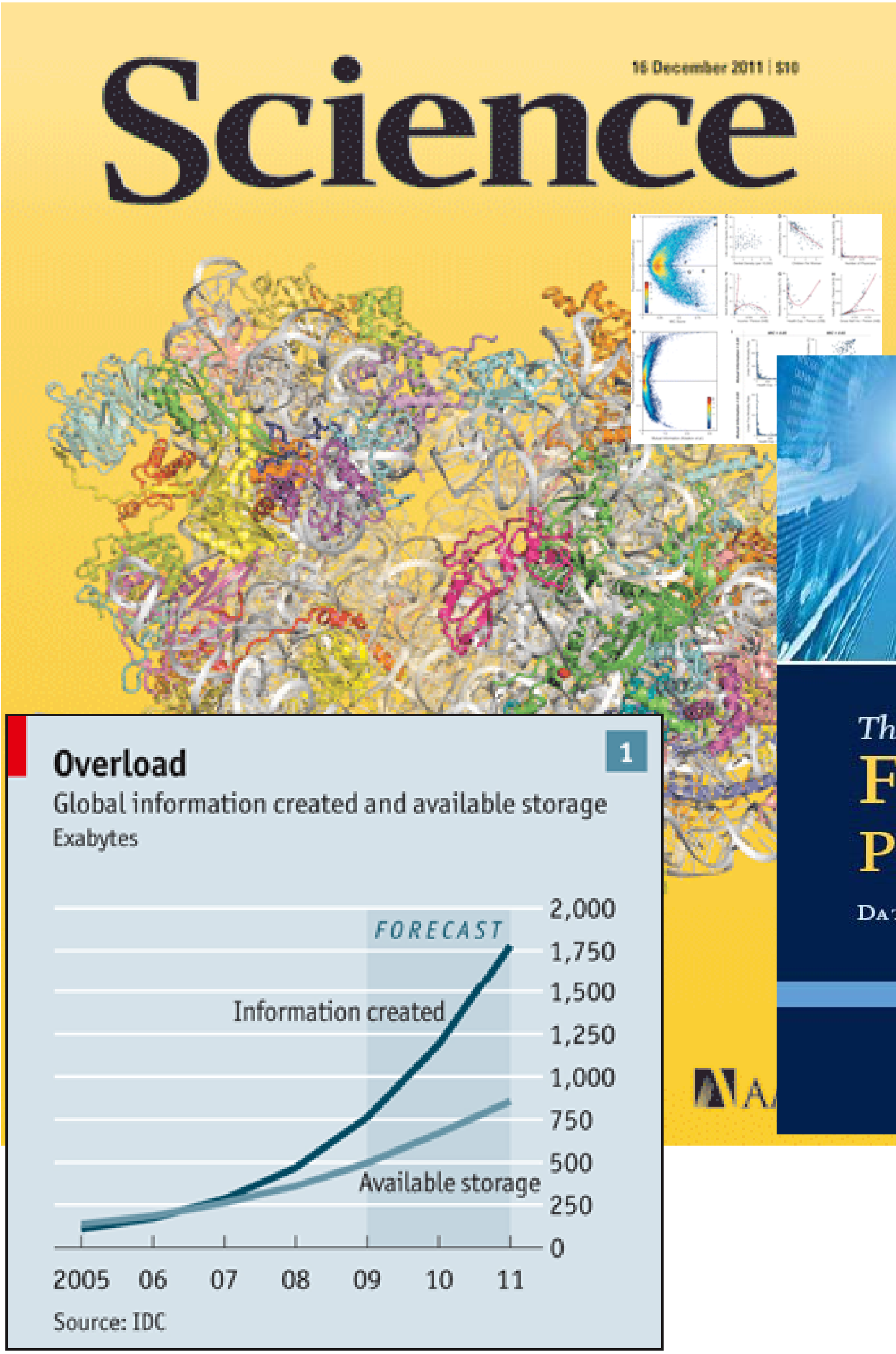
# Understanding Climate Change:
# *A Data-Driven Approach*

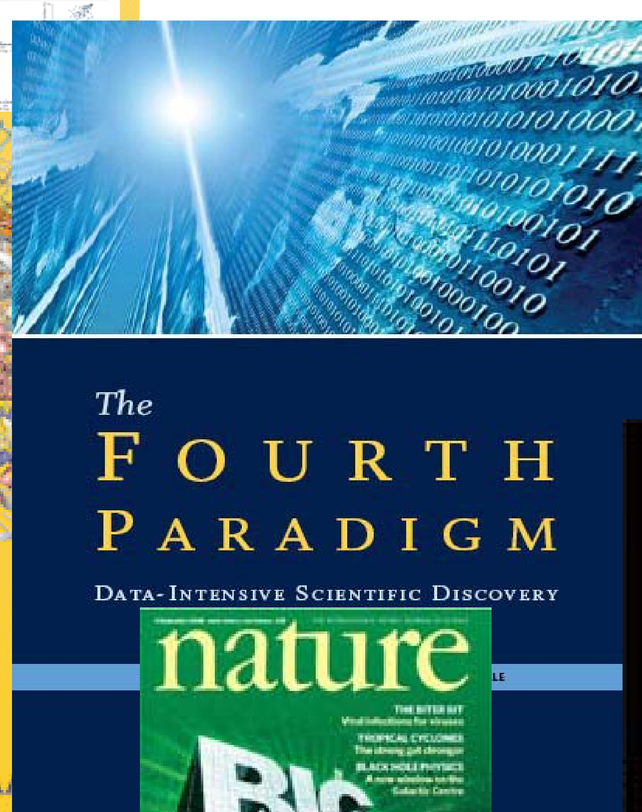

Alok Choudhary, Northwestern University

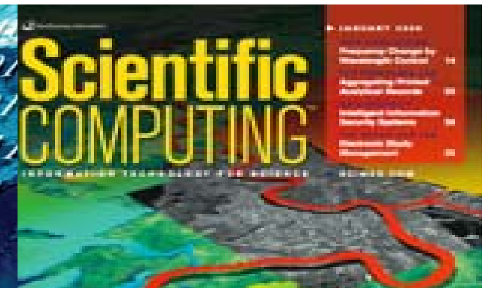
Nagiza F. Samatova, NC State and ORNL

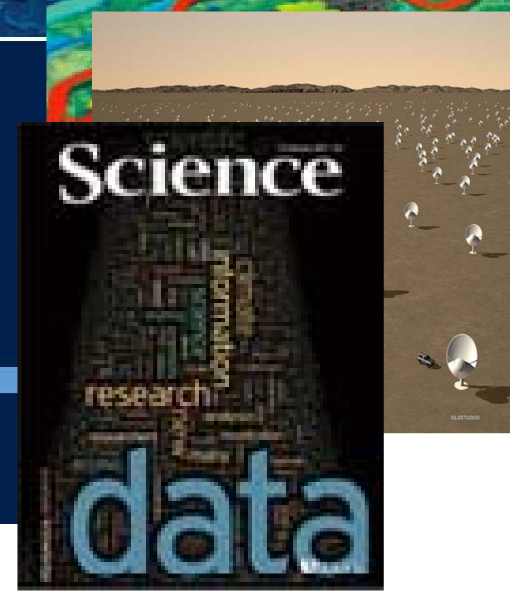
choudhar@eecs.northwestern.edu
samatova@cs.ncsu.edu

# Science and Society Transformed by Data



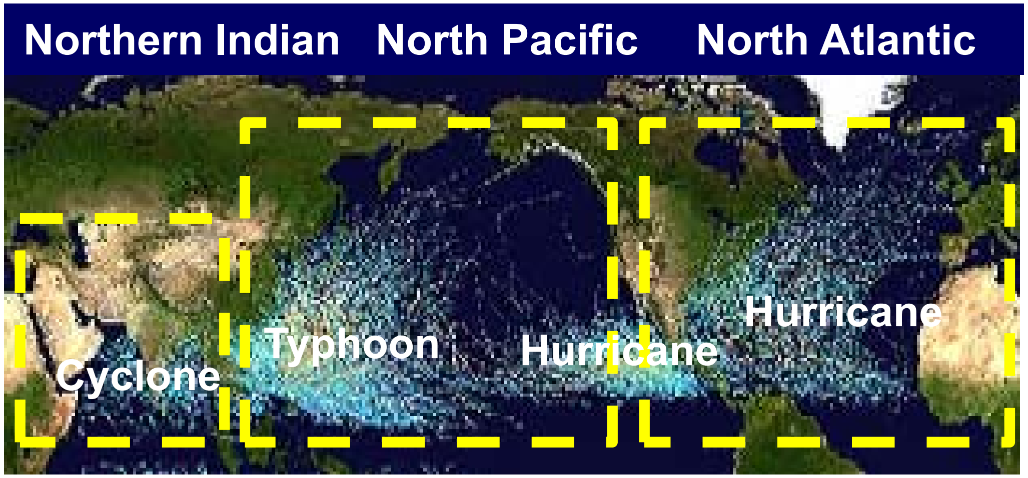



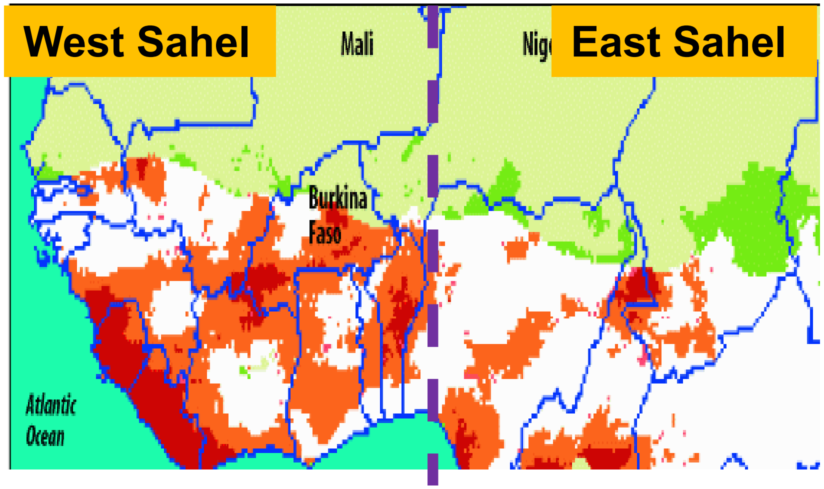




# Example Use Cases: Extreme Events Prediction

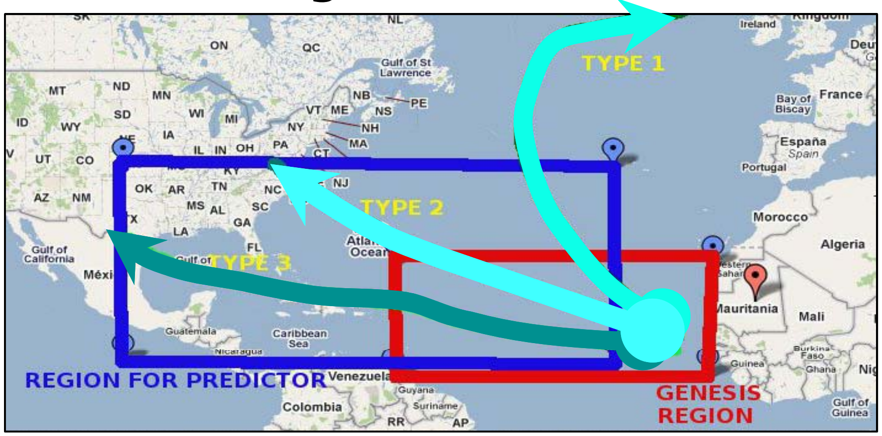
## NH Tropical Cyclone (TC) Activity




## Climate-Meningitis Outlook




## Forecasting NA Hurricane Tracks

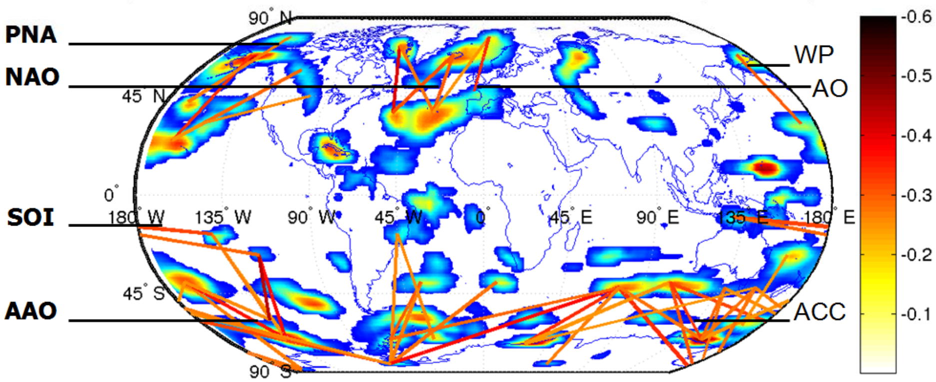


## Climate System Complexity

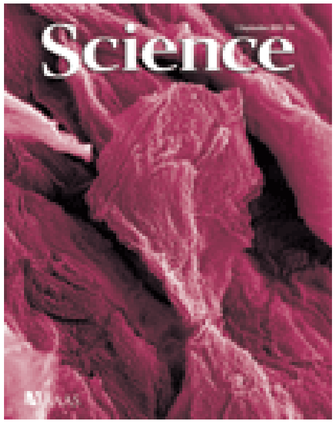
### The Complexity of Climate Systems Comes from Interconnections.

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

NCEP (National Centers for Environmental Prediction) Reanalysis Data

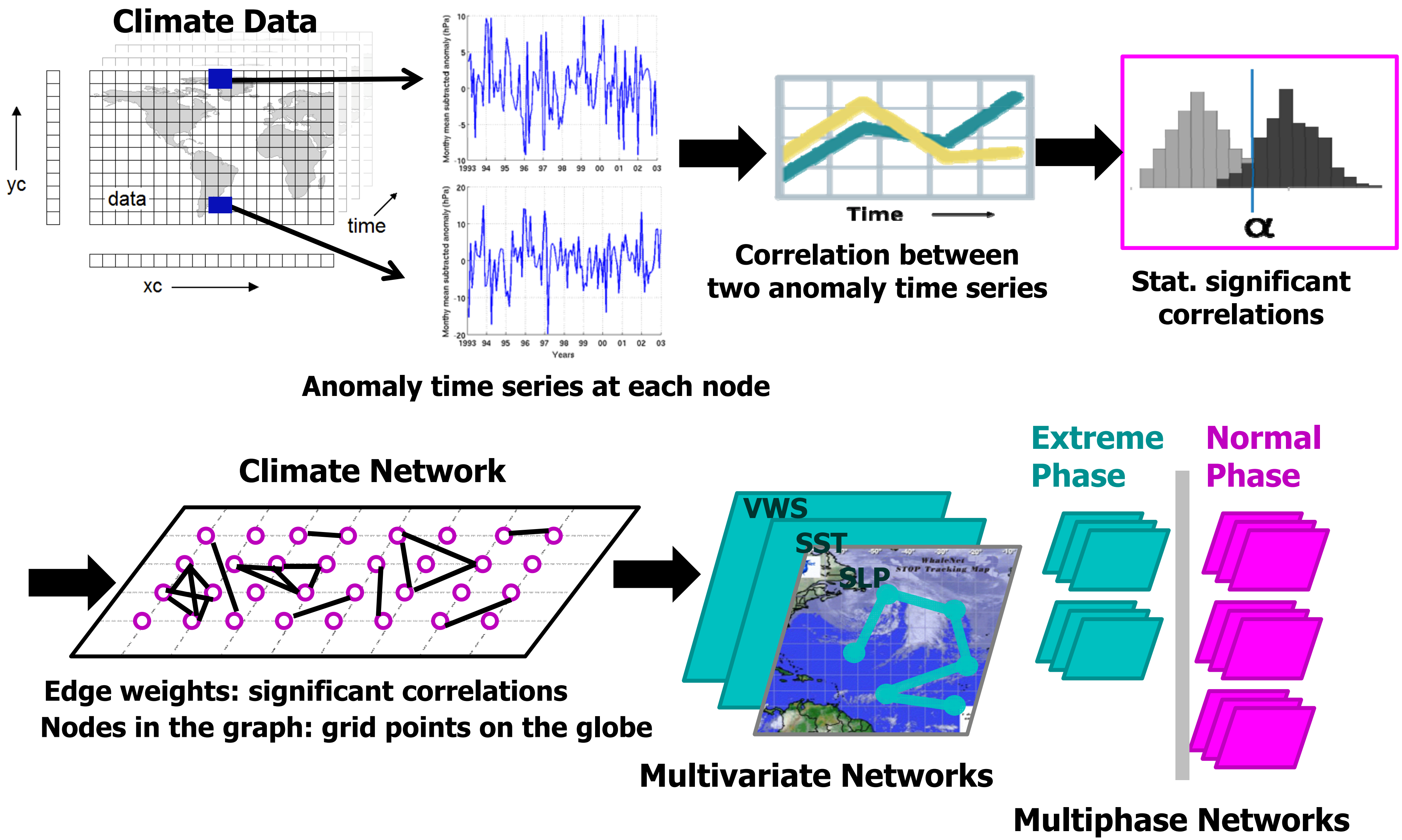



<u>**Challenge:**</u>
How to "connect the dots", that is, to construct *predictive phenomenological m*odels explaining **structure-dynamics-function relationships** in the complex climate system.



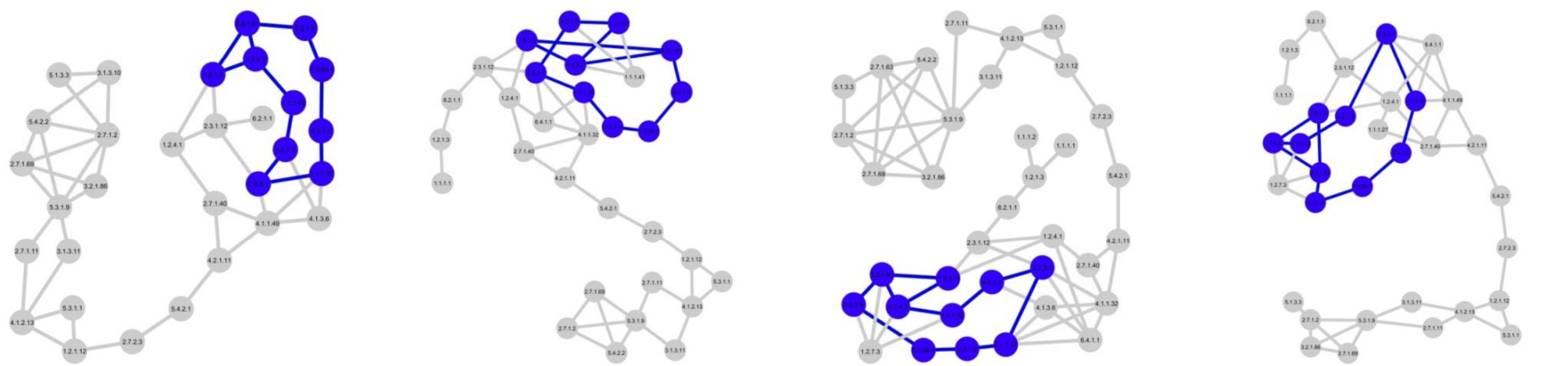

From Simplicity to Complexity
*Science 3 September 2010: 1125.*

# Modeling a Climate System as a Network

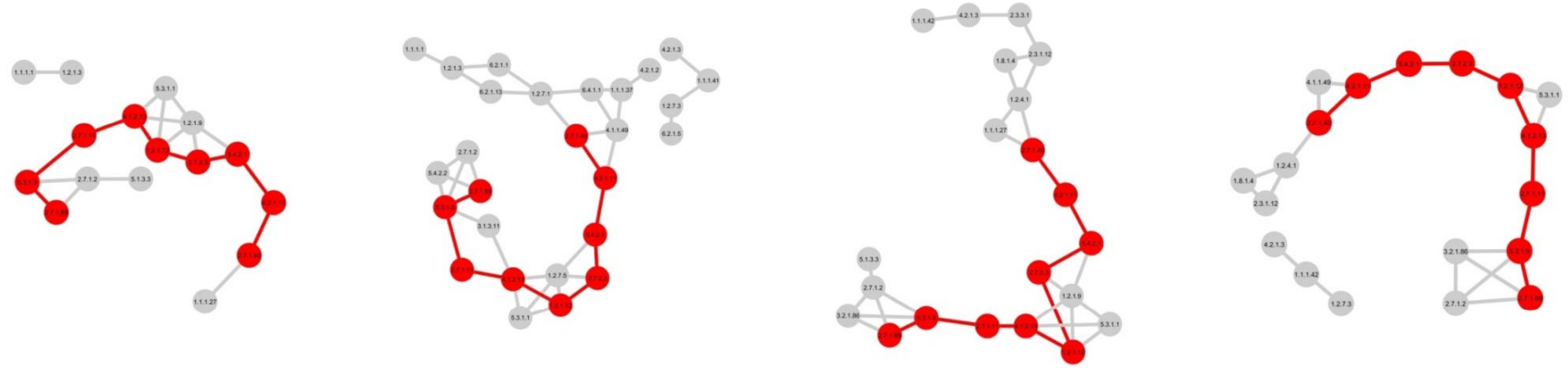


# Subgraphs Common to Extreme Event Climate Networks

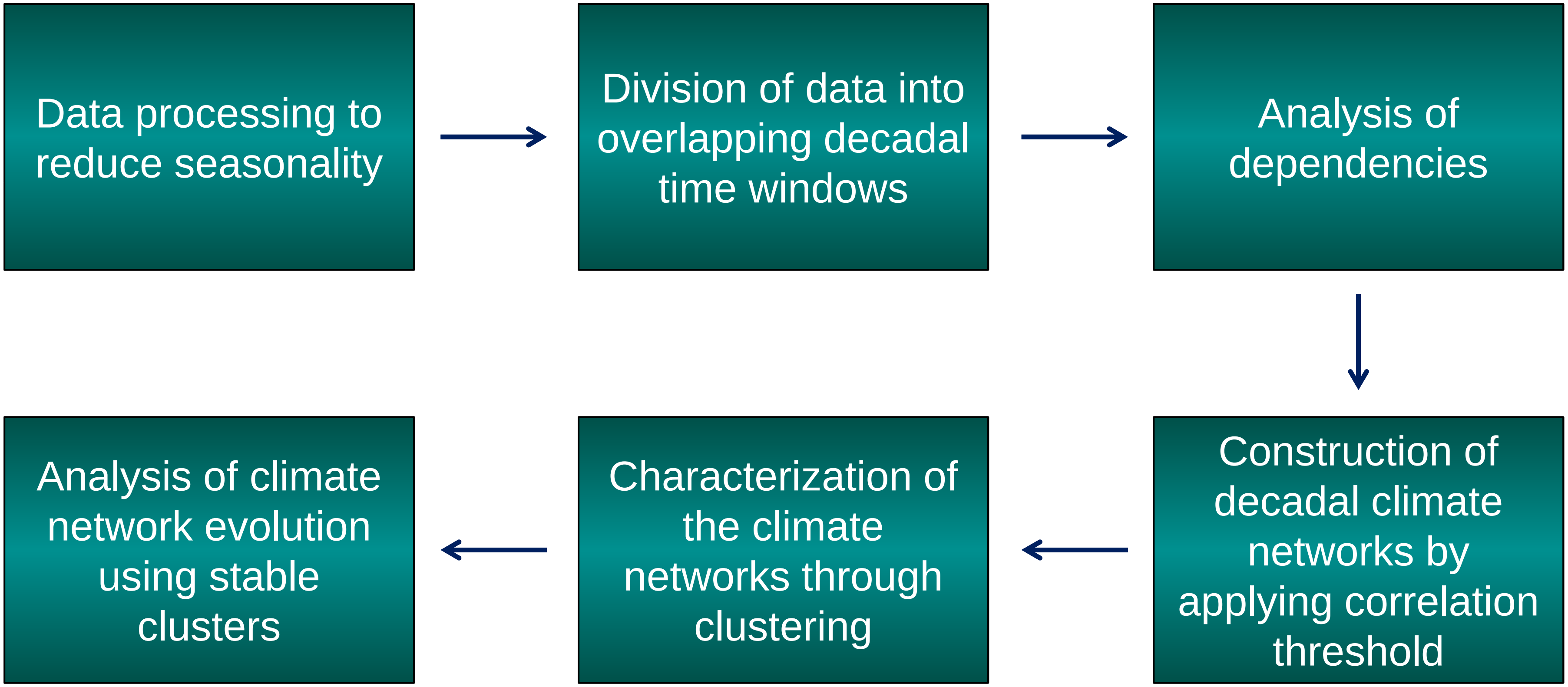
## Networks for Climate Systems during Extreme Events



## Networks for Climate Systems during Normal Events

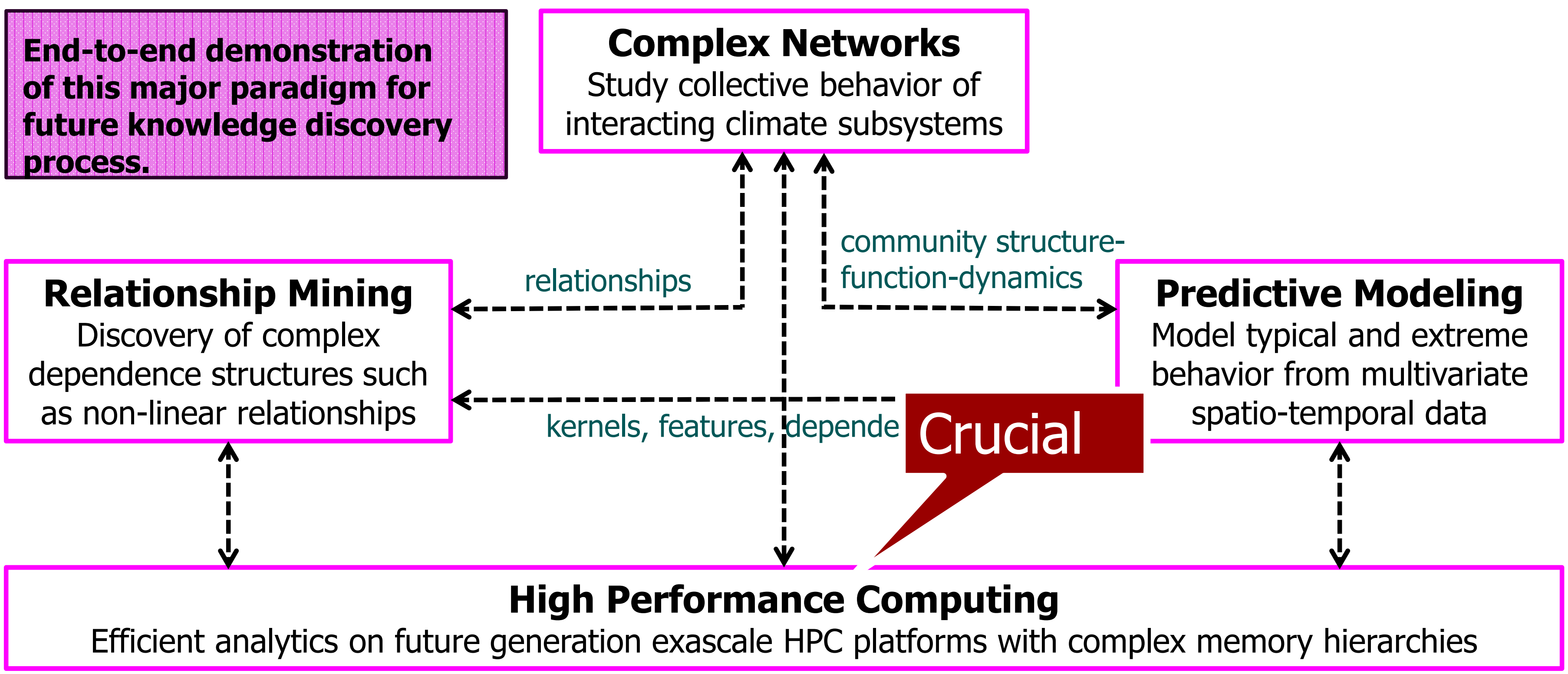


# Identifying patterns in the evolution of the climate system – Example : Analysis of Decadal Trends in Climate

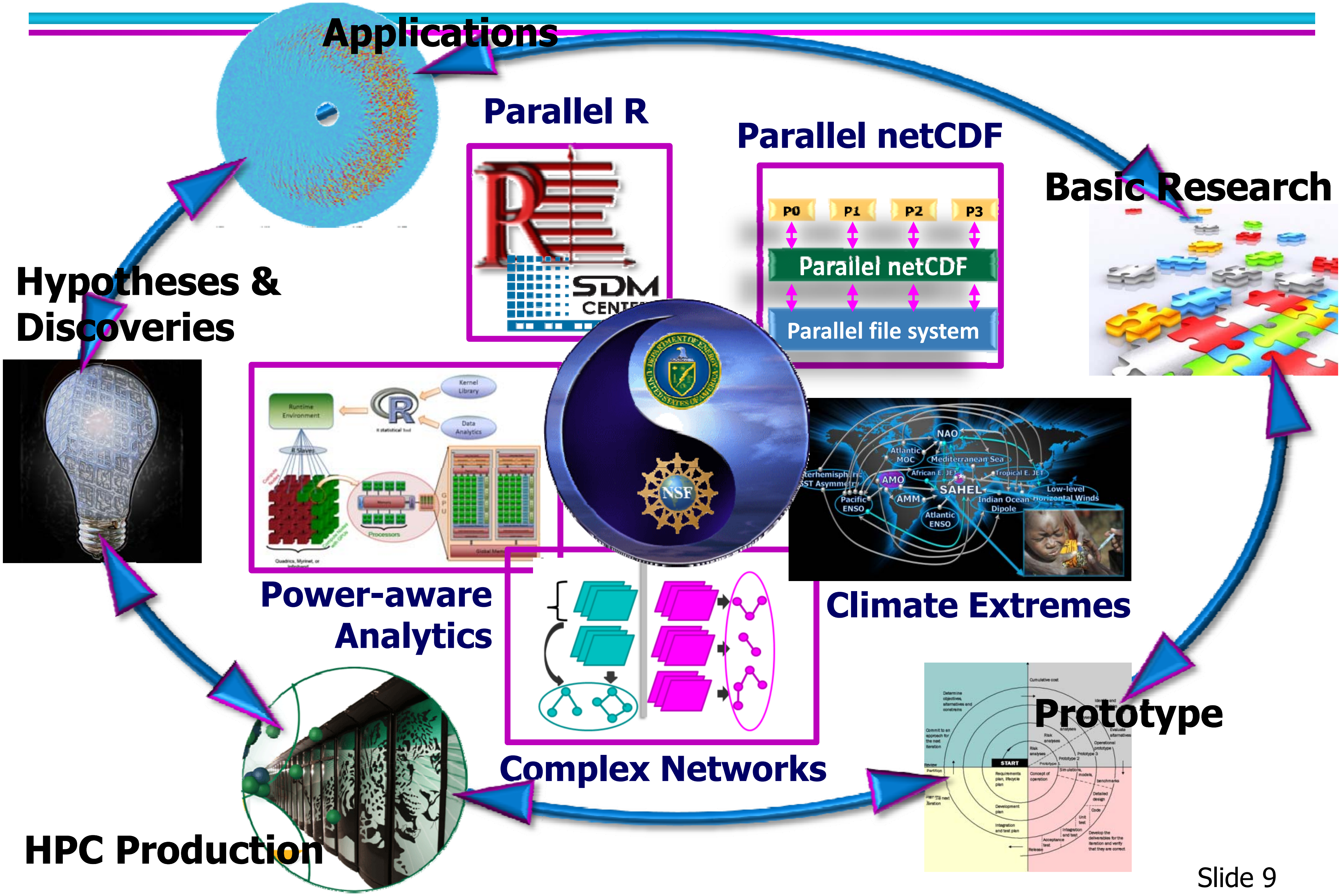


# Enabling Transformative Computer Science Research

**Enabling large-scale data-driven science for complex, multivariate, spatio-temporal, non-linear, and dynamic systems:**

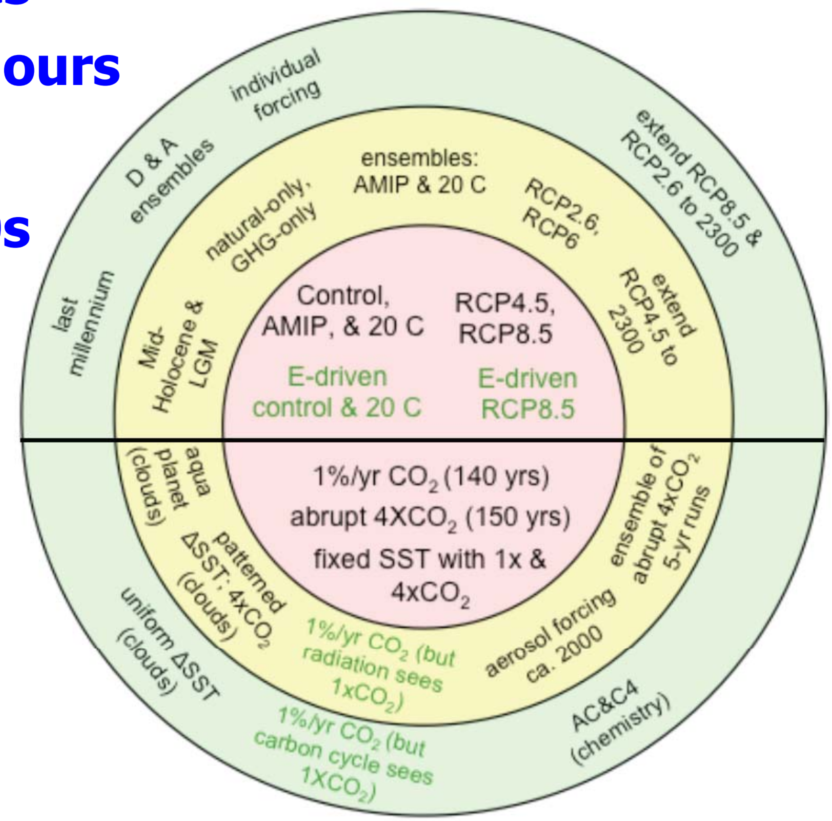


# A Complementary Interplay of R&D Portfolios



# Illustrative Case for HPC: CMIP3 → CMIP5

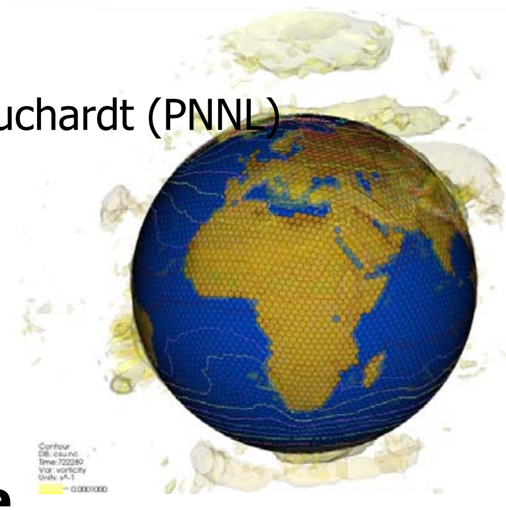
- **Coupled Model Inter comparison Project**
- **Spatial resolution: 1 – 0.25 degrees**
- **Temporal resolution: 6 hours – 3 hours**
- **Models: 24 - 37**
- **Simulation experiments: 10s - 100s**
  - Control runs & hindcast
  - Decadal & centennial-scale forecasts
- **Covers 1000s of simulation years**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



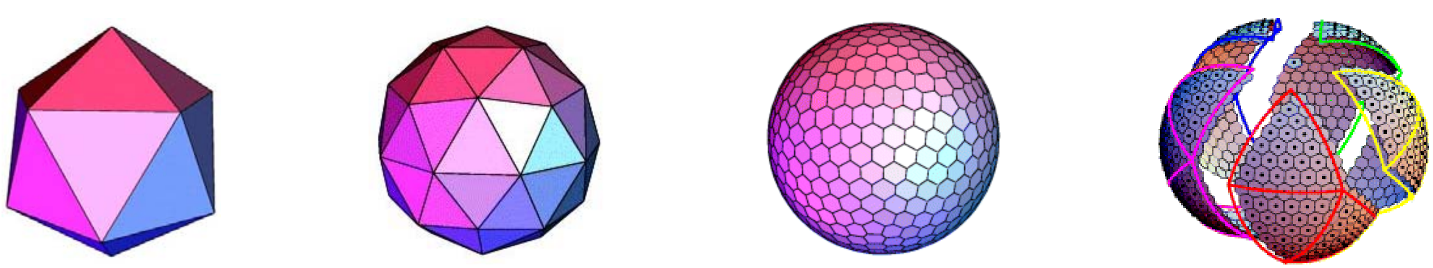

**Summary of CMIP5 model experiments, grouped into three tiers**

# 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**

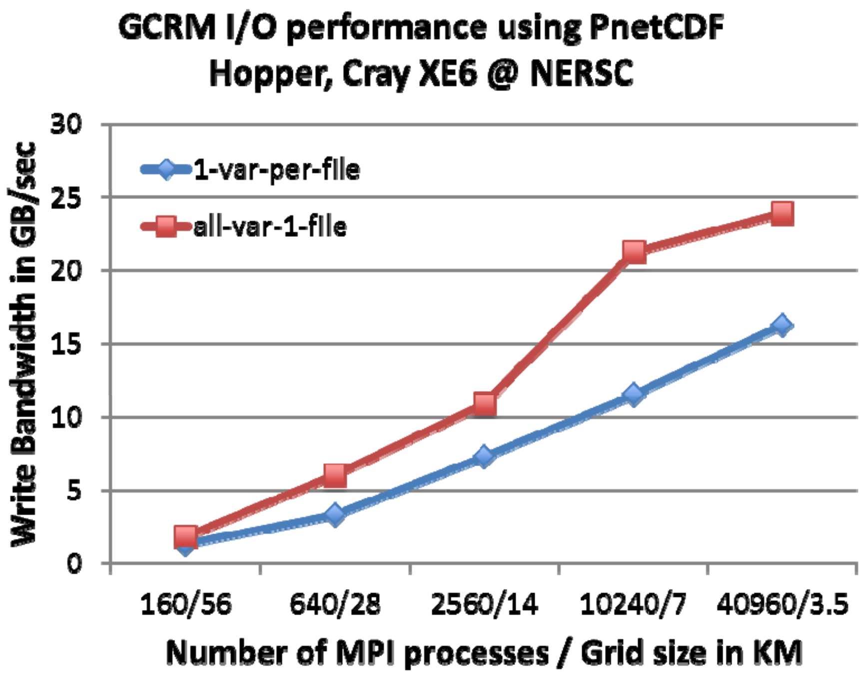


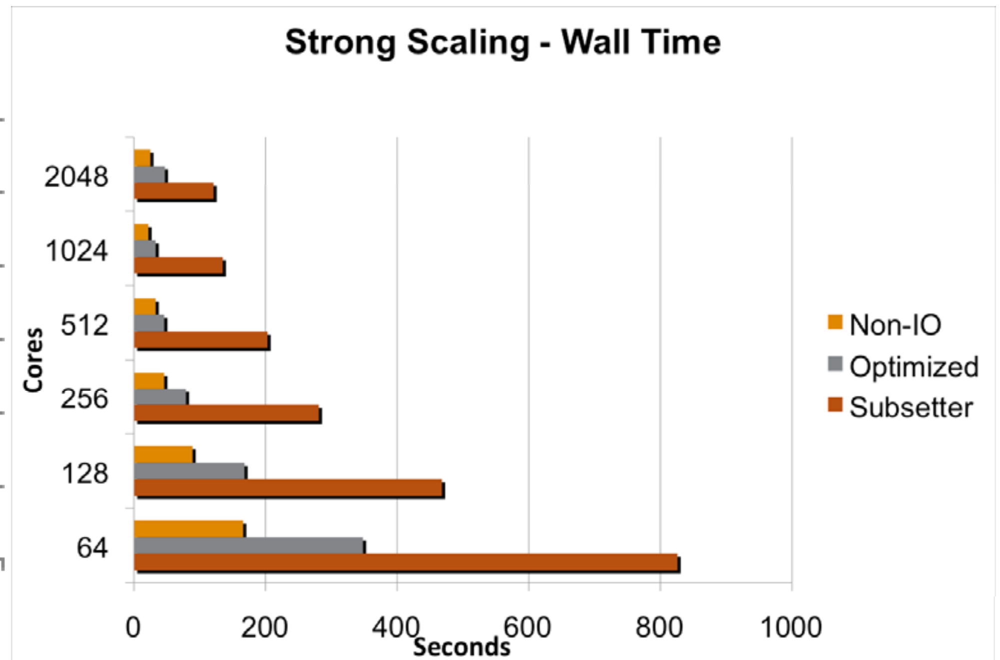
**I/O was previously a major bottleneck: The only reason execution at this scale became possible was due to I/O scaling.**



# Illustrative Results

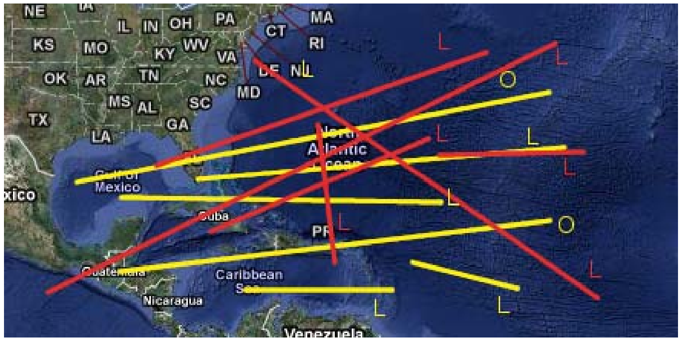
- **Improved I/O throughput**
  - Using PnetCDF optimizations, massive scalability
  - For 3.5 km grid resolution, grid size is 41.9M cells with 256 vertical layers
  - Data analysis read and simulation checkpoint



# 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

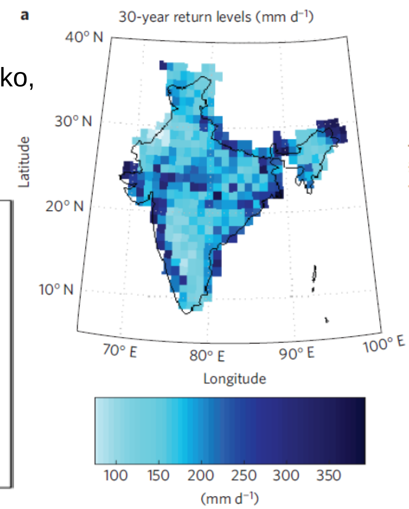



Significant correlations for hurricane prediction

(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)

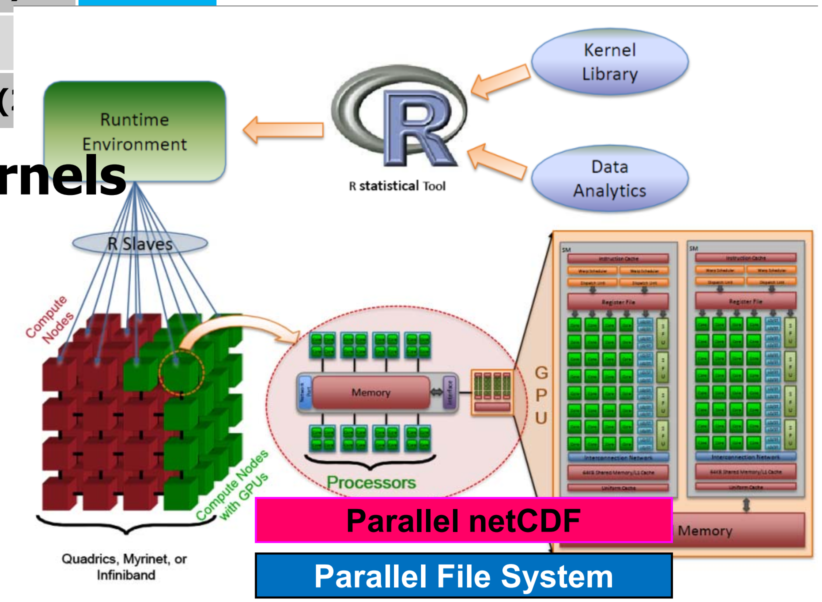
# Enabling Large-scale Analytics: An HPC Library of Data Analysis Kernels

## Performance typically dominated by a few computational *kernels.*

| Application | Top 3 Kernels | | | Σ (%) |
|---|---|---|---|---|
| | Kernel 1 (%) | Kernel 2 (%) | Kernel 3 (%) | |
| K-means | Distance (68) | Center (21) | minDist (10) | 99 |
| Fuzzy K-means | Center (58) | Distance (39) | fuzzySum (1) | 98 |
| BIRCH | Distance (54) | Variance (22) | Redist (10) | 86 |
| HOP | Density (39) | Search (30) | Gather (23) | 92 |
| Naïve Bayesian | probCal (49) | Variance (38) | dataRead (10) | 97 |
| ScalParC | Classify (37) | giniCalc (36) | Compare (24) | 97 |
| Apriori | Subset (58) | dataRead (14) | Increment (8) | 80 |
| Eclat | Intersect (39) | addClass (23) | invertC (10) | |
| SVMlight | quotMatrix (57) | quadGrad (38) | quotUpdate ( | |

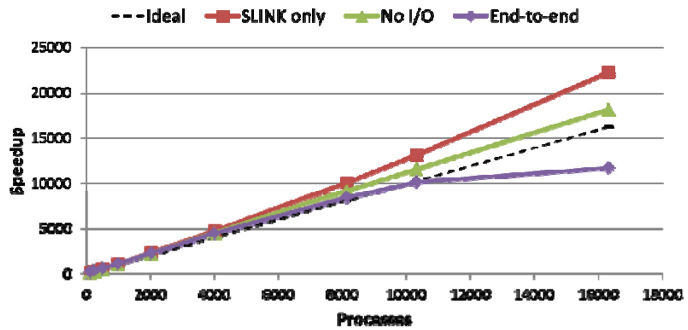
## Library of highly optimized, scalable kernels

- 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

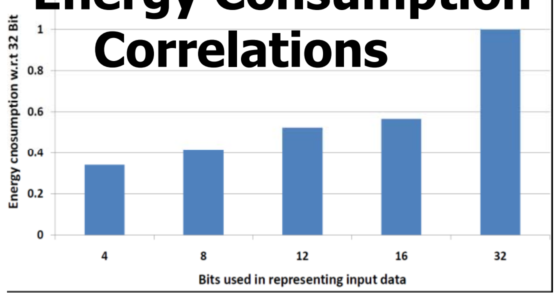
- **Parallel hierarchical clustering**
  - Speedup of 18,000 on 16k processors
  - I/O significant at large scale



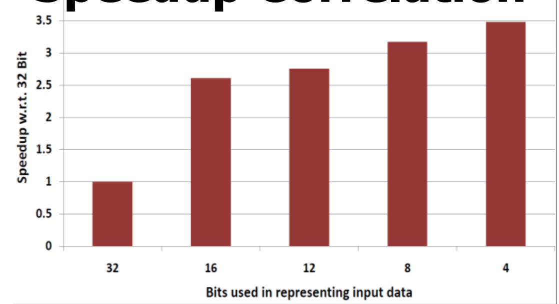

## Power-aware analytics

- **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

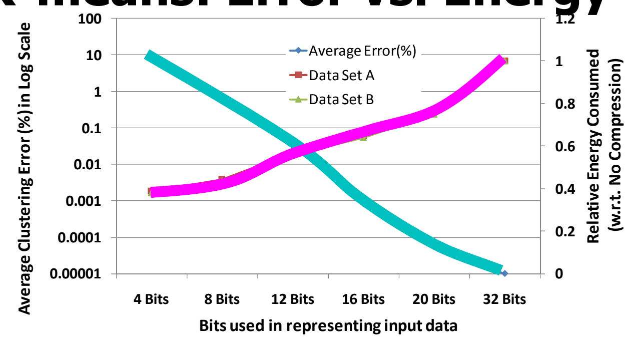
**Energy Consumption Correlations**




**Speedup Correlation**




**K-means: Error vs. Energy**



# Data Mining and Analytics – Broader Impact

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

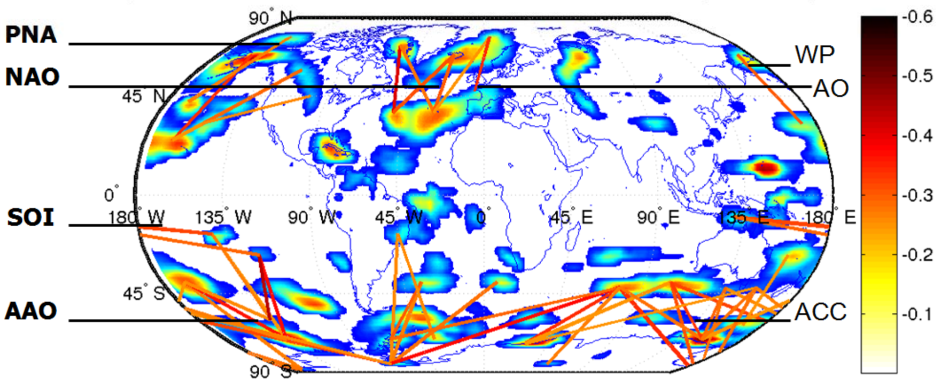
# Examples and Results

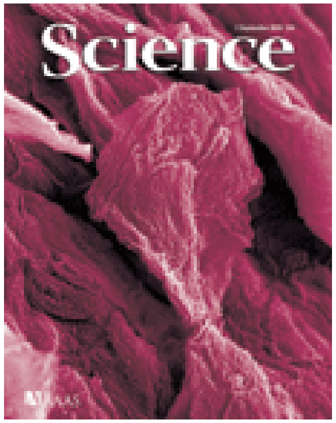
# Climate System Complexity

## 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).**

NCEP (National Centers for Environmental Prediction) Reanalysis Data




**Challenge:**

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

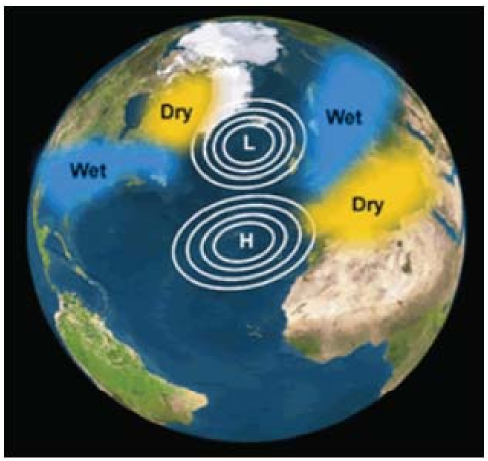
From Simplicity to Complexity

*Science 3 September 2010: 1125.*

# What are Climate Indices?

## Climate indices are defined to quantify climatic phenomena

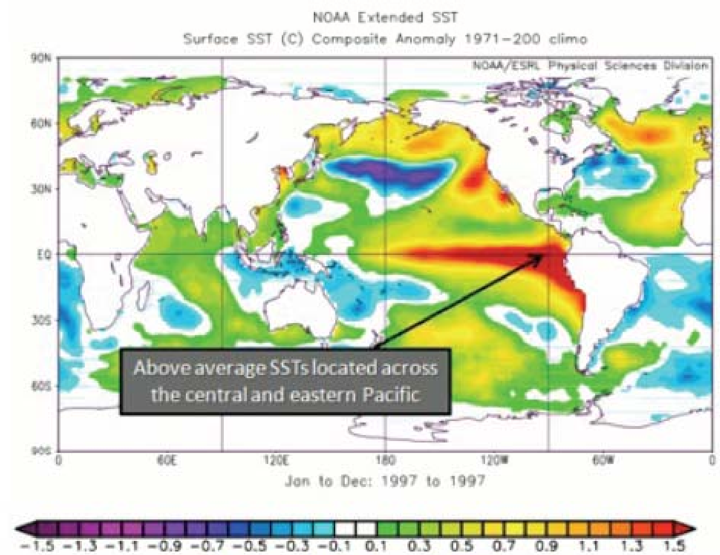
Many of them are defined in terms of teleconnection patterns or dipoles




North Atlantic Oscillation

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

NAO




El Niño (Warm Phase)

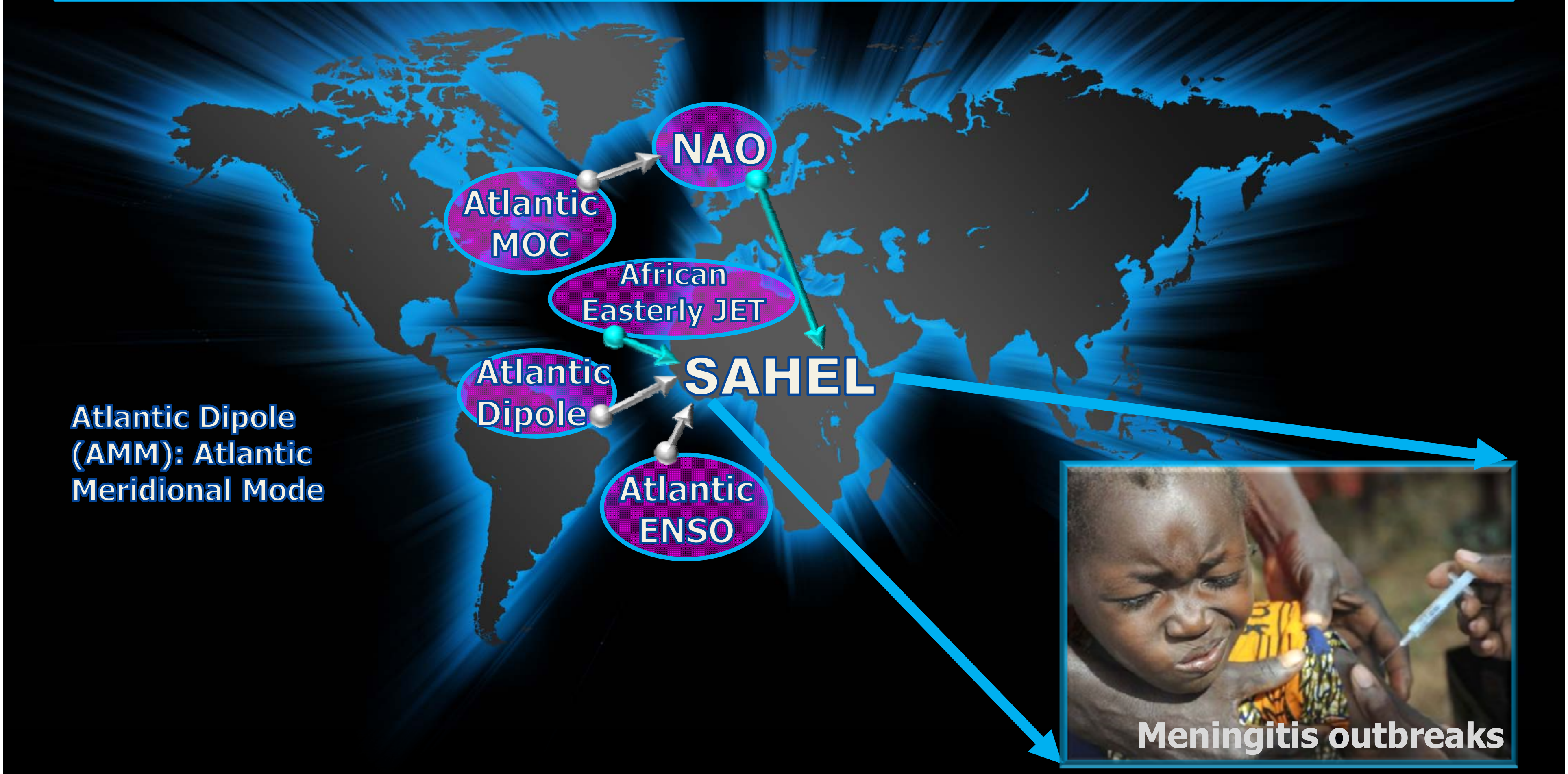
Teleconnection pattern - above average Sea Surface Temperature across the tropical Pacific

leads to drought like conditions in the Sahel region

Nino1+2 | Nino3 | Nino4 | Nino3.4

ENSO index family

**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.**

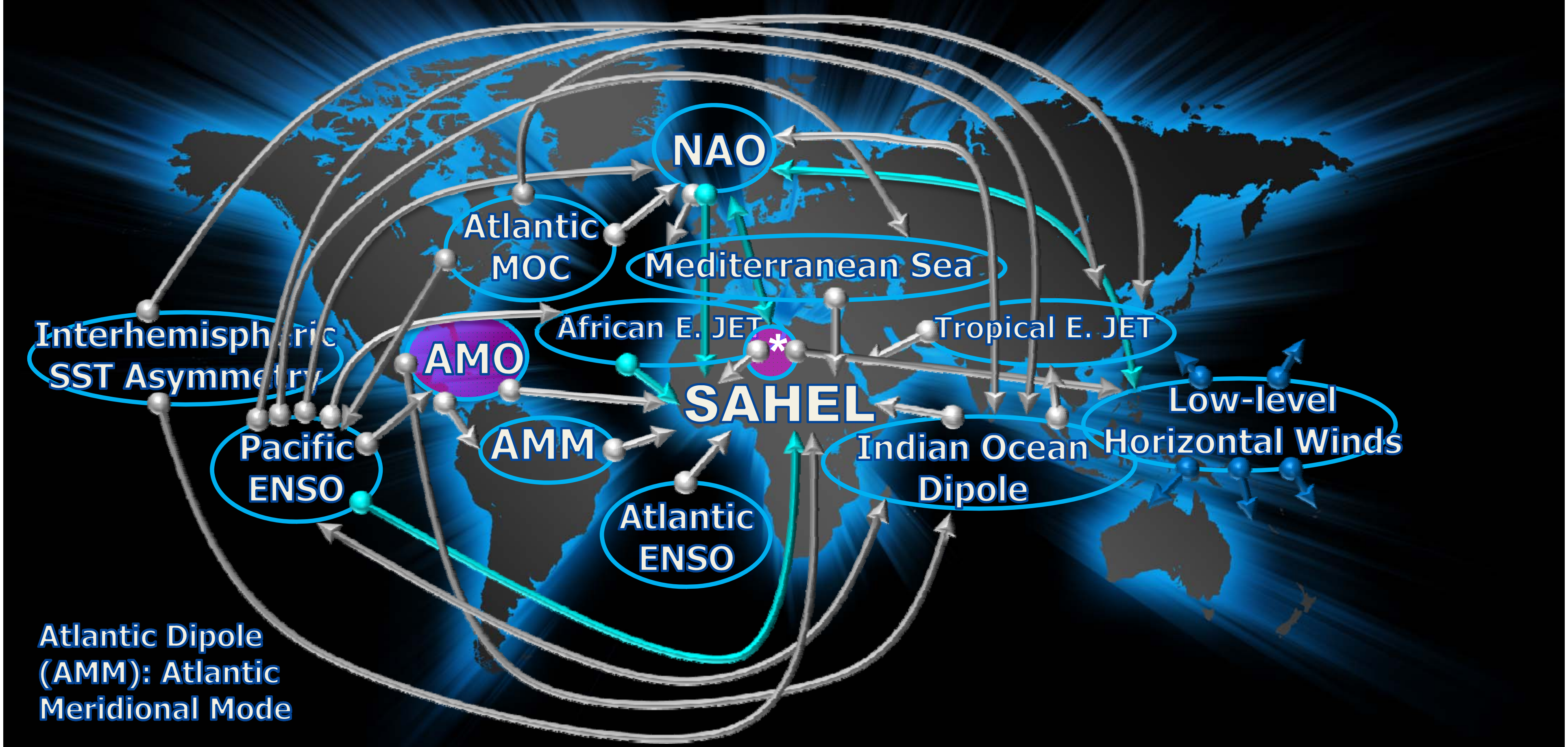



Direct/indirect causality; Documented mechanisms w/ confidence

Mechanisms not fully understood

# 1986-2009 Studies to Understand Key Climate Drivers & Dynamic Factors/Mechanisms Affecting the West African Climate.

## Can data-driven approaches expedite such discoveries?




Direct/indirect causality; Documented mechanisms w/ confidence

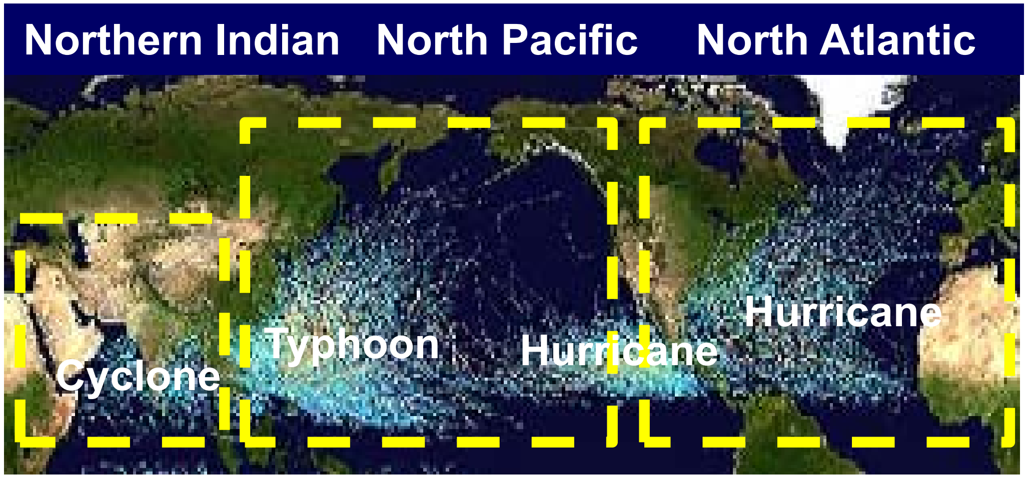
Mechanisms not fully understood

Hadley & Walker circulations

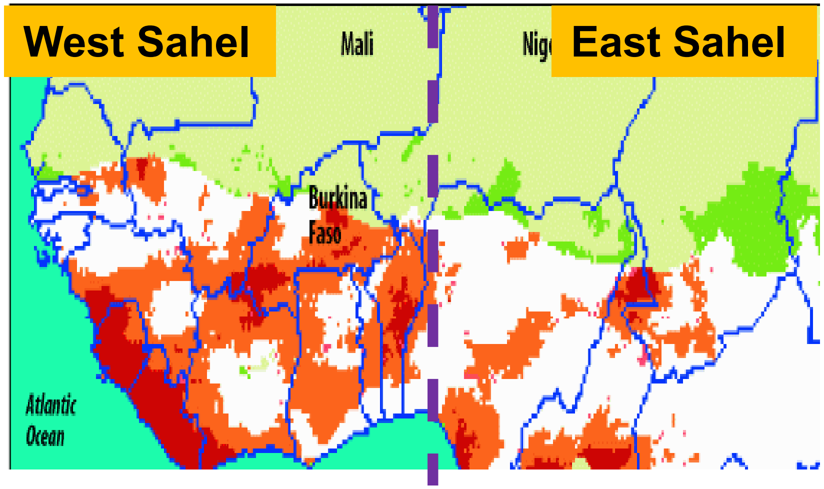
* North African Orographic Forcing

# Example Use Cases: Extreme Events Prediction

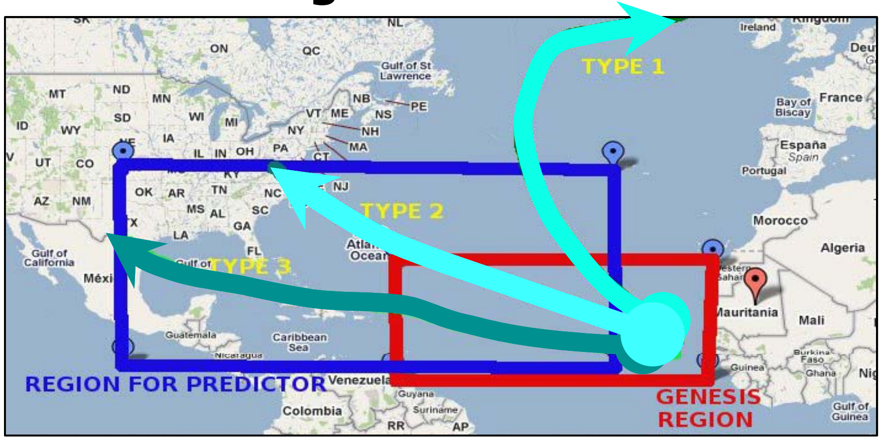
## NH Tropical Cyclone (TC) Activity



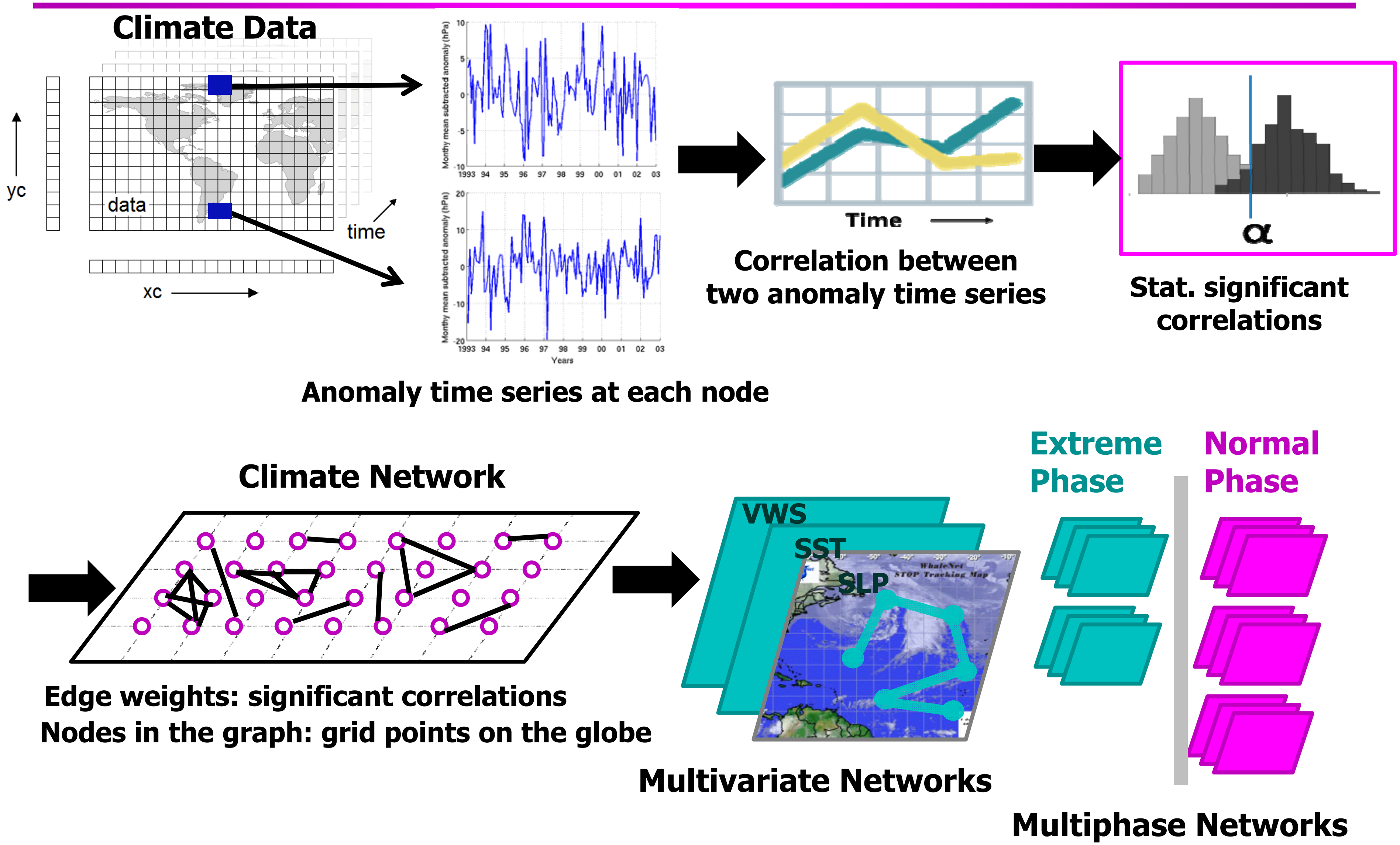

## Climate-Meningitis Outlook




## Forecasting NA Hurricane Tracks

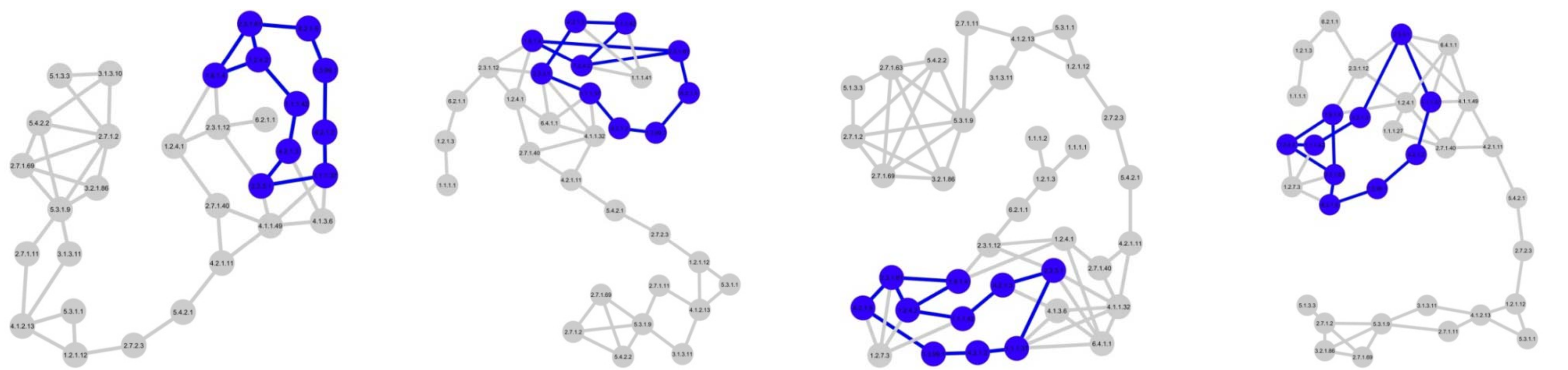


# Modeling a Climate System as a Network

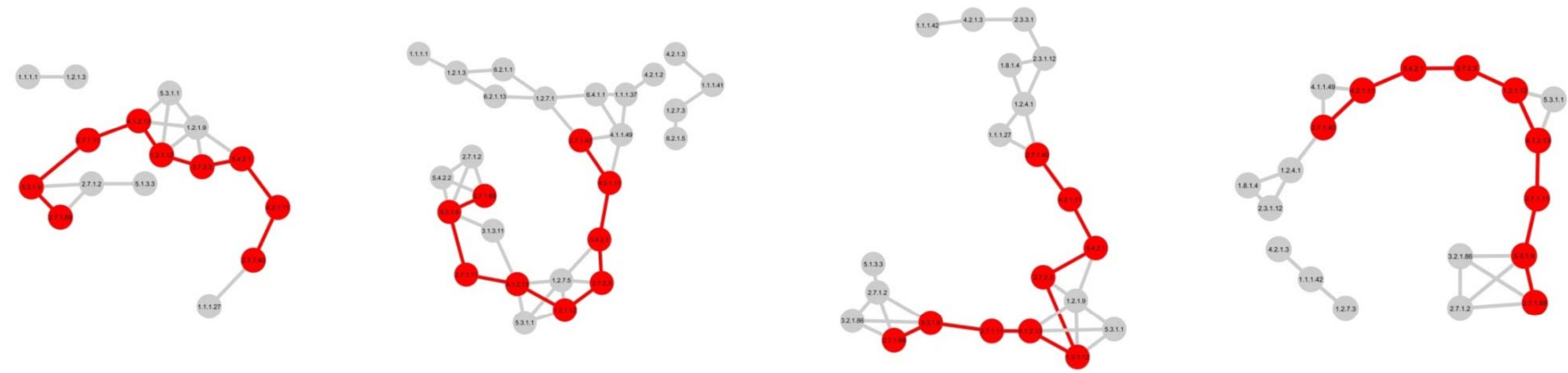


# Subgraphs Common to Extreme Event Climate Networks

## Networks for Climate Systems during Extreme Events

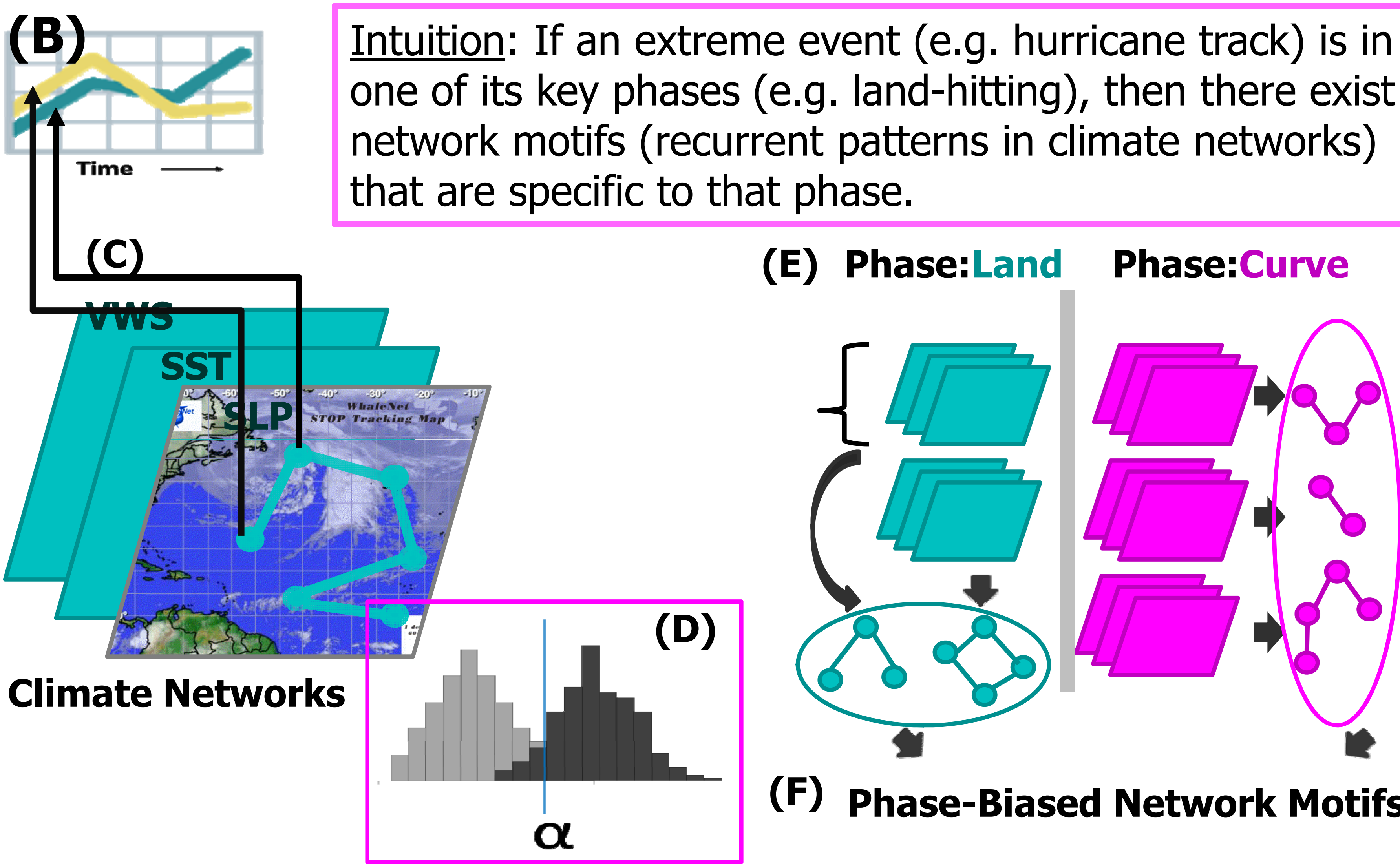


## Networks for Climate Systems during Normal Events

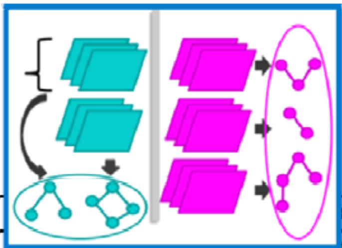


# Extreme Event Forecasting via Contrast-based Network Motif Discovery

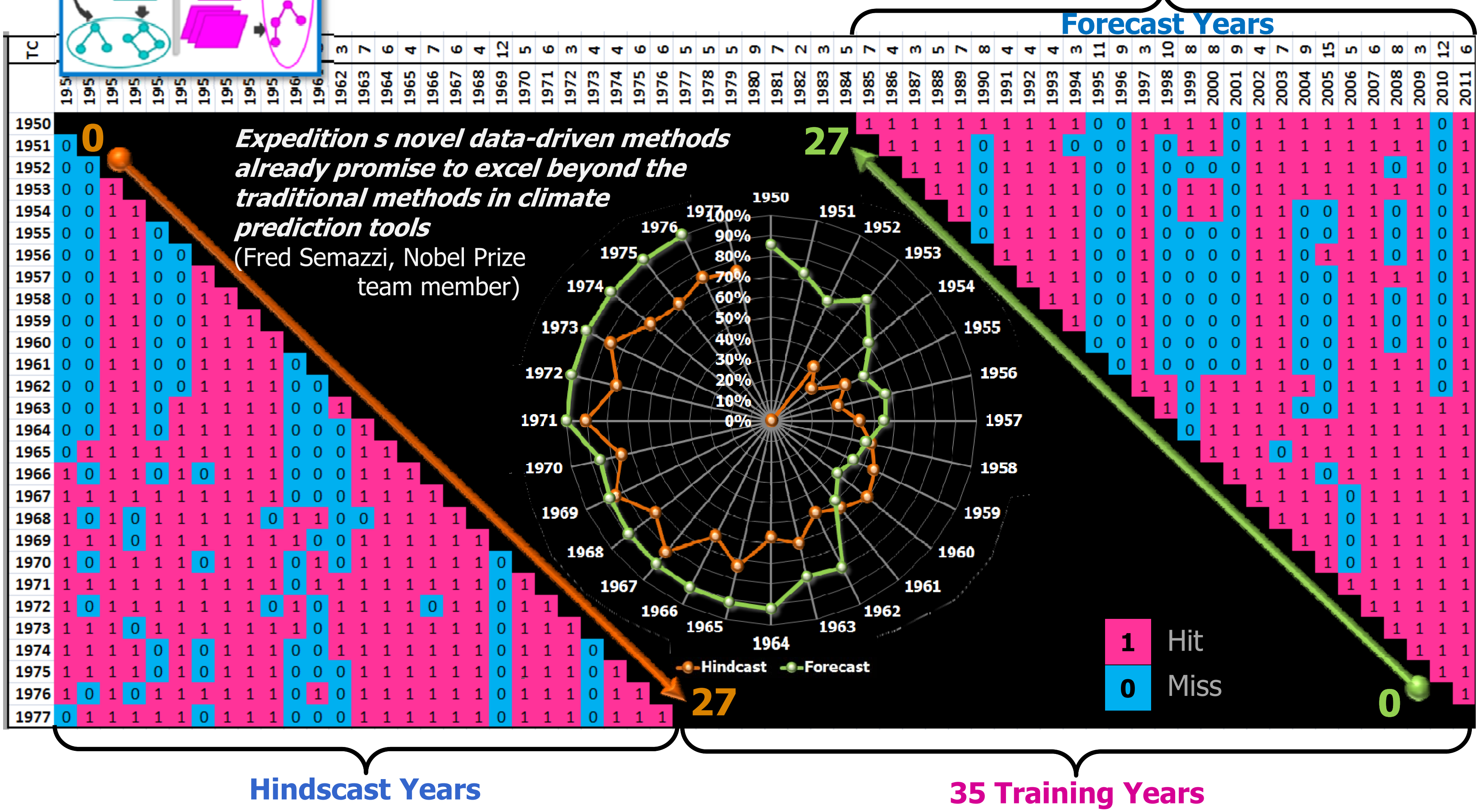
Intuition: If an extreme event (e.g. hurricane track) is in one of its key phases (e.g. land-hitting), then there exist network motifs (recurrent patterns in climate networks) that are specific to that phase.



# Robust & Accurate Seasonal Hurricane Forecasts through Comparative Climate Networks Analytics

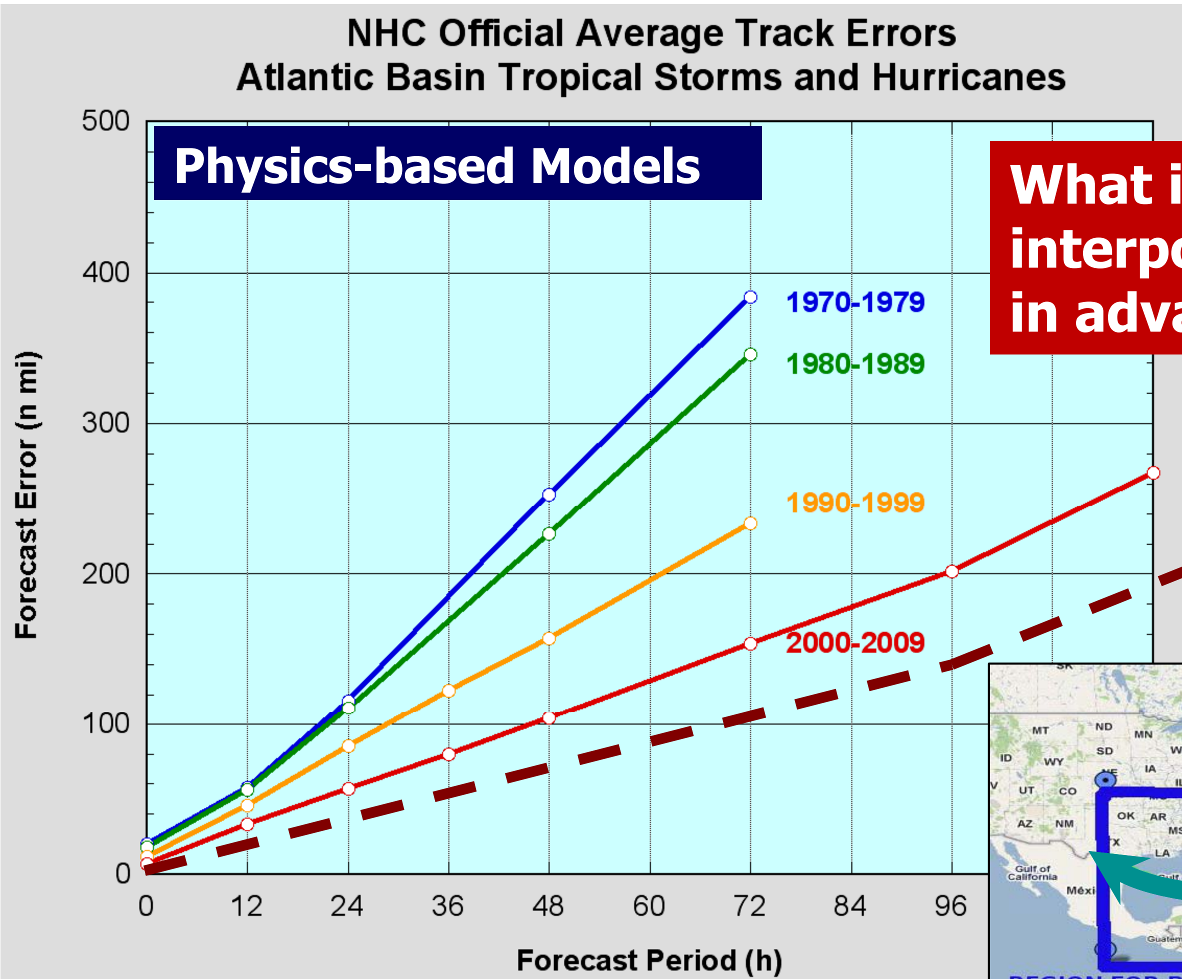


## Comparative analysis of climate networks leverages the DOE-funded network theory & scalable algorithms.



# Forecasting Hurricane Tracks

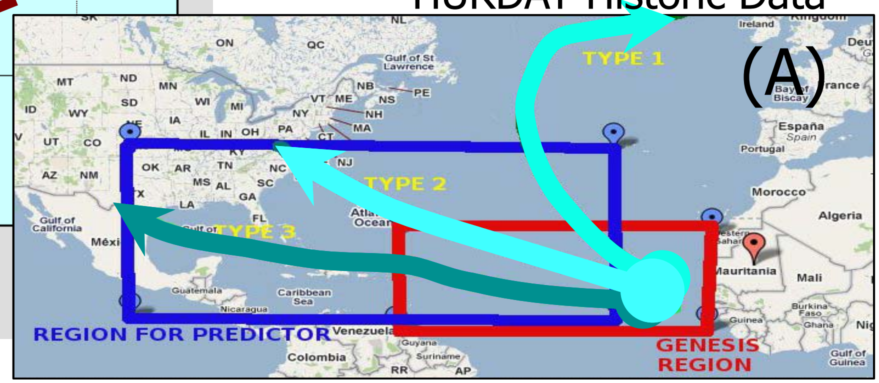
## Improving but have mean error (>185km) beyond 48 h




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

**~500 km**

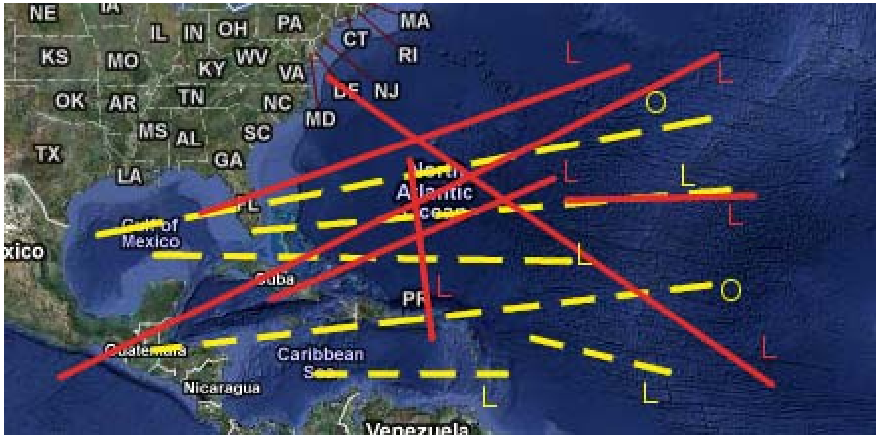
HURDAT Historic Data




Src: NOAA/NHC: http://www.nhc.noaa.gov/verification/index.shtml

## 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.




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

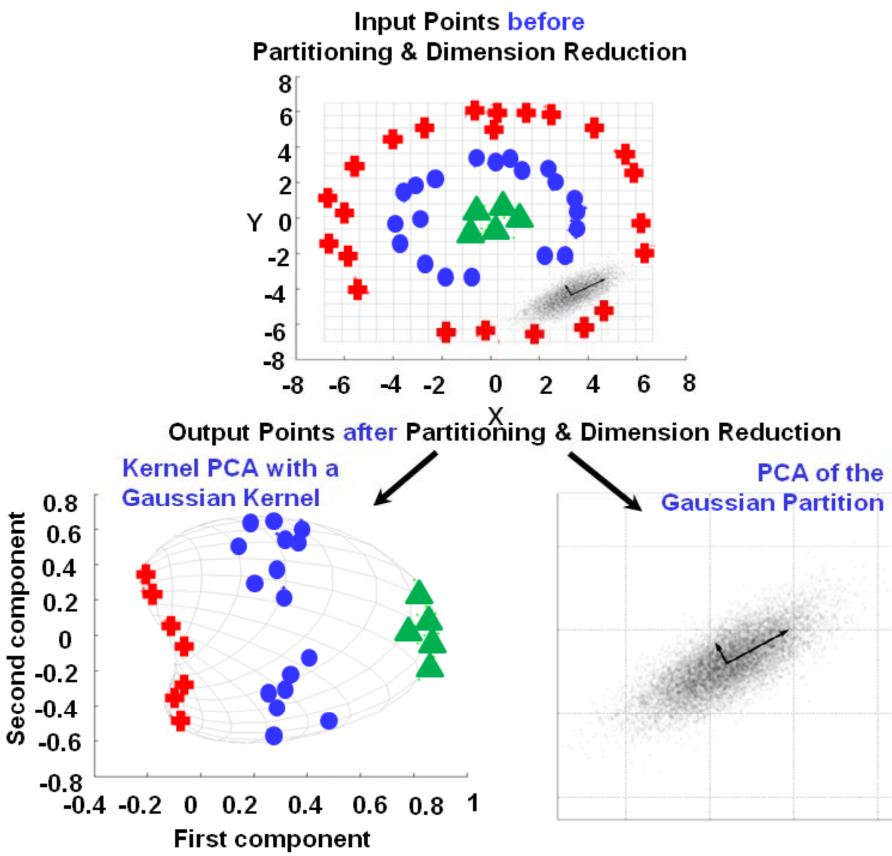
- Nearly **east-oriented SLP** edges suggest horizontal pressure gradient configuration in the same direction.
- Based on Buys Ballot's law, this pressure gradient would be associated with **wind flow in the north-south direction**.
- Onshore wind anomaly flow would promote favorable conditions for landfall;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

**Performance of Land-hitting vs. Offshore**

| | LOO | | | 10-FOLD | |
|---|---|---|---|---|---|
| | **SLP** | **SST** | **SLP+SST** | **SLP** | **SST** |
| **Accuracy** | 0.88 | 0.90 | 0.92 | 0.90 | 0.90 |
| **Sensitivity** | 0.91 | 0.96 | 0.97 | 0.95 | 0.97 |
| **Specificity** | 0.77 | 0.76 | 0.81 | 0.80 | 0.74 |
| **Precision** | 0.90 | 0.90 | 0.92 | 0.92 | 0.90 |
| **F1-meas.** | 0.90 | 0.93 | 0.94 | 0.93 | 0.93 |

# Hierarchical Modularity of Complex Systems:
## Multilevel Paradigm via Divide-and-Conquer Strategy

**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.**




**FORECASTER**

### 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**.

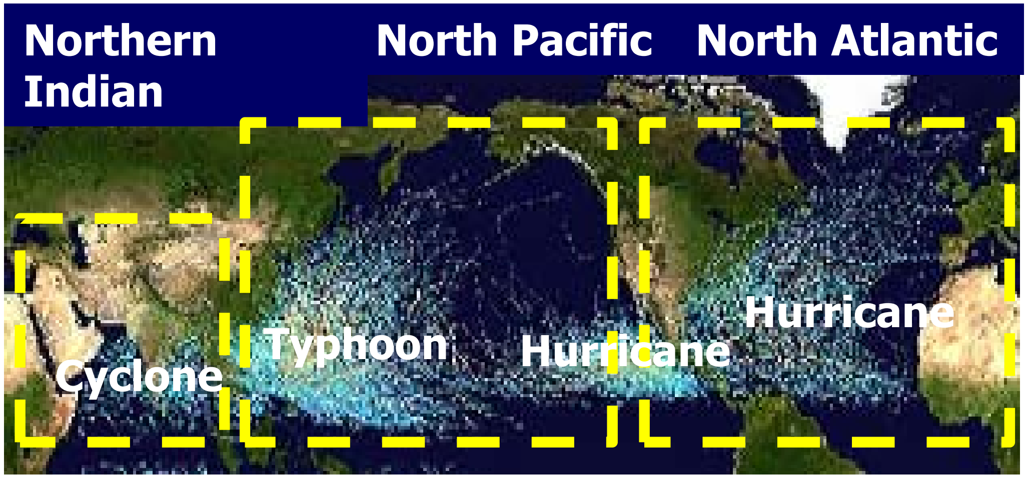
### Conquer Step:

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

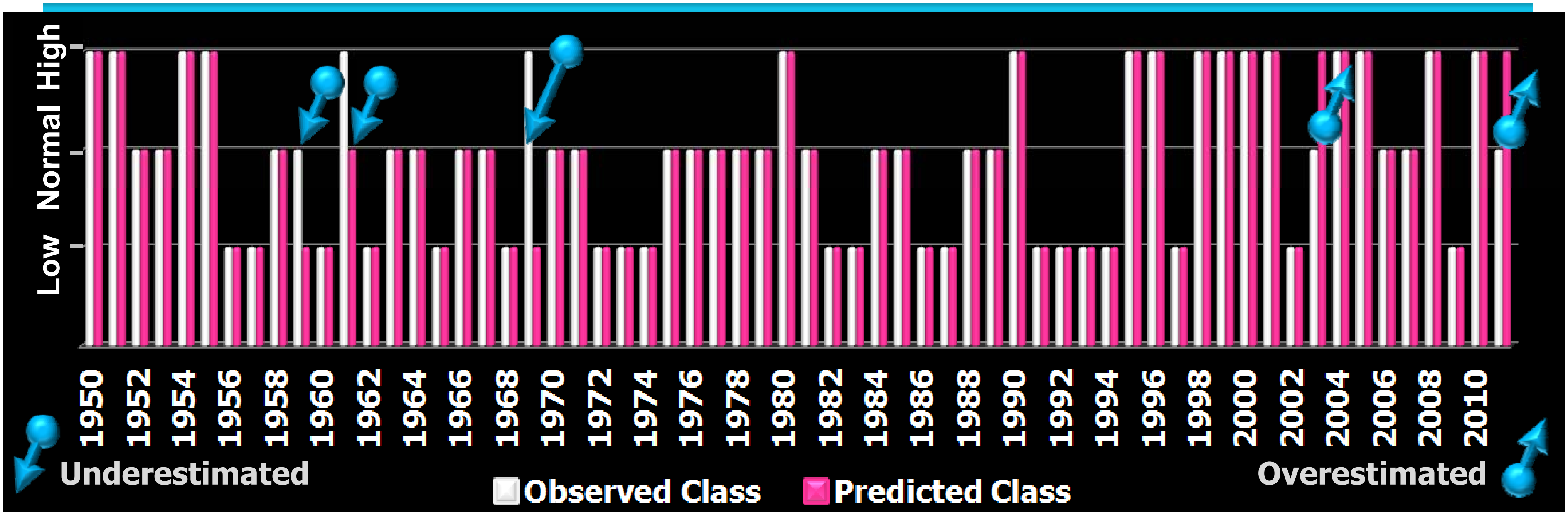
**There is an inherent interplay (e.g., feedback) between regional scale subsystems and the global scale system. Ignoring these relationships by focusing on a specific region is a simplification.**

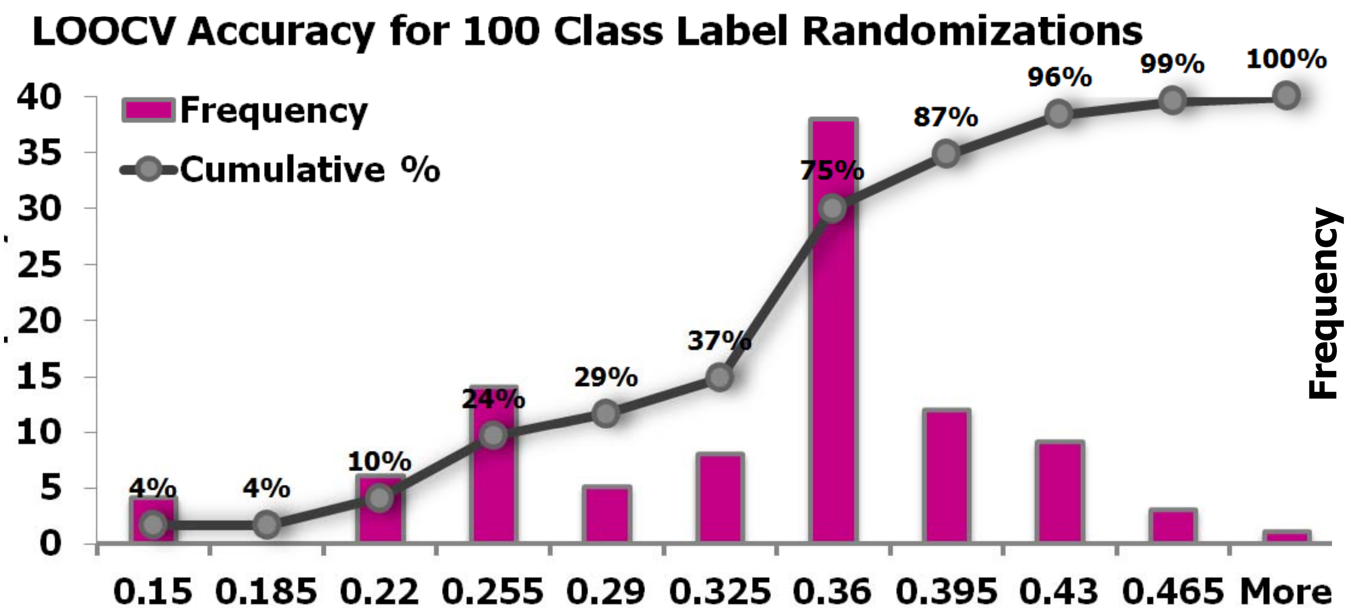
**DETECTOR**

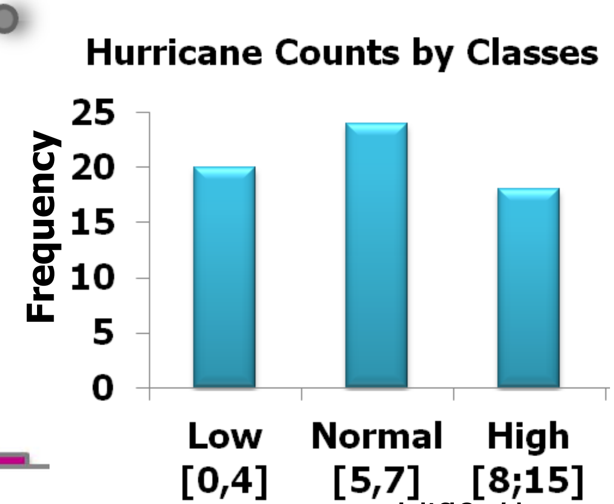



We could use these relationships for detecting the prediction errors and/or possibly correcting them.

# 92% Accuracy w/ Leave One Out Cross Validation
## Seasonal Hurricane Activity



# Hurricane Activity Class Forecast vs. State-of-art

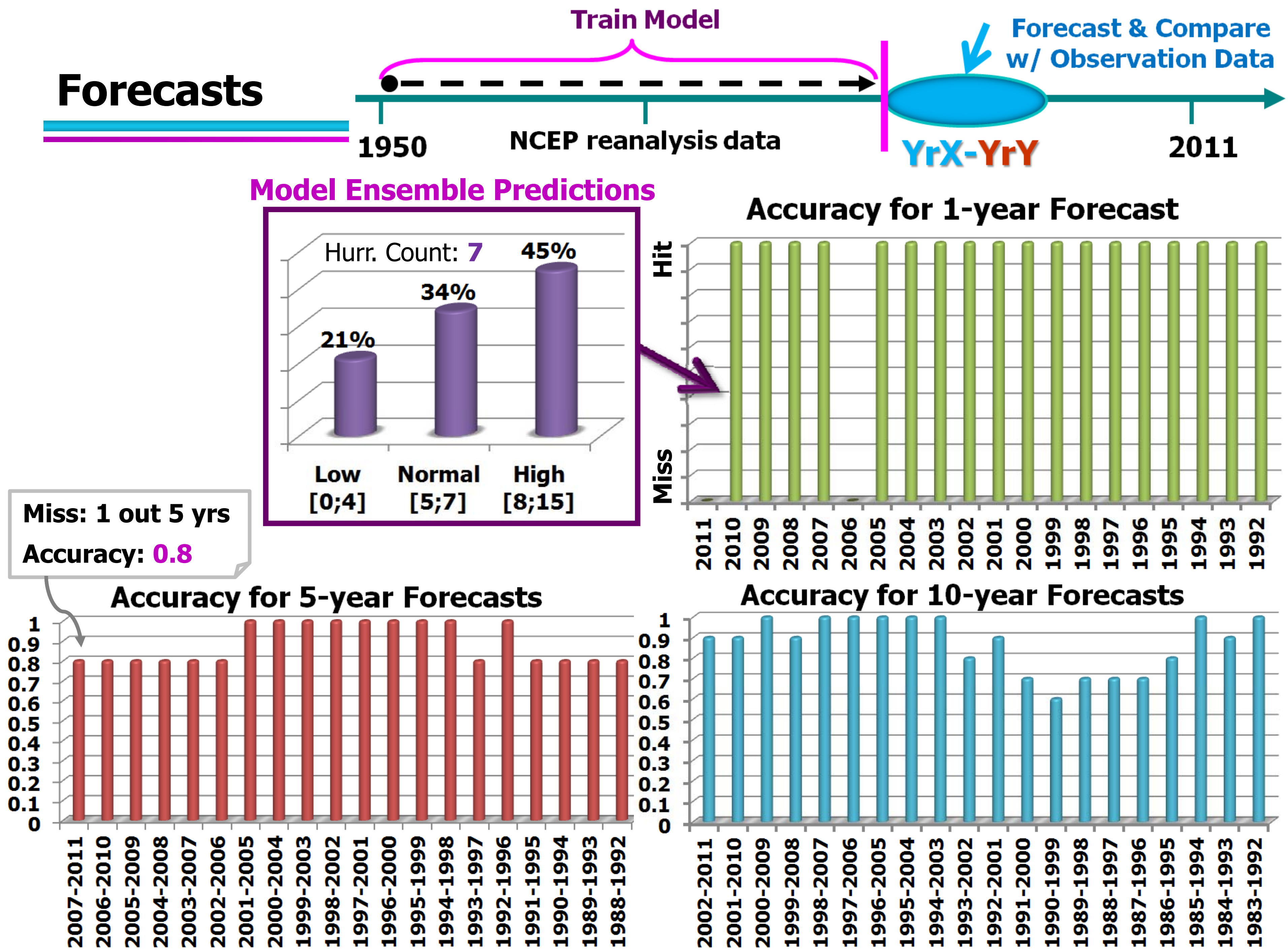
## FORECASTER Performance on North Atlantic Hurricane

| Metric | FORECASTER NC State | [1], 2009 Colorado | [2], 2010 GA Tech | Random Forest | Bagging | Boosting |
|---|---|---|---|---|---|---|
| **Accuracy (%)** | **93.3** | 64.0 | 65.5 | 76.7 | 73.3 | 75.0 |
| **HSS** | **0.90** | 0.45 | 0.49 | 0.66 | 0.60 | 0.62 |
| **PSS** | **0.92** | 0.44 | 0.50 | 0.65 | 0.63 | 0.63 |
| **GSS** | **0.96** | 0.50 | 0.68 | 0.65 | 0.67 | 0.66 |

**ML-based** **Regression** **Hybrid**

[1] P. J. Klotzbach and W. M. Gray, "Twenty-five years of Atlantic basin seasonal hurricane forecasts (1984-2008)," Geophys. Res. Lett., vol. 36, pp. L09 711, 5pp, May 2009.
[2] H. M. Kim and P. J. Webster. Extended-range seasonal hurricane forecasts for the North Atlantic with a **hybrid dynamical-statistical model**. Geophys. Res. Lett., 37(21):L21705, 2010.

**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;
**GS**: Gerrity score, occurrences substantially less frequent.



Forecasts
Train Model
Forecast & Compare
w/ Observation Data
1950
NCEP reanalysis data
YrX-YrY
2011
Model Ensemble Predictions
Hurr. Count: 7
21%
34%
45%
Low
[0;4]
Normal
[5;7]
High
[8;15]
Accuracy for 1-year Forecast
Hit
Miss
2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1994 1993 1992
Miss: 1 out 5 yrs
Accuracy: 0.8
Accuracy for 5-year Forecasts
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
2007-2011 2006-2010 2005-2009 2004-2008 2003-2007 2002-2006 2001-2005 2000-2004 1999-2003 1998-2002 1997-2001 1996-2000 1995-1999 1994-1998 1993-1997 1992-1996 1991-1995 1990-1994 1989-1993 1988-1992
Accuracy for 10-year Forecasts
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
2002-2011 2001-2010 2000-2009 1999-2008 1998-2007 1997-2006 1996-2005 1995-2004 1994-2003 1993-2002 1992-2001 1991-2000 1990-1999 1989-1998 1988-1997 1987-1996 1986-1995 1985-1994 1984-1993 1983-1992

# Effectiveness of DETECTOR + FORECASTER

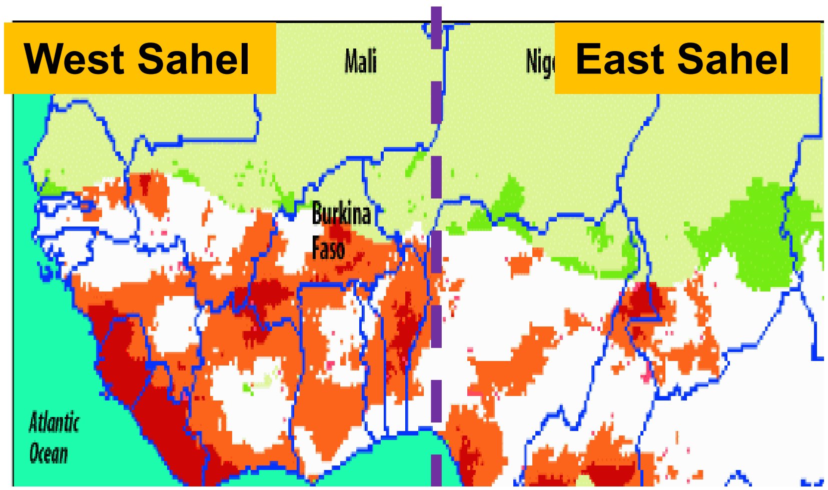
## Regional subsystems and global system interplays

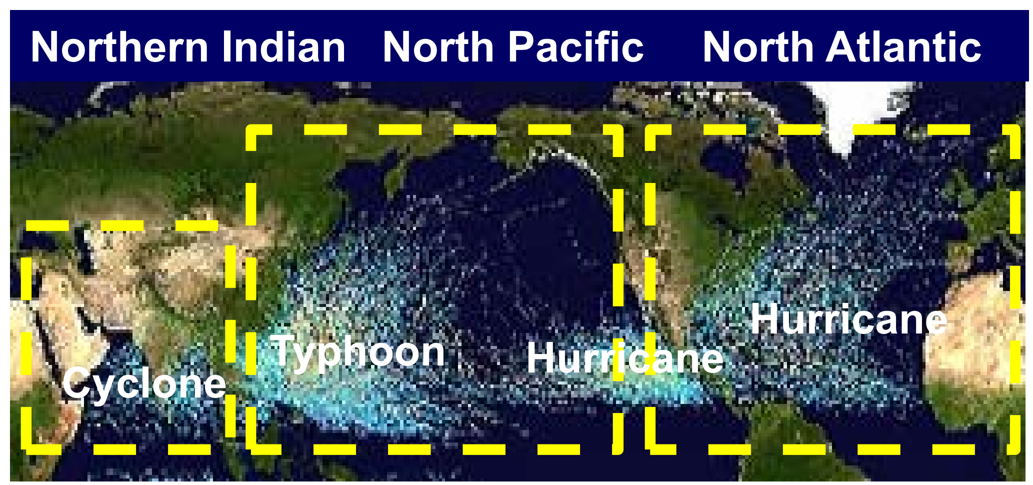
| Task | System | FORECASTER | DETECTOR + FORECATER |
|---|---|---|---|
| STCP | NH | 90.0 | **95.0** |
| | NA1 | 88.3 | **93.3** |
| SHP | NA2 | 93.3 | **98.6** |
| | LNA | 86.7 | **93.4** |
| NARP | SH | 88.9 | **94.5** |
| | WS | 90.7 | **96.3** |

**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.



# Predicted Network Motifs Agree with Climate Indices Related to Hurricane Activity

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

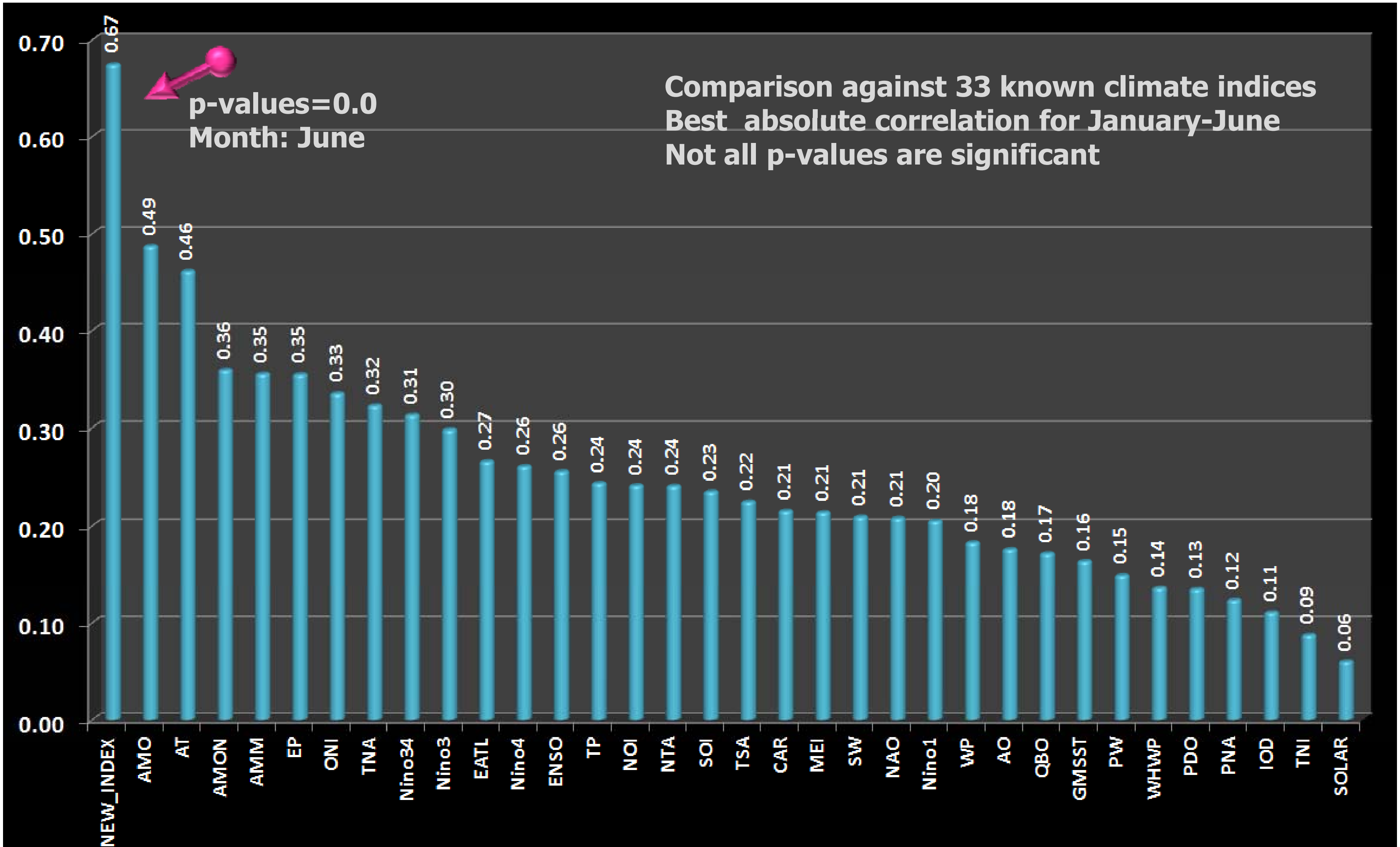
## Published Facts

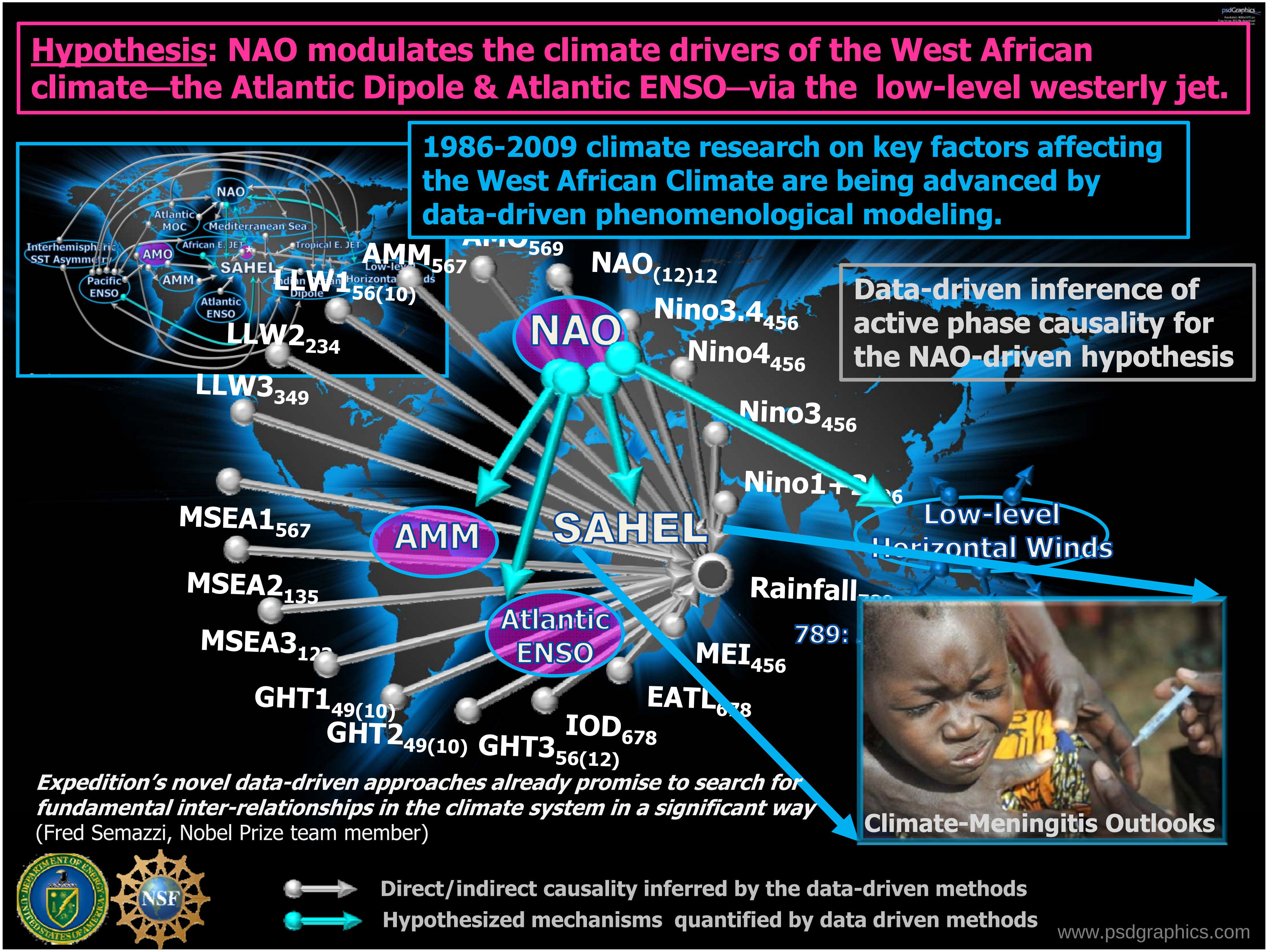
- Nino3 SSTs correlate with Atlantic hurricane activity
- ENSO modulates NA TCs
- SSTs in MDR contribute to hurricanes in MDR region
- NAO June correlates with NA hurricane tracks
- Shifts in the PDO phase can have significant implications for Atlantic hurricane activity

## New Hypotheses

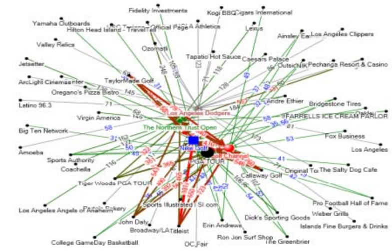
Atlantic multi-decadal Oscillation (AMO) and Arctic Oscillation (AO) indices might affect the North Atlantic tropical cyclone activities

# 0.67 Spearman Rank-order Correlation between Network-based Climate Index & Hurricane Activity



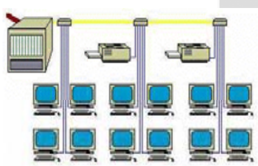

Hypothesis: NAO modulates the climate drivers of the West African climate—the Atlantic Dipole & Atlantic ENSO—via the low-level westerly jet.
1986-2009 climate research on key factors affecting the West African Climate are being advanced by data-driven phenomenological modeling.
Data-driven inference of active phase causality for the NAO-driven hypothesis
NAO
Atlantic MOC
Mediterranean Sea
Interhemispheric SST Asymmetry
AMO
African E. JET
Tropical E. JET
SAHEL
AMM
Pacific ENSO
Atlantic ENSO
AMM567
LLW1 56(10)
LLW2 234
LLW3 349
AMO569
NAO(12)12
Nino3.4 456
Nino4 456
Nino3 456
NAO
SAHEL
AMM
Atlantic ENSO
Low-level Horizontal Winds
MSEA1 567
MSEA2 135
MSEA3
GHT1 49(10)
GHT2 49(10)
GHT3 56(12)
IOD678
MEI456
789:
Climate-Meningitis Outlooks
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)
NSF
Direct/indirect causality inferred by the data-driven methods
Hypothesized mechanisms quantified by data driven methods
www.psdgraphics.com

# Summary: Discovering Knowledge from Massive Data – Next Frontier for HPC

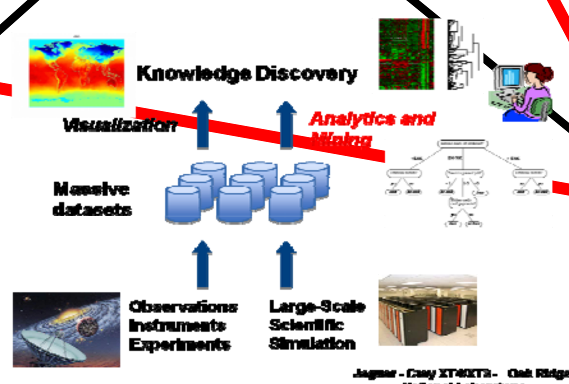


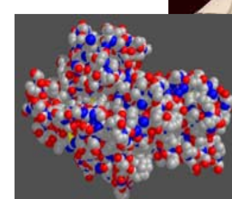
Business

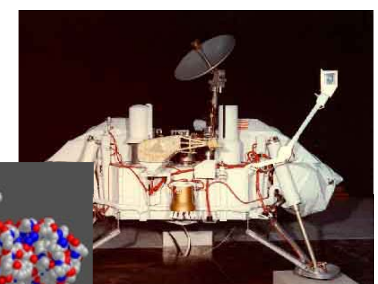

Data management, High-End Analytics, Data Mining, and Network Mining


Engineering






Science