition’s novel data-driven approaches already promise to search for fundamental inter-relationships in the climate system in a significant way (Fred Semazzi, Nobel Prize team member)
Summary: Discovering Knowledge from Massive Data – Next Frontier for HPC

Data management, High-End Analytics, Data Mining, and Network Mining

Business

Engineering

Science