Annual Report for Period: 09/2011 - 08/2012 **Submitted on:** 08-31-2012
Principal Investigator: Kumar, Vipin **Award ID:** 1029711
Organization: University of Minnesota

Title: Collaborative Research: Understanding Climate Change: A Data Driven Approach

Project Participants

Name: Kumar, Vipin
Worked for more than 160 Hours: Yes
Contribution to Project: Provide overall leadership on the project. Direct and guide research on mining climate data using graph-based methods for dipole discovery; time-series analysis for monopole discovery; various approaches for discovery and analysis of teleconnection patterns, atmosphere-ocean-land interactions, and long-term trends.

Name: Banerjee, Arindam
Worked for more than 160 Hours: Yes
Contribution to Project: Research on machine learning and data mining with emphasis on structure learning and predictive modeling with sparse statistical climate networks, graphical models for abrupt change detection, and paleoclimatic reconstructions using matrix completion methods.

Name: Boriah, Shyam
Worked for more than 160 Hours: Yes
Contribution to Project: Research on change detection algorithms, specifically on scalable techniques for global change detection of deforestation, urbanization and other significant land cover change events that have interrelationships with climate.

Name: Chatterjee, Singdhansu B.
Worked for more than 160 Hours: Yes
Contribution to Project: Research on statistical methods and theoretical components relating to change detection in climate networks, extreme climate modeling, modeling of tropical storm patterns, model selection, multi-model ensembles, attribution and climate change impact, spatio-temporal dependent climate data analysis, Bayesian and resampling theory.

Name: Das, Debasish
Worked for more than 160 Hours: Yes
Contribution to Project: Precipitation extremes characterization and predictive modeling based on climate model simulations and sensor-based observations, as well as new methods that can bring together a variety of new methods in extreme value theory, dependence analysis, spatial and spatiotemporal data mining, as well as dominant physical processes.

Name: Ganguly, Auroop
Worked for more than 160 Hours: Yes
Contribution to Project: Overall direction of climate extremes research, specifically, characterization of extremes and regional change, multi-model evaluation and uncertainty, and enhanced data-driven projections, as well as research in water impacts of climate change. Contributions to statistical or machine learning graphical and network based models for predictive modeling in climate.

Name: Foley, Jon
Worked for more than 160 Hours: No
Contribution to Project: Expertise and guidance for research on abrupt change detection in environmental systems, including land use changes and interaction among climate and ecosystems.

Name: Knight, Joseph F.
Worked for more than 160 Hours: Yes
Contribution to Project: Developed methodologies to use multi-temporal satellite image data for land use mapping and change detection. Supervised graduate student research.

Name: Liess, Stefan
Worked for more than 160 hours: Yes
Contribution to Project: Analysis of global climate networks in observations, reanalysis data, model projections, and multi-model ensembles. Detection of teleconnections with focus on their behavior in climate change scenarios and their impact on abrupt climate change such as droughts. Prediction of tropical cyclone activity with statistical methods and research on impact of climate change on cyclogenesis and tropical cyclone intensity.

Name: Shekhar, Shashi
Worked for more than 160 hours: Yes
Contribution to Project: Research on mining spatiotemporal patterns relating to climate change, including detecting abrupt change intervals in spatial (e.g., ecotones) and temporal (e.g., droughts) climate data, spatial decision tree for geographical classification for land cover, and analyzing the spatial autocorrelation in ecoclimate data (e.g., NDVI) with respect to data resolution.

Name: Snyder, Peter K.
Worked for more than 160 hours: Yes
Contribution to Project: Research on use of wavelet analysis for the detection of abrupt changes in environmental systems, including abrupt precipitation changes, both with the observational record and with CMIP5 climate model output for the 21st century. Evaluation of historical CMIP5 model performance and dynamical downscaling of future projections for select models to investigate extreme precipitation events with climate change.

Name: Steinbach, Michael
Worked for more than 160 Hours: Yes
Contribution to Project: On the technical side, contributed to the research related to the discovery of dipoles from data and the relationship of hurricanes to Sea Surface Temperature (SST). Also share some responsibilities for various administrative tasks related to the project.

Name: Steinhaeuser, Karsten
Worked for more than 160 Hours: Yes
Contribution to Project: Direction of research on correlation-based networks and other graph-based methods for descriptive analysis and predictive modeling of climate phenomena from observed and model-simulated data, study of ocean-land interactions; perform share of administrative duties related to the project.

Post-Doc

Graduate Student

Name: Chatterjee, Soumyadeep
Worked for more than 160 Hours: Yes
Contribution to Project: Develop structured sparse regression methods for prediction of land variables such as temperature and precipitation based on ocean variables. Establish rates of convergence of such methods for low-sample high-dimensional regimes.

Name: Chowdhury, Aritra
Worked for more than 160 Hours: No
Contribution to Project: Review of existing methods of reconstructions of past climate based on proxies. Investigation of suitability of matrix completion methods for paleoclimatic reconstructions.

Name: Corcoran, Jennifer
Worked for more than 160 Hours: No
Contribution to Project: Develop new land cover/use mapping methods based on decision trees that can be used to create improved inputs to climate models.

Name: Faghmous, James
Worked for more than 160 Hours: No
Contribution to Project: Designed efficient network construction algorithms for entire group. Mentored three undergraduate students. Designed data-analytic methods to investigate SST-Hurricane relationship. Identified and built useful climate datasets for group research. Managed collaboration across Co-PIs for our hurricane research.

Name: Fu, Qiang
Worked for more than 160 Hours: Yes
Contribution to Project: Developed scalable methods, based on graph-structured linear programs, for detecting significant droughts from spatiotemporal precipitation data. Developed parallel implementations which scales with use of multiple cores.

Name: Harding, Keith
Worked for more than 160 Hours: Yes
Contribution to Project: Evaluation of historical CMIP5 model performance and dynamical downscaling of future projections for select models to investigate extreme precipitation events with climate change.

Name: Heyman, Megan
Worked for more than 160 Hours: Yes
Contribution to Project: Statistical dimension reduction techniques and statistical inference and data analysis in high-dimensional problems. Research projects include alternative techniques of penalized regression estimation that may lead to better quality statistical inference, resampling techniques for penalized regression, and using probabilistic matrix factorization techniques for dimension reduction.

Name: Jiang, Zhe
Worked for more than 160 Hours: Yes
Contribution to Project: Develop efficient computational approaches for learning spatial decision tree models for geographical classification (e.g., land cover classification, wetland mapping).

Name: Karpatne, Anuj
Worked for more than 160 Hours: No
Contribution to Project: Working on change detection local regression methods for plant trait prediction, and relationship mining using climate indices.

Name: Kawale, Jaya
Worked for more than 160 Hours: Yes
Contribution to Project: Developing data mining and signal processing techniques for analysis of climate dipoles.

Name: Kodra, Evan
Worked for more than 160 Hours: Yes
Contribution to Project: Characterization and predictive analysis of heat waves, cold snaps and heavy precipitation from multiple climate and earth system models, along with methods for model combinations and uncertainty characterization for climate extremes and impacts of regional climate change.

Name: Kumar, Arjun
Worked for more than 160 Hours: Yes
Contribution to Project: Developing data mining and signal processing techniques for analysis of dipoles and monopoles (oscillations).

Name: Kumar, Devashish
Worked for more than 160 Hours: Yes
Contribution to Project: Precipitation extremes characterization and predictive modeling based on climate model simulations and sensor-based observations, as well as new methods that can bring together a variety of new methods in extreme value theory, dependence analysis, spatial and spatiotemporal data mining, as well as dominant physical processes.

Name: Lu, Ying
Worked for more than 160 Hours: Yes
Contribution to Project: Techniques for detecting change in parameters in generalized linear models and other exponential family-related statistical frameworks, techniques for detecting change in extreme value distributions, and techniques for detecting change in relationship between complex, dependent climatic variables.

Name: Mithal, Varun
Worked for more than 160 Hours: No
Contribution to Project: Worked on a number of change detection algorithms, most recently Bayesian change detection algorithms. He is also modeling spatial dependence through Markov Random Fields.

Name: Nandy, Abhishek
Worked for more than 160 Hours: Yes
Contribution to Project: Statistical model selection and model averaging methods. Research focus is mainly on comparing models with dissimilar parameters, or on not comparable parameter spaces.

Name: Tolen, Joshua
Worked for more than 160 Hours: No
Contribution to Project: Multimodel evaluation and uncertainty for regional temperature and precipitation.

Name: Rampi, Lian
Worked for more than 160 Hours: No
Contribution to Project: Developing improved wetland mapping methods based on geographic object based image analysis that can be used to create improved inputs to climate models.

Name: Yuan, Sen
Worked for more than 160 Hours: Yes
Contribution to Project: Characterization of quantiles and extreme quantiles as minimizers of convex contrast functions, their multivariate generalization, and techniques for statistical inference relating to multivariate extreme quantiles. Extension to regression and dependent data framework is currently under study.

Name: Zhou, Xun
Worked for more than 160 Hours: Yes
Contribution to Project: Designed and optimized efficient computational approaches to the discovery of spatiotemporal change intervals from eco-climate data, e.g., periods of abrupt precipitation change, ecotones such as Sahel, etc. Developed approaches to quantifying the sensitivity of spatial autocorrelation in eco-climate data at varying resolutions.

Undergraduate Student

Name: Blank, Mace
Worked more than 160 hours: Yes
Contribution to Project: Develop tool with graphical user interface for investigating relationships between ocean indicators and land climate, weather events, natural disasters, etc.

Name: Breidenbach, Laina
Worked more than 160 hours: Yes
Contribution to Project: Climate data preprocessing for use in graph-based analyses and predictive modeling.

Name: Das, Debadrita
Worked for more than 160 Hours: Yes
Contribution to Project: Climate model bias and multimodel agreement for global precipitation assessments.

Name: Haasken, Ryan
Worked more than 160 hours: Yes
Contribution to Project: Design algorithms to quickly analyze the relationship between Atlantic sea surface temperatures and Atlantic hurricane activity.

Name: Le, Matthew
Worked more than 160 hours: Yes
Contribution to Project: Data preparation and analysis for studying the impact of Pacific ocean warming on Atlantic Hurricane activity.

Name: Ormsby, Dominick
Worked more than 160 hours: Yes
Contribution to Project: Data processing, research on complex networks and dipole detection; develop tool with graphical user interface to visualize and investigate dipoles.

Name: Smith, Graham
Worked more than 160 hours: Yes
Contribution to Project: Analyze evolution of climate networks; analysis and visualization of trends in climate data.

Name: Styles, Luke
Worked more than 160 hours: Yes
Contribution to Project: Develop algorithms for ocean eddy detection and monitoring in sea surface height data; develop tool with graphical user interface for data exploration and analysis.
Name: Welter, Roland
Worked more than 160 hours: Yes
Contribution to Project: Climate data preprocessing for use in graph-based analyses and predictive modeling.

Technician, Programmer

Other Participants

Name: Beard, Joel
Worked for more than 160 Hours: Yes
Contribution to Project: Research on high-dimensional causality and attribution techniques and data-depth techniques.
*Joel was supported by a teaching assistantship and therefore not paid from NSF funds.

Name: Gibson, Nikai
Worked for more than 160 Hours: Yes
Contribution to Project: Data exploration and generation of case studies for ocean eddy detection.
*Nikai is a summer student who was paid from UMN funds.

Name: Dietz, Lindsey
Worked for more than 160 Hours: Yes
Contribution to Project: Eliciting the pattern of tropical storms that develop in the Atlantic basin, using a blend of physics-based prior knowledge and data from tropical storms and hurricanes. The relationship between the minimum central pressure and the maximum sustained windspeed is under study.
*Lindsey was supported by a teaching assistantship and therefore not paid from NSF funds.

Name: Middleton, James
Worked for more than 160 Hours: Yes
Contribution to Project: Investigate relationships between ocean indicators and land climate, weather events, natural disasters, etc.
*James is a summer student who was paid from UMN funds.

Name: Pearson, Zachary
Worked for more than 160 Hours: Yes
Contribution to Project: Develop regression methods for prediction of land variables such as temperature and precipitation based on ocean variables.
*Zachary is a summer student who was paid from UMN funds.

Name: Sumler, Rahni
Worked for more than 160 Hours: Yes
Contribution to Project: Develop methods for detecting and visualizing spatio-temporal 'hot spots' from historical climate records
*Rahni is a summer student who was paid from UMN funds.

Organizational Partners

Northeastern University (NEU), North Carolina State University (NCSU), Northwestern University (NWU), North Carolina A & T University (NCAT)

Other Collaborators or Contacts

Subhadip Bandyopadhyay (Infosys India), Arnab Bhattacharjee (Dundee, UK), Yang Cheng (Census), Sloan Coats (Columbia), John Drake (UTK), David Erickson (ORNL), Subimal Ghosh (IIT Bombay, India), Eric Kihn (NOAA-NGDC), Martin Klein (Census), Ken Knapp (NOAA-NCDC), Partha Lahiri (Maryland-College Park), Taps Maiti (Michigan State), Anatoli Melechko (NCSU), Michel dos Santos Mesquita (Bjerknes Centre), Diganta Mukherjee (Usha Martin Academy, India), Nitai Mukhopadhyay (Virginia Commonwealth), Richard Seager (Columbia), Jason Smerdon (Columbia), John Tillinghast (Census), Bak Tran (Census), Tom Wilbanks (ORNL), Tommy Wright (Census), Hao Ren Wu (Michigan State)

Activities and Findings

Education Activities: Training and Development
The education and mentoring opportunities from this Expeditions project go significantly above and beyond what can be achieved with a single PI-driven or smaller-scale collaborative project. The interdisciplinary expertise of the project partners enables the development of cutting-edge skills and capabilities at the intersection of computer science, computational statistics, network science and high-performance computing on one hand, with global and regional climate modeling and assessments of climate impacts on natural and built environments on the other. The following provides several specific examples of interdisciplinary training opportunities and cross-institutional collaborations resulting from this project

Advising and Mentoring:
The students and postdoctoral researchers involved in the project have been exposed to several important aspects of research: they participated in research meetings and discussions, wrote technical papers, and gave public presentations about research work. Furthermore, these students have been trained in the fields of data mining, statistics, climate data analysis, pattern recognition, image processing, and the design and implementation of algorithms including the use of distributed platforms. Regular weekly project meetings at individual institutions as well bi-weekly conference calls including all project partners provide students an opportunity to develop collaborative research skills for working in interdisciplinary environments and addressing grand challenge problems facing society. Members of this Expeditions project have acquired

or strengthened necessary leadership and team-building skills that will help them achieve their professional and research goals.

So far, the members of the Expedition team advised and mentored 31 PhD, 3 MS, and 16 undergraduate students, 5 research scientists, and 3 Postdoctoral fellows, and 8 internships. They have graduated a total of 4 PhD students and 2 MS students.

Many students are co-supervised by the faculty from different disciplines. Examples include but not limited to the following:

- Evan Kodra (NEU), a graduate student with a background in statistics, is working directly with Auroop Ganguly (NEU), a civil and environmental engineer with prior experience in climate extremes and uncertainty. Together, they are collaborating with Snigdhansu Chatterjee (UMN) to develop new methods in statistics and Arindam Banerjee (UMN) to understand how new advances in machine learning can be leveraged to further develop and refine these approaches.
- Fred Semazzi (climate scientist) and Nagiza Samatova (computer scientist, data mining) at NCSU co-advise several PhD students and co-serve on the students' PhD. Jointly with the researchers from UMN (Vipin Kumar) and NWU (Alok Choudhary, William Hendrix) these students have co-published several papers in premier and top journals and conferences (e.g., *Data Mining and Knowledge Discovery*, *SIAM Data Mining*). Their accomplishments have been highlighted through the DOE, Office of Science, News Release (e.g., http://ascr-discovery.science.doe.gov/universities/hurricane1.shtml).
- Abdollah Homaifar (Electrical and Computer Engineering, machine learning specialist) along with Ken Knapp and Erin Kihn at NOAA's National Climatic Data Center and the National Geophysical Data Center respectively are co-advising three PhD students from NCAT.
- Research training for three graduate students from the School of Statistics, and several others from other departments at the University of Minnesota.

Expedition-trained and -graduated PhD students are currently faculty and staff at universities including Northwestern University (USA); Prince of Songkla University, Pattani Campus, Pattani (Thailand); and Kwame Nkrumah University of Science and Technology (Ghana, Africa). Other students are currently employed by the Oak Ridge National Lab (1) and Northwestern University (1).

Broader Impact:
The Expedition team has strong commitment in promoting diversity and K-12 in mathematics, science, and computing. One of our important goals is to engage students (including women and those from underrepresented minorities) to create interest and awareness of the work in this area.

To date, we have mentored a total of 10 female researchers including 1 postdoc, 4 PhD, 2 MS, 3 undergraduates. Two of the PhD female students have recently graduated; one secured a faculty position in Thailand.

The Expeditions team has been collaborating with the Interdisciplinary Scientific Environmental Technology (ISET) program, whose objective is to provide opportunities for underrepresented students to study climate or environmental sciences and the related technologies. Through ISET and other programs we have mentored 13 researchers from underrepresented minority groups including 2 female PhD students, 1 PhD male student, and 10 undergraduate students. This includes seven undergraduate students from NCAT's Electrical and Computer Engineering and Computer Science Departments who have participated in summer internships at UMN through the ISET program during 2011 and 2012. These students are acquiring valuable knowledge and skills that will enable them to continue working on the project at the conclusion of their internships, which in turn will strengthen the collaboration and generate new opportunities for interaction between the partner institutions. For example, one of the NCSU

undergraduate students funded through the NSF REU supplement award has expressed his appreciation of the provided opportunities to his research mentor, Nagiza Samatova:

> *I just wanted to thank you for offering me the invaluable opportunity of working on projects in your lab. I have learned a great deal during my time here and am now fully committed to pursuing a career in research. The guidance that you and your MS/PhD students have provided me has been the key in this regard. My sincerest thanks go out to all of you for your efforts in helping me grow as both a scientist and a student.*

We have supported 11 undergraduate students through the NSF REU supplement award. These undergraduates have been involved in the writing of workshop, conference, and journal papers, and book chapters. Some of these students have attended international conferences (ICDM'11 (Vancouver, Canada), *CLUSTER*'12 (Beijing, China) to present their research findings at those conferences. As an example of international student outreach, the Expeditions project funded an undergraduate summer intern (Debadrita Das) from the Civil Engineering department of the Indian Institute of Technology at Kharagpur. Members of the Expedition also supervise undergraduate honors theses, e.g., Nagiza Samatova (NCSU) is advising 3 undergraduate honors theses focusing on 'Big Data' analytics and predictive climate informatics.

Industry and Government Lab Impact:
The team work and the acquired research skills and multi-disciplinary education have prepared the Expedition students to be an attractive workforce for industry. Several of our students have been invited to participate in the industry and national government lab research projects through summer internship programs. For example, students have interned with IBM, Microsoft, Argonne National Lab, and Oak Ridge National Lab, BlueCross BlueShield, and Inmar. Several of them were able to directly build upon their current research with the Expedition and provide a significant contribution to their industry projects. For example, Saurabh Pendse, while at Microsoft, built a production system for indexing and compression (based on his research effort) and was given an opportunity to present his work to the VP. Chaunté Lacewell, interned at Intel Corporation as part of Intel Collaborator Program and has received the highest rating in the group (Outstanding).

Likewise, the students interning at national labs have been working on technology transfer from the research prototypes conducted under the Expeditions project to the production software that is planned to be part of the DOE labs supercomputing facilities. In a more International example, one of our students, James Faghmous, won a Ph.D. student James Faghmous won a 2011 NSF Nordic Research Opportunity Grant to visit Norway to conduct research at one of the world's top climate research institutes, the Bjerknes Center for Climate Research (BCCR) in Bergen, Norway. As a result, a new area of research on ocean eddies was initiated within our group.

Instructional Development:
The Expedition project provided our faculty an outstanding opportunity to enrich the university curricula, both at the undergraduate and graduate levels, in a number of ways.

In order to provide a ***Climate Education Community Interface***, Prof. Fred Semazzi at NCSU has founded the Professional Science Master's (PSM) degree program in Climate Change & Society (CCS). The degree program is

intended for students interested in careers in planning or policy and professionals working in government agencies or private sector firms concerned with any aspect of planning or setting policies affected by global climate change.

Capitalizing on the Expeditions' research activities on predictive modeling and complex networks, Nagiza Samatova (NCSU), has created a special concentration of the graduate (CSC 522) and undergraduate (CSC 422) courses on Graph Mining with an annual enrollment of more than 60 students. The undergraduate and graduate students in that course have written and co-edited a textbook titled "***Practical Graph Mining with R***" that is in press with the Chapman & Hall/CRC Press under the Data Mining and Knowledge Discovery Series. All the proceeds from the sale of the students' textbook will go to the NC State Department of Computer Science to promote early scientific career development among undergraduate and graduate students.

Likewise, Vipin Kumar and Michael Steinbach (UMN) have been enriching their highly popular CSci 5523 Data Mining course, which enrolls almost 100 students from a wide variety of departments, by emphasizing topics on scientific data mining, and mining climate and earth science data. In addition the discussion of relevant topics during the course, many examples used to illustrate various techniques come from the earth science and climate domains. Furthermore, one entire lecture is devoted to topics related to ongoing work in our climate and earth science, e.g., change detection and complex networks.

Snigdhansu Chatterjee (UMN) has incorporated large take-home projects on climate data analysis for groups of students in courses taught at the University of Minnesota (Stat 5601 – Nonparametric Methods in Statistics; Stat 5401 – Applied Multivariate Statistics). Chatterjee also designed a directed reading course for graduate student Ying Lu on change detection and climate-related topics and directed a summer research internship for graduate student Joel Beard. He initiated a course on advanced topics in Statistics, titled Climate Statistics. This course was labeled Stat 8931, and is likely to be offered repeatedly in the next few years.

Outreach

Our team has been involved in a large number of outreach activities aimed at a wide range of audiences. These efforts include engaging other researchers and organizations in discussions and collaborations; organizing conferences, workshops, and sessions at academic meetings; delivering invited talks and presentations in different venues; releasing open-source implementations of tools and software; and so on. The full breadth of these activities is described in more detail below.

Collaborations & Community Engagement

Several members of the Expeditions team are participating in community outreach programs targeting undergraduate students in an effort to drive interest in research and furthering the sciences. Presentations have been given through community programs to audiences including parents, teachers, students and administrators – for example, Vipin Kumar delivered a lecture "Data Mining for Global Change" as part of the NASA Summer Short Course for Earth System Modeling and Supercomputing to a group of graduate students pursuing interdisciplinary studies.

In addition, individual institutions are developing new partnerships with other organizations. For instance, NCSU is working with the African Center of Meteorological Application for Development (ACMAD) regional center and its health-sector affiliates to transform research results on isolating distinct climate modes of spatially homogeneous behavior for climate-meningitis occurrence into operational products for

the prediction of meningitis epidemics. This activity will enable this project to support the recently created Global Framework of Climate Services operational program (http://www.climatesciencewatch.org/2009/09/10/new-global-framework-for-climate-services-should-strengthen-preparedness; http://unfccc.int/resource/docs/2010/smsn/igo/093.pdf). The findings will also be disseminated via other alternatives such as seminars to key institutions in Africa, e.g., the Kwame Nkrumah University of Science & Technology (KNUST), Ghana, where a climate science program was recently established. In the long term, this will help facilitate joint collaborative research. Thus, this project has promoted cultural awareness and integration of ideas from different sides of the global spectrum.

Another example is the ongoing effort by NCAT to build a strong connection with the National Climatic Data Center (NCDC). Team members have traveled to the NCDC to meet with key scientists. During this visit, students presented their research work and received feedback regarding feasibility of the joint work. Currently, we are in the process of assigning a co-advisor from the NCDC for our students and continue to hold a monthly teleconference with an NCDC scientist.

Our Expeditions team members are also engaged in project-related national and international activities in the scientific community at large. Fredrick Semazzi is a member of the Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre, which was established by the Task Group on Data and Scenario Support for Impact and Climate Analysis to facilitate the timely distribution of consistent scenarios of changes in climate and related environmental and socio-economic factors for use in climate impacts assessments. Auroop Ganguly was invited to be a reviewer for the IPCC's Fifth Assessment Report – Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) under preparation by the IPCC's Working Group II on Impacts, Adaptation and Vulnerability, and he served on a review panel of the United Nations Environmental Programme (UNEP) on Environmental Impacts of Climate Change.

Conference, Workshop, and Session Organization

An integral part of our project – both for disseminating our research and receiving feedback as well as learning about related work – is an annual workshop we have organized at the end of each project year. The First Workshop on Understanding Climate Change from Data was held on August 15-16, 2011 and the Second Workshop on August 6-7, 2012; both workshops had approximately 100 attendees. Each year we invited prominent researchers and practitioners from computer science, Earth and atmospheric sciences, geography, civil and environmental engineering, statistics, and other disciplines to participate in this event. We also invited submissions from graduate students and postdoctoral researchers from other institutions, whose attendance we supported via travel fellowships. Combined with contributions from our team members, the workshops have featured exciting programs including talks on topics relating to the goals and visions of the project, panel discussions on the present state and future of this research area, and poster sessions to showcase the students' work in a setting that encourages interaction between attendees. Feedback about these workshops has been overwhelmingly positive, and we plan to continue this annual workshop in future years.

In addition to our own workshop, team members have been involved in the organization of numerous other conferences, workshops, and sessions held at other academic meetings or events. In particular, in an effort to proactively reach out to the geoscience community, several team members are convening sessions at the Fall Meeting of the American Geophysical Union (AGU) later this year. This annual meeting attracts over 20,000 attendees is widely regarded as the most prominent gathering of Earth and space scientists worldwide. The following is a list of all past and planned events (co-)organized by members of the Expeditions team:

December 2012: Auroop Ganguly and Peter Snyder co-convene a session on Climate Extremes and Impacts: Can Big Data Mining and Fusion Help Reduce Uncertainties? at the AGU Fall Meeting

December 2012: Stefan Liess and Karsten Steinhaeuser co-convene a session on Advanced Methods for Pattern Recognition and Data Prospecting for Big Data at the AGU Fall Meeting

December 2012: Shyam Boriah and Auroop Ganguly co-convene a session on Spatio-temporal Change Detection and The Data Infrastructure of Environmental Observatories at the AGU Fall Meeting

December 2012: Vipin Kumar co-convenes a session on New Mechanisms, Feedbacks, and Approaches for Improving Predictions of the Global Carbon Cycle in Earth System Models at the AGU Fall Meeting

October 2012: Karsten co-organizes the Conference on Intelligent Data Understanding

September 2012: Karsten Steinhaeuser and Arindam Banerjee co-organize the *2nd International Workshop on Climate Informatics*

August 2012: The Expeditions Team hosted the ***2nd Annual Workshop on Understanding Climate Change from Data***

August 2012: Vipin Kumar co-organized the *ACM SIGKDD Workshop on Data Mining Applications in Sustainability*

August 2012: Debasish Das, Auroop Ganguly, and Karsten Steinhaeuser co-organized the *6th ACM SIGKDD Workshop on Knowledge Discovery from Sensor Data*

July 2012: Snigdhansu Chatterjee corganized a session on *Discovering Climate Patterns Using Large Data Analysis* at the Joint Statistical Meetings

June 2012: Auroop Ganguly co-convened a special session on *Modeling and Analytics for Hydrologic Impact Assessment due to Climate Change* at the International Conference on Computational Methods in Water Resources

April 2012: Karsten Steinhaeuser co-convened a session on *Complex Networks: Theory and methods applied to geophysical systems* at the EGU General Assembly

March 2012: Arindam Banerjee co-organized the Institute for Mathematics and Its Applications (IMA) *Workshop on Machine Learning: Theory and Computation*

December 2011: Auroop Ganguly, Vipin Kumar, Michael Steinbach, and Karsten Steinhaeuser co-organized the *3rd IEEE ICDM Workshop on Knowledge Discovery from Climate Data*

November 2011: Karsten Steinhaeuser co-organized the *Climate Knowledge Discovery Workshop* at Supercomputing

October 2011: Karsten Steinhaeuser co-organized the *NASA Conference on Intelligent Data Understanding*

August 2011: The Expeditions Team hosted the ***First Annual Workshop on Understanding Climate Change from Data***

August 2011: Shashi Shekhar co-chaired the *12th International Symposium on Spatial and Temporal Databases*

August 2011: Auroop Ganguly and Karsten Steinhaeuser co-organized the *5th ACM SIGKDD Workshop on Knowledge Discovery from Sensor Data*

July 2011: Arindam Banerjee co-organized the *ICML Workshop on Machine Learning for Global Challenges*

December 2010: Auroop Ganguly, Vipin Kumar, Michael Steinbach, and Karsten Steinhaeuser co-organized the *2nd IEEE ICDM Workshop on Knowledge Discovery from Climate Data*

Invited Talks

Many of our team members have given invited talks, lectures, and tutorials in a wide range of venues ranging from conferences and workshops at academic meetings to symposia and workshops organized by academic institutions, government organizations, or private industry.

August 2012: Snigdhansu Chatterjee gave a talk on "Using Multivariate Quantiles for High-Dimensional Inference in Survey Data Problems" at the US Census Bureau

August 2012: Vipin Kumar delivered the Innovation Award lecture at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining

July 2012: Alok Choudhary gave the keynote address "Discovering Knowledge from Massive Networks and Science Data – Next Frontier for HPC" at the Department of Energy Computational Science Graduate Fellowship Conference

July 2012: Vipin Kumar delivered a lecture "Data Mining for Global Change" as part of the NASA Summer Short Course for Earth System Modeling and Supercomputing

July 2012: Nagiza Samatova gave an invited talk "Accurate Forecasting of Adverse Spatio-Temporal Extreme Events" at the 2012 Institute for Computing in Science (ICiS) summer workshop on "*Graph and Hypergraph Problems in Computational Science: Applications and Algorithms*"July 2012: Snigdhansu Chatterjee gave a talk "Multivariate Generalized Spatial Quantiles and Applications" at the Second Institute of Mathematical Statistics – Asia Pacific Rim Meetings

June 2012: Snigdhansu Chatterjee gave a talk "Uncertainty quantification in change detection" at the IMS/ASA Spring Research Conference on *Enabling the Interface between Statistics and Engineering*

June 2012: Karsten Steinhaeuser gave an invited talk "Exploring Data Mining and Machine Learning Methods for Hydrology" at the International Conference on Computational Methods in Water Resources

June 2012: Vipin Kumar gave an invited talk "Understanding Global Change from Data" at the ARO Workshop on Big Data at Large

June 2012: Shashi Shekhar gave an invited talk "Spatial Big Data" at the ARO Workshop on Big Data at Large

June 2012: Nagiza Samatova gave an invited talk "On the Path to Sustainable, Scalable, & Energy-efficient Data Analytics" at the IEEE International Green Computing Conference

May 2012: Shashi Shekhar gave a keynote speech "Spatial Big Data" at the 11th ACM SIGMOD International Workshop on Data Engineering for Wireless and Mobile Access

May 2012: Auroop Ganguly gave a talk "Weather Extremes and Climate Change With Novel Computational Analysis & Modeling" at AIR Worldwide

May 2012: Karsten Steinhaeuser gave a presentation "Understanding Global Change from Data" to the EarthCube Data Discovery, Access, and Mining group

May 2012: Alok Choudhary gave the keynote address "Discovering Knowledge from Massive Social Networks and Science Data – Next Frontier for HPC" at the International Symposium on Cluster, Cloud and Grid Computing

May 2012: Shashi Shekhar gave a talk "Spatial Big Data" at the NSF OCI Workshop on Big Data Benchmarking

April 2012: Shashi Shekhar gave a talk "Spatial Data Mining" at the NIH-AAG Symposium on Geospatial Frontiers in Health and Social Environments

April 2012: Alok Choudhary gave a talk "On Data Intensive Computing and Exascale" at the International Exascale Software Project

March 2012: Shashi Shekhar gave a Distinguished Colloquium "Spatial Data Mining" at the Management Science and Information Systems Department at Rutgers University

March 2012: Nagiza Samatova and Alok Choudhary gave an invited talk "Understanding Climate Change: A Data-Driven Approach" to the DOE ASCAC Advisory Committee

March 2012: Shashi Shekhar gave a Distinguished Colloquium "Spatial Data Mining" at the Computer Science Department of Iowa State University

Februrary 2012: Fred Semazzi gave a talk "State of Seasonal climate prediction and applications" at the Thirtieth Greater Horn of Africa Climate Outlook Forum

January 2012: Snigdhansu Chatterjee gave a talk "Modeling climate characteristics using distribution-free small area methodology" at the 22nd Annual Conference of The International Environmetrics Society

January 2012: Snigdhansu Chatterjee gave a talk "Modeling climate characteristics using distribution-free small area methodology" at the Contemporary Issues and Applications of Statistics

December 2011: Snigdhansu Chatterjee gave a talk “Modeling climate characteristics using distribution-free small area methodology” at the Statistical Concepts and Methods for the Modern WorldDecember 2011: Auroop Ganguly gave an invited talk “Computational Data Sciences for Assessment and Prediction of Climate Extremes” at the AGU Fall Meeting

December 2011: Auroop Ganguly gave an invited talk “Precipitation Extremes with Climate Variability and Change” at the AGU Fall Meeting

November 2011: Karsten Steinhaeuser gave an invited talk “Knowledge Discovery with Networks for Climate Science” at the Climate Knowledge Discovery Workshop held at Supercomputing

November 2011: Arindam Banerjee gave an invited talk “Probabilistic Graphical Models for Climate Data Analysis” at the Climate Knowledge Discovery Workshop held at Supercomputi

November 2011: Vipin Kumar gave the keynote talk “Mining Scientific Data: Past, Present, and Future” at the IEEE International Conference on Tools with Artificial Intelligence

November 2011: Shashi Shekhar gave a talk “High-Performance Spatial Data Mining” at the Workshop on Challenges and opportunities in High-performance and Distributed GIS, ACM SIG-Spatial Intl. Conference on GIS

October 2011: Shashi Shekhar gave an invited talk “Geo-Social Computing” at the CDC/FDA Workshop on Food Safety Biosurveillance at Michigan State University

October 2011: Arindam Banerjee gave an invited talk at the Conference on Intelligent Data Understanding

September 2011: Michael Steinbach gave an invited talk “Understanding Climate Change: A Data Driven Approach” at the DIMACS sponsored workshop: US-China Collaborations in Computer Science and Sustainability

August 2011: Arindam Banerjee gave a talk “Machine Learning for Climate Sciences” at the 1st International Workshop on Climate Informatics

July 2011: Vipin Kumar gave a talk “Understanding Climate Change: Opportunities and Challenges for Data Intensive Computational Science” at the Center for Scalable Application Development Software Scientific Data and Analytics for Extreme Scale Computing Workshop

July 2011: Snigdhansu Chatterjee gave a talk “A statistical study of climate change: Analysis of temperature records of Arctic seawater data” at the Indian Statistical Institute

July 2011: Shashi Shekhar gave a keynote speech “Spatial and Spatio-temporal Data Mining” at the ISPRS Symposium on Spatial-Temporal Analysis and Data Mining

July 2011: Auroop Ganguly gave a presentation "Can machine learning help translate the science of climate change to information relevant for preparedness and policy?" at the Grand Challenges Workshop at the International Conference on Machine Learning

June 2011: Shashi Shekhar gave a talk “Spatial Data Mining and Sustainability” at the NSF Workshop on Information and Communication Technologies for Sustainability

June 2011: Vipin Kumar presented an invited talk at NASA's annual Earth Science Technology Forum (ESTF) in the special session on Revolutionary Information Systems Technology & the Impact on NASA Earth Science

April 2011: Shashi Shekhar gave a talk “What is Special about Mining Spatial Data” at the Computer Science Colloquium at Montana State University

April 2011: Vipin Kumar gave an invited keynote talk “Understanding Climate Change: Opportunities and Challenges for Computer Science” at the annual Information and Computing Technology (ICT) Conference

March 2011: Karsten Steinhaeuser gave an invited talk at the Climate Knowledge Discovery Workshop held at the DKRZ (German Climate Computing Centre)

March 2011: Vipin Kumar gave an invited talk “Discovery of Patterns in Global Earth Science Data using Data Mining” for the Distinguished Lecture Series at the College of Computing and Informatics

March 2011: Vipin Kumar gave an invited plenary talk “Discovery of Patterns in Global Earth Science Data using Data Mining” at the SIAM Conference on Computational Science and Engineering

February 2011: Shashi Shekhar gave a talk "Geo-Social Media Revolution: Hype or Reality?" at the annual retreat of the Integrated Media Systems Center (an NSF ERC) at the University of Southern California

February 2011: Alok Choudhary gave a talk "Exascale Supercomputing: Challenges and Approaches to Data Intensive Computing and Analytics" at University of Tsukba, RIKEN, AICS, Kobe Supercomputing Center, the Tokyo Supercomputing Center, the University of Tokyo, and the Tokyo Institute of Technology Matsuoka Lab

February 2011: Alok Choudhary gave a talk "High-Performance Data Mining: An Essential Paradigm for Knowledge Discovery" at AIST, Tsukuba, and Tokyo

December 2010: Vipin Kumar gave an invited keynote talk "Monitoring of Changes in the Global Forest Cover Using Data Mining" at the 2010 International Conference on Electronic-Business Intelligence

December 2010: Vipin Kumar gave an invited keynote talk "Discovery of Patterns in Global Earth Science Data using Data Mining" at the ICDM Workshop on Spatial and Spatiotemporal Data Mining

December 2010: Alok Choudhary gave a talk "High-Performance Data Mining: An Essential Paradigm for Knowledge Discovery" at the University of Florida's Barr Systems Distinguished Lecture Series

November 2010: Shashi Shekhar gave a keynote speech "Spatial Data Mining" at the Annual GIS Day celebration at the University of Notre Dame

November 2010: Alok Choudhary gave a keynote talk "High-Performance Data Mining: An Essential Paradigm for Knowledge Discovery" at the International Conference on Multimedia Information Networking and Security

November 2010: Alok Choudhary gave a keynote talk "High-Performance Data Mining: An Essential Paradigm for Knowledge Discovery" at the 2010 Workshop Multimedia Technologies

October 2010: Vipin Kumar gave an invited talk at Draper Laboratory's Global Climate Monitoring Symposium

October 2010: Vipin Kumar gave an invited keynote talk "Monitoring of the Changes in the Global Forest Cover Using Data Mining" at NASA's Conference on Intelligent Data Understanding

September 2010: Alok Choudhary gave a lecture "High-Performance Data Mining: An Essential Paradigm for Knowledge Discovery" at the Northwestern University Electrical Engineering and Computer Science Department's Meet the Faculty Lecture Series

Awards

A number of Expedition team members have received awards.

August 12, 2012

Vipin Kumar received the 2012 ACM SIGKDD Innovation Award for technical excellence and contributions that have had a lasting impact on the field of Knowledge Discovery and Data Mining. The award was presented during the opening plenary session of the 2012 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) in Beijing, China. His awards talk highlighted Expeditions related work.

June 20, 2012

Research on refining hurricane forecasts by Nagiza Samatova, Alok Choudhary, and colleagues was highlighted by the DOE ASCR Discovery web-magazine.

June 14, 2012

Research Associate Karsten Steinhaeuser was selected as one of 34 DIssertations Initiative for the advancement of Climate Change ReSearch (DISCCRS) scholars and will participate in the DISCCRS VII Symposium later this year. DISCCRS is a program co-sponsored by NSF and NASA with the goal of fostering interdisciplinary research and preparing early career scientists for interdisciplinary research careers, in particular in areas related to climate change.

April 26, 2012
PhD student Soumyadeep Chatterjee won the Best Student Paper Award at the SIAM International Conference on Data Mining (SDM) 2012 for his paper titled "Sparse Group Lasso: Consistency and Climate Applications" co-authored by Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee and Auroop Ganguly. The paper proves statistical consistency bounds for estimators with hierarchical sparse regularizors and shows utility of using such estimators for high dimensional regression problems in climate science.

February 25, 2012
PhD student Jaya Kawale won the Best Graduate Student Poster Award at the Midwest Women in Computing Celebration for her work on "A Graph Based Approach to Find Teleconnections in Climate Data." As a winner of this regional prize, Jaya will receive funding to attend the national 2012 Grace Hopper Celebration.

October 19, 2011
PhD student Jaya Kawale won the Best Student Paper award for her paper titled "Data Guided Discovery of Dynamic Climate Dipoles" at the NASA Conference on Intelligent Data Understanding (CIDU 2011). The paper was a joint work with authors from the University of Minnesota, North Carolina State University and Oak Ridge National Laboratory. The paper shows the utility of a graph-based approach to find teleconnections in climate data and uses the method as an alternate way for analyzing the performance of the various climate models used to study the global climate system.

September 19, 2011
Auroop Ganguly was nominated to the American Meteorological Society (AMS) Committee on Artificial Intelligence Applications to Environmental Science and was named an Associate Editor of the journal Water Resources Research published by the American Gephysical Union (AGU).

September 7, 2011
Shashi Shekhar was named a Resident Fellow of the Institute on the Environment, an award given annually to faculty who pursue interdisciplinary research, teaching and collaborations to provide environmental solutions.

August 18, 2011
PhD student Jaya Kawale won the 2011 Explorations in Science through Computation Student Award for her work on Discovering Teleconnections in Climate Data through Data Mining. The award will be presented at the 2011 Supercomputing Conference (SC11) awards ceremony.

April 27, 2011
Ph.D. student James Faghmous won a 2011 NSF Nordic Research Opportunity Grant to visit Norway to conduct research at one of the world's top climate research institutes, the Bjerknes Center for Climate Research (BCCR) in Bergen, Norway.

April 13, 2011
Nature Research Highlights "Climate Change: Cold Spells in a Warm World", featuring work done by Auroop Ganguly.

April 12 - 13, 2011
MSNBC, Yahoo News and Our Amazing Planet feature: "Chilling: Cold Snaps Will Persist in Warming World", featuring work done by Auroop Ganguly.

Tools & Software

Expedition's team members have also released open source software products to the scientific community at large. Northwestern has been actively developing Parallel NetCDF, a parallel implementation for reading and writing NetCDF files, which is the universally-accepted standard for storing and exchanging climate data. The latest version of PnetCDF (1.3.0) was released in June, and includes new features such as support for 64-bit data types and arrays with more than 4 billion elements, as well as a non-blocking I/O API. Northwestern has also released shared memory (OpenMP) parallel implementations of density-based and hierarchical clustering algorithms (DBSCAN and SLINK, respectively). All three of these software products are available on the Northwestern website at http://cucis.ece.northwestern.edu.

The following software libraries and tools that have been used both within the Expedition project and have also been demonstrated to have a broader impact for other application domains such as those in biology. Some libraries are still in the final testing stages for the production release but are available upon request:

- **(α,β)-motif finder**: Tool for identification of functional modules (e.g., dense subgraphs) that are predicted to be related to the target phase of the system (e.g., low or high hurricane activity seasons) (http://freescience.org/cs/ABClique/)
- **MCE-parallel**: A scalable, parallel algorithm for the NP-hard clique enumeration problem (http://freescience.org/cs/parallelclique/)
- **Bi-clustering**: Software for identification of system's phase-related modules and their putative cross-talks (http://freescience.org/cs/bi-clustering/)
- **DENSE** (Dense ENriched Subgraph Enumeration): Tool for identifying dense subgraphs that are ``enriched" by query nodes representing knowledge *priors* (e.g., locations for known climate indices) (http://freescience.org/cs/DENSE/)
- **SPICE** (System Phenotype-related Interplaying Components Enumerator: Iterative enumerator of statistically significant and system's phase-relevant subsystems and can be applied to both network data and vector data (http://freescience.org/cs/SPICE)
- **ISABELA** (In-situ Sort-And-B-spline Error-bounded Lossy Abatement), Library for in-situ, embarassingly parallel B-spline based lossy compression of scientific floating point data with user-controlled accuracy bounds
- **ISABELA-QA**: A memory and storage light-weight parallel query processing engine over ISABELA-compressed scientific data capable of multi-core, multi-node, GPU executions
- **ALACRI^{2}TY** (Analytics-driven Lossless dAta Compression for Rapid In-situ Indexing, sToring, and querYing Fast and memory light-weight query processing (and lossless compression) engine for scientific floating point data that is optimized for heterogeneous access pattern
- **ISOBAR** (In-Situ Orthogonal Byte Aggregate Reduction (ISOBAR) Compression: A preconditioner-based, high-throughput lossless compression technique for hard-to-compress scientific datasets (http://freescience.org/cs/ISOBAR)
- **Prm_causality**: A tool for data-driven, semi-automatic inference of plausible phenomenological models (http://freescience.org/cs/prm_causality/)
- **FORECASTER** (Forecast Oriented Feature Elimination-based Classification of Adverse Spatio-Temporal Extremes): A forecast-oriented feature elimination-based ensemble of classifiers for robust forecasting of adverse spatio-temporal extremes).
- **DETECTOR** (Forecast Error Detection and Correction): A library for detecting and correcting prediction errors in extreme event forecasts

The project's novel data-driven approaches already promise to excel beyond the traditional methods in climate prediction tools and search for fundamental inter-relationships in the climate system in a significant way. The Expedition project has a focus on creating bridges and special partnerships with

research and operational climate communities to maximize the benefits. One example is an ongoing effort to create a hybrid model for hurricane landfall prediction, which aims to combine physical models with machine learning techniques to achieve improved performance. Considering the ground-breaking nature of Expedition's data-driven discoveries we have adopted a four component-interface framework to engage potential user communities. This section also discusses several pathways for expansion of our outreach activities.

This interface was inspired by the ongoing NCAR-NCSU Google funded Meningitis project. Meningococcal meningitis is one of the deadliest and most terrifying illnesses in Africa. Because of the relative scarcity of the currently available vaccine and its limited efficacy, mass vaccinations in a region are initiated only when an epidemic is already underway in the region or a neighboring region. Bacterial meningitis epidemics occur regularly in the "meningitis belt" of sub-Saharan Africa extending from Gambia to Ethiopia, but the prediction of their location, frequency, and duration has been elusive. Meningitis is a serious infectious disease affecting 21 countries across Africa. It kills hundreds of thousands of people in one year. Three hundred million people are at risk, with 700,000 cases in the past 10 years and 10-50 % fatality rate. The UCAR-NCSU-Google project focuses on the short-term forecasts of one to two weeks to guide the vaccination activities has made significant contribution to strategies for dispensing the vaccine which is in short supplies.

Here we describe one of the flagship projects of the Expedition outreach efforts. It contributes to the urgent need to address a major societal problem in the health sector. We have developed a new climate index for the prediction of the epidemic levels of the meningitis decease. The new index is predicted with high skill thus paving way for more effective management of meningitis outbreaks through a more effective vaccination campaign guided by climate information. This outreach component of the Expedition subproject builds on the break-through research by Samatova and Semazzi's joint research program in the application of supervised machine learning methodology in computer science for the prediction of the Sahelian climate during the meningitis season, and integrates the resulting improvements for making optimal and high impact decisions into the campaign for the control of meningitis epidemics in Sahelian Africa. The most robust and useful weather/meningitis relationship comes from the strong correlation between the start of the rainy season and the abrupt decline in the transmission of the disease. The actionable climate information is that if the onset of the high relative humidity season could be forecast at the district-level scale (at which public health managers make decisions), officials could prioritize vaccine deployment to areas of predicted continued low humidity where meningitis transmission will persist. The prediction on seasonal time scales will be used to optimize the vaccination campaign by improving the logistical considerations for the vaccination activities including distribution of hardware assets, e.g. vehicles, vaccination supplies for the campaign and human capital. Through collaboration with NCAR, Kwame Nkrumah University of Science and Technology (KNUST), the African Center for Meteorological Applications for Development (ACMAD), and public health organizations in West Africa seeking to optimize the effectiveness of the vaccination campaign, this subproject will help to create the operational seasonal prediction products for the regional operation center of ACMAD to serve the meningitis vaccination over the Sahelian region. The Expedition mission is to conduct basic science, not to build capacity. The project will achieve its outreach objectives by using a common data format with ACMAD via a Toolkit Portal that is common with ACMAD to ensure seamless exchange of data and information.

Other outreach mechanisms

In addition to our strategy to support the operationalization of our research outcomes, the following methods will also be used to support outreach activities, (i) journal publications for the research community and other publications generated by the project since its inception, (ii) through national and international partnerships (Table 1).

Name of Outreach Partner	**Type of Enabled Outreach**
National Centre for Atmospheric Research (**NCAR**)	NSF Expedition project rainfall and humidity prediction for the NCAR/Google project on the vaccination of meningitis epidemics which severely impacts up to *250,000 people* annually; next frontier *data analytics and reduction involving massive PB HPC problems*
United Nations (UN) World Meteorological Organization **(WMO**)	Operationalization of NSF Expedition project experimental seasonal climate prediction methodology for WMO *African Centre of Meteorological Applications for Development* (ACMAD) and other WMO Regional Climate Centers (RCCs) serving over *3 billion people*.
National Hurricane Centre **(NHC)**	Operationalization of NSF Expedition project experimental hurricane prediction (*prospects for lead time greater than ten days before landfall)*-methodology for NHC to reduce vulnerability of US coastal population; efficient *hurricane intensity* estimation methods - *50% improvement*
National Climate Data Centre **(NCDC)**	Creation of *new climate indices* and introduction of NSF Expedition data driven methodologies into the NCDC's analysis tool-kit of the highly multi-dimensional and most voluminous open source global climate data archive; highly efficient graphical methods for *abrupt climate detection*
World Climate Research Program **(WCRP)**	Fundamental research on *understanding of casual pathways of the Sahelian climate predictability and causality, change detection of Atlantic hurricanes*
Intergovernmental Panel on Climate Change **(IPCC)**	Trends in rainfall *extremes* over India during last half-century; causality of the Sahelian climate variability; change detection of Atlantic hurricanes
North Carolina State Univ. (**NCSU**) & Univ. of Minnesota (**UMN**)	Infusion of Expedition discoveries into university climate science curriculum; *NCSU Climate PSM program & UMN Climate Statistics advanced course (http://climate-psm.meas.ncsu.edu/)*

Table 1: Examples of Future Expedition Outreach Activities

Refereed Journal and Conference Publications

Accepted/Published: (ordered by date)

J.F. Knight, B.P. Tolcser, "Remote classification of wetlands using decision trees", Photogrammetric Engineering and Remote Sensing (2012), accepted.

S. Lakshminarasimhan, N. Shah, S. Ethier, S.-H. Ku, C.S. Chang, S. Klasky, R. Latham, R. Ross, N.F. Samatova, "ISABELA: Effective in-situ compression of spatio-temporal data", Concurrency and Computation: Practice and Experience Journal (2012), accepted.

M.T Dugda, A. T. Workineh, A. Homaifar and J.H. Kim, "Receiver Function Inversion Using Genetic Algorithms", Bulletin of the Seismological Society of America (2012), accepted.

Z. Li, P. Qiu, S. Chatterjee, Z. Wang. "Using p-values to design statistical process control charts", Statistical Papers (2012), accepted.

C. Monteleoni, G.A. Schmidt, F. Alexander, A. Niculescu-Mizil, K. Steinhaeuser, M. Tippett, A. Banerjee, M.B. Blumenthal, A.R. Ganguly, J.E. Smerdon, M. Tedesco. "Climate Informatics", in T. Yu, N. Chawla, S. Simoff (Eds.), Computational Intelligent Data Analysis for Sustainable Development, CRC Press (2012), accepted.

A. Ganguly, E, Kodra, S. Chatterjee, A. Banerjee, and Habib Najm, "Computational data sciences for actionable insights on climate extremes and uncertainty," in T. Yu, N. Chawla, S. Simoff (Eds.), Computational Intelligent Data Analysis for Sustainable Development, CRC Press (2012), accepted.

J. Jenkins, E. Schendel, S. Lakshminarasimhan, D.A. Boyuka III, T. Rogers, S. Ethier, R. Ross, S. Klasky, N.F. Samatova, "Byte-precision Level of Detail Processing for Variable Precision Analysis", Supercomputing [SC], Salt Lake City, Utah, USA (Nov 2012).

W. Hendrix, M. Patwary, A. Agrawal, W. Liao, A. Choudhary, "Parallel Hierarchical Clustering on Shared Memory Platforms", High Performance Computing Conference (HiPC), Pune, India (December 2012).

M. Patwary, D. Palsetia, A. Agrawal, W. Liao, F. Manne, and A. Choudhary, "A New Scalable Parallel DBSCAN Algorithm Using the Disjoint-Set Data Structure", Supercomputing [SC], Salt Lake City, UT, USA (Nov 2012).

P. Mohan, X. Zhou, S. Shekhar, "Quantifying Resolution sensitivity of spatial autocorrelation: A Resolution Correlogram approach", In Proc. of International Conference on Geographic Information Science [GIScience], Columbus, OH, USA (Sept 2012).

Z. Gong, T. Rogers, J. Jenkins, H. Kolla, S. Ethier, J. Chen, R. Ross, S. Klasky, N.F. Samatova, "MLOC: Multi-level Layout Optimization Framework for Compressed Scientific Data Exploration with Heterogeneous Access Patterns", Proc.41st International Conference on Parallel Processing [ICPP], Pittsburgh, PA (Sept 2012).

J. Jenkins, I. Arkatkar, S. Lakshminarasimhan, N. Shah, E.R. Schendel, S.Ethier, C.S. Chang, J. Chen, H. Kolla, S. Klasky, R. Ross, N. F. Samatova. "Analytics-driven Lossless Data Compression for Rapid In-situ Indexing, Storing, and Querying", 23rd International Conference on Database and Expert Systems Applications [DEXA], Vienna, Austria (Sept 2012).

N. Shah, E.R. Schendel, S. Lakshminarasimhan, S.V. Pendse, T. Rogers, N.F. Samatova, "Improving I/O Throughput with PRIMACY: Preconditioning ID-Mapper for Compressing Incompressibility", IEEE International Conference on Cluster Computing [Cluster], Beijing, China (Sept 2012).

J. Jenkins, J. Dinan, P. Balaji, N.F. Samatova, R. Thakur, "Enabling Fast, Noncontiguous GPU Data Movement in Hybrid MPI+GPU Environments", IEEE International Conference on Cluster Computing [Cluster], Beijing, China (Sept 2012).

J. Kawale, S. Chatterjee, D. Ormsby, K. Steinhaeuser, S. Liess, V. Kumar. "Testing the Significance of Spatio-temporal Teleconnection Patterns." ACM SIGKDD Conference on Knowledge Discovery and Data Mining [KDD], Beijing, China (Aug 2012).

D. Das, A. Ganguly, Z. Obradovic, A. Banerjee, "Towards understanding dominant processes in complex dynamical systems: Case of precipitation extremes", ACM SIGKDD Workshop on Knowledge Discovery from Sensor Data [SensorKDD], Beijing, China (Aug 2012).

D. Das, E. Kodra, A. Ganguly, Z. Obradovic, "Mining Extreme Values: Climate and Natural Hazard", ACM SIGKDD Workshop on Data Mining Applications in Sustinability [SustKDD], Beijing, China (Aug 2012).

D. Das, A. Ganguly, S. Chatterjee, V. Kumar, Z. Obradovic, "Spatially Penalized Regression for Dependence Analysis and Prediction of Rare Events: A Case for Precipitation Extremes", ACM SIGKDD Workshop on Data Mining Applications in Sustinability [SustKDD], Beijing, China (Aug 2012).

J. H. Faghmous, Y. Chamber, F. Vikeboe, S. Boriah, M. d. S. Mesquita, S. Liess, V. Kumar, "Novel and scalable spatio-temporal technique for ocean eddy monitoring", 26th Conference on Artificial Intelligence [AAAI], Toronto, Canada (July 2012).

H. Shan, J. Kattge, P. B. Reich, A. Banerjee, F. Schrodt, and M. Reichstein, "Gap Filling in the Plant Kingdom---Trait Prediction using Hierarchical Probabilistic Matrix Factorization," International Conference on Machine Learning [ICML], Edinburgh, Scotland (June 2012).

E.R. Schendel, S.V. Pendse, J. Jenkins, D.A. Boyuka II, Z. Gong, S. Lakshminarasimhan, Q. Liu, S. Klasky, R. Ross, N.F. Samatova, "ISOBAR hybrid compression-I/O interleaving for large-scale parallel I/O optimization", The 21st International ACM Symposium on High-Performance Parallel and Distributed Computing [HPDC], Delft, the Netherlands (June 2012).

S. Lakshminarasimhan, P. Kumar, W. Liao, A. Choudhary, V. Kumar, N.F. Samatova, "On the Path to Sustainable, Scalable, and Energy-efficient Data Analytics: Challenges, Promises, and Future Directions", 2012 International Green Computing Conference [IGCC], San Jose, CA, USA (June 2012).

Z. Gong, S. Lakshminarasimhan, J. Jenkins, H. Kolla, S. Ethier, J. Chen, R. Ross, S. Klasky, N.F. Samatova, "Multi-level layout optimization for efficient spatio-temporal queries on ISABELA-compressed data", The 26th IEEE International Parallel and Distributed Processing Symposium [IPDPS], Shanghai, China (May 2012).

E. Parish, E. Kodra, K. Steinhaeuser, A.R. Ganguly, "Estimating future global per capita water availability based on changes in climate and population", Computers & Geosciences, 42:79-86 (May 2012).

S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, A.R. Ganguly. "Sparse Group Lasso: Consistency and Climate Applications", SIAM International Conference on Data Mining [SDM], Anaheim, CA, USA (April 2012). *Best Student Paper Award*

Q. Fu, A. Banerjee, S. Liess, and P. Snyder. "Drought Detection for the Last Century: A MRF-based Approach", SIAM International Conference on Data Mining [SDM], Anaheim, CA, USA (April 2012).

S.V. Pendse, I. Tetteh, F. Semazzi, V. Kumar, N.F. Samatova, "Data-driven, semi-automatic inference of phenomenological physical models: Application to Eastern Sahel rainfall", SIAM International Conference on Data Mining [SDM], Anaheim, CA, USA (April 2012).

E.R. Schendel, Y. Jin, N. Shah, J. Chen, C.S. Chang, S.-H. Ku, S. Ethier, S. Klasky, R. Latham, R. Ross, N.F. Samatova, "ISOBAR preconditioner for effective and high-throughput lossless data compression", The 28th IEEE International Conference on Data Engineering [ICDE], Washington, DC (April 2012).

G. Fetanat, A. Homaifar, K. Knapp, "Tropical cyclone intensity estimation using temporal analysis and spatial features in satellite data," 30th Conference on Hurricanes and Tropical Meteorology, Ponte Vedra Beach, FL, USA (April 2012).

M. Gebril, R. Buaba, A. Homaifar, E. Kihn, "Satellite Imagery Retrieval: Features & Metrics Evaluation", IEEE Aerospace Conference, Big Sky, MT, USA (March 2012).

S. Ghosh, D. Das, S.-C. Kao, A.R. Ganguly, "Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes", Nature Climate Change, 2:86-91, (Feb 2012). doi:10.1038/nclimate1327.

E. Kodra, S. Ghosh, A.R. Ganguly, "Evaluation of global climate models for Indian monsoon climatology", Environmental Research Letters, 7:014012, (Feb 2012). doi:10.1088/1748-9326/7/1/014012.

Y. Jin, S. Lakshminarasimhan, N. Shah, Z. Gong, C.S. Chang, J. Chen, S. Ethier, H. Kolla, S.-H. Ku, S. Klasky, R. Latham, R. Ross, K. Schuchardt, N.F. Samatova, "S-preconditioner for multi-fold data reduction with guaranteed user-controlled accuracy", IEEE International Conference on Data Mining [ICDM], Vancouver, Canada (Dec 2011).

J. Kawale, S. Chatterjee, A. Kumar, S. Liess, M. Steinbach, V. Kumar, "Anomaly construction in climate data: issues and challenges", NASA Conference on Intelligent Data Understanding [CIDU], Mountain View, CA, USA (Nov 2011).

X. Zhou, S. Shekhar, P. Mohan, S. Liess, P. K. Snyder, "Discovering Interesting Sub-paths in Spatiotemporal Datasets: A Summary of Results", ACM SIGSPATIAL International Conference on Advances in Geographic Information System, Chicago, IL, USA, p. 44-53 (Nov 2011).

S. Lakshminarasimhan, J. Jenkins, I. Arkatkar, Z. Gong, H. Kolla, S.-H Ku, S. Ethier, J. Chen, C.S. Chang, S. Klasky, R. Latham, R. Ross, N.F. Samatova, "ISABELA-QA: Query-driven Analytics with ISABELA-compressed Extreme-Scale Scientific Data", Supercomputing [SC], Seattle, WA (Nov 2011).

J. Kawale, S. Liess, A. Kumar, M. Steinbach, A. Ganguly, N. F. Samatova, F. Semazzi, P. K. Snyder, V. Kumar, "Data guided discovery of dynamic climate dipoles", NASA Conference on Intelligent Data Understanding [CIDU], Mountain View, CA, USA (April 2011). *Best Student Paper Award*

Z. Chen, W. Hendrix, N.F. Samatova, "Community-based anomaly detection in evolutionary networks", Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies (Oct 2011). doi: 10.1007/s10844-011-0183-2

Publications in Year 1 of the Project (Sep 2010 - Aug 2011)

J. Jenkins, I. Arkatkar, J. D. Owens, A. Choudhary, N. F. Samatova, " Lessons Learned from Exploring the Backtracking Paradigm on the GPU", 17th International European Conference on Parallel and Distributed Computing [Euro-Par], Bordeaux, France (2011).

S. Lakshminarasimhan, N. Shah, S. Ethier, S. Klasky, R. Latham, R. Ross and N. F. Samatova, "Compressing the Incompressible with ISABELA: In-situ Reduction of Spatio-Temporal Data", Proc. 17th International European Conference on Parallel and Distributed Computing [Euro-Par] (2011).

T. Pansombut, W. Hendrix, Z. J. Gao, B. E. Harrison, N. F. Samatova, "Biclustering-Driven Ensemble of Bayesian Belief Network Classifiers for Underdetermined Problems", Proc. of the 22nd International Joint Conference on Artificial Intelligence [IJCAI] (2011).

H. Sencan, Z. Chen, W. Hendrix, T. Pansombut, F. Semazzi, A. Choudhary, V. Kumar, A.V. Melechko, N. F. Samatova, "Classification of Emerging Extreme Event Tracks in Multi-Variate Spatio-Temporal Physical Systems Using Dynamic Network Structures: Application to Hurricane Track Prediction", Proceedings of the 22nd International Joint Conference on Artificial Intelligence [IJCAI] (2011).

E. Kodra, S. Chatterjee, A.R. Ganguly, "Challenges and opportunities toward improved data-guided handling of global climate model ensembles for regional climate change assessments", ICML Workshop on Machine Learning for Global Challenges (2011).

A. Agovic, A. Banerjee, S. Chatterjee, "Probabilistic matrix addition," Proc. 28th Int'l Conference on Machine Learning [ICML] (2011).

S. Lakshminarasimhan, J. Jenkins, I. Arkatkar, Z. Gong, H. Kolla, S-H Ku, S. Ethier, J. Chen, CS Chang, S. Klasky, R. Latham, R. Ross and N. F. Samatova, "ISABELA-QA: Query-driven Data Analytics over ISABELA-compressed Scientific Data", Supercomputing (2011).

J. Kawale, M. Steinbach, V. Kumar, "Discovering Dynamic Dipoles in Climate Data", SIAM International Conference on Data Mining [SDM] (2011).

N. Mukhopadhyay, S. Chatterjee, "High dimensional data analysis using multivariate generalized spatial quantiles", Journal of Multivariate Analysis, vol.102, no. 4, p. 768-780 (2011).

Z. Chen, K.A. Wilson, Y. Jin, W. Hendrix, N.F. Samatova, "Detecting and Tracking Community Dynamics in Evolutionary Networks", IEEE ICDM, SIASP Workshop, p. 318-327 (2010).

Submitted:

R. Buaba, A. Homaifar, E. Kihn, "Optimal Parameterization for Exact Euclidean Locality Sensitive Hashing for Fast Running time", Elsevier, International Journal of Approximate Reasoning, submitted.

R. Buaba, A. Homaifar, E. Kihn, "Optimal Load Factor for Exact Euclidean Locality Sensitive Hashing to Guarantee Low Memory Space", Elsevier, International Journal of Approximate Reasoning, submitted.

Z. Chen, W. Hendrix, G. Han, I.K. Tetteh, A. Choudhary, F.H.M. Semazzi, N.F. Samatova, "Detecting Predictive and Physically Interpretable Communities in Contrast Groups of Networks: Application to Adverse Spatio-Temporal Extremes", Data Mining and Knowledge Discovery Journal, (2nd Revision), submitted.

W. Hendrix, M. Patwary, A. Agrawal, W. Liao, A. Choudhary, "Parallel Hierarchical Clustering on Shared Memory Platforms", High Performance Computing Conference (HiPC), submitted.

Z. Jiang, S. Shekhar, P. Mohan, J.Knight, J. Corcoran, "Learning Spatial Decision Tree for Geographical Classification: A Summary of Results", ACM SIGSPATIAL GIS, submitted.

J. Kawale, S. Liess, A. Kumar, M. Steinbach, A. Ganguly, N. F. Samatova, F. Semazzi, P. K. Snyder, V. Kumar, "A graph based approach to find teleconnections in climate data", Statistical Analysis and Data Mining Journal, submitted.

C. Lacewell, A. Homaifar, Y.L. Lin, "Tracing the Origins and Propagation of Pre-Tropical Storm Debby Mesoscale Convective Systems using Pattern recognition and Image Fusion", Meteorology and Atmospheric Physics, submitted.

E.M. Rangel, W. Hendrix, A. Agrawal, W. Liao, A. Choudhary, "Stochastically Emerging Medoids: Application of Classical Problems in Probability Theory for Clustering Massive Data Sets", Neural Information Processing Systems Conference [NIPS], submitted.

Web/Internet Site

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

Contributions

Contributions within Discipline

We have developed numerous data mining and machine learning methods for the analysis of climate data. Specifically, these include methodological advances such as novel algorithms for change detection, statistical models for identifying dependency structures, and graph-based approaches for representing and analyzing complex systems; and computational advances such as scalable implementations of important data mining kernels for high-performance systems (e.g., multi-cores, GPU processors) and compression techniques for spatio-temporal data. The Section on 'Research Activities and Findings' provides a more technical discussion of the individual contributions.

Contributions to Science and Engineering

This project is one example of the application of the Fourth Paradigm, that is, the advancement of a scientific discipline through data-driven discovery. Although our research is specifically focused on developing an improved understanding of climate processes, the techniques and tools developed are applicable to other domains that share similar data characteristics, e.g., multi-dimensional (space-time), high degree of complexity, etc. We will make the tools and datasets resulting from this project available on our project website and will continue to build and strengthen collaborations with researchers in other disciplines.

Contributions to Human Resources Development

Primary contributions to date have involved the education and training of researchers in the areas of data mining and climate-related sciences. As discussed in the Section on 'Educational Activities', the project members involved have expanded their knowledge, improved their research skills and benefited from mentoring opportunities. The activities of the project are catalyzing higher levels of engagement and interactions at (and among) the partner institutions through interdisciplinary collaborations spanning computational and climate sciences. WORKSHOP ATTENDANCE

Contributions to Resources for Research and Education

The engagement of faculty and students – especially from underrepresented groups in climate-related disciplines – is creating a new pool of experts who are globally-active while building institutional capacity for international engagement. Training of multidisciplinary scientists who are aware of impactful applications such as climate change science and extreme climate events prepares the next generation of teachers and researchers capable of solving grand challenge problems for our society. In order to disseminate our findings and generate interest in the broader community, team members are integrating research results into their courses.

Contributions beyond Science and Engineering

Our methods for supervised and semi-supervised learning developed through this project have advanced estimation and prediction skill for mean processes and extreme events in climate, but are also being tested for robustness in other application domains. Moreover, the results of our research will also be relevant to other organizations and agencies outside of the sciences and engineering domain. For example, the research being conducted has shown great promise for enhancing skill in the prediction of hurricane activity over the Atlantic and Western Pacific Ocean basins. Our strategic collaboration with the National Climatic Data Center (NCDC) and other outreach institutions has the potential to inspire paradigm shifts in the use of hurricane prediction information to benefit the insurance industry, FEMA, and other climate sensitive risk management sectors. We expect these and similar overarching contributions to materialize over the remainder of the project's life.

Research Activities and Findings

Annual Report for Period: 09/2011 - 08/2012 **Submitted on:** 08-31-2012
Principal Investigator: Kumar, Vipin **Award ID:** 1029771
Organization: University of Minnesota

Title: Collaborative Research: Understanding Climate Change: A Data Driven Approach

Objectives of the Project

Climate change is the defining environmental challenge now facing our planet. Whether it is rising temperatures, the increasing frequency or intensity of hurricanes, extreme droughts and floods, or rising sea levels, the social, economic and environmental consequences are great as the resource-stressed planet nears 9 billion inhabitants by mid-century. Yet there is considerable uncertainty regarding the social and environmental impacts due to the limited capabilities of existing physics-based models of the Earth system. Consequently, important questions relating to food security, water resources, biodiversity, and other socio-economic issues over relevant spatial and temporal scales remain unresolved. Therefore, a new and transformative approach is required to better understand the climate system and the potential impacts of climate change.

Data-driven approaches that have been highly successful in other scientific disciplines hold significant potential for application in environmental sciences. This Expeditions project aims to address key challenges in understanding climate change and impacts by developing methods that leverage the abundance of climate and ecological data available from satellite and ground-based sensors, the observational record for atmospheric, oceanic, and terrestrial processes, and physics-based climate model simulations. To realize this ambitious goal, novel methodologies appropriate to climate change science are being developed.

These innovative approaches will help provide new understanding of the complex nature of the Earth system and the mechanisms contributing to the adverse consequences of climate change, such as increased frequency and intensity of hurricanes, precipitation regime shifts, and the propensity for extreme weather events that result in environmental and socioeconomic disasters. Methodologies developed as part of this project will be used to advance scientific knowledge, to gain actionable insights, and to inform policymakers.

The following highlights some of the key activities and findings.

Table of Contents

Contributors are listed for each task with one of the following codes indicating level of involvement and affiliation:
F – Faculty R – Research Associate P – Postdoctoral Fellow G – Graduate Student U – Undergraduate Student
C – Collaborator

1. PREDICTIVE MODELING

Climate data seriously challenges the state-of-the-art in predictive modeling. The challenge comes from both the nature of the data itself as well as the nature of problems that need to be solved. As a result, there are arguably no off-the-shelf predictive modeling methods which can even begin to meaningfully analyze climate data and make predictions based on the same. In particular, the long range spatial and temporal dependencies in climate data cannot be effectively captured by Markov models with local dependencies that are often used in domains such as image, speech, video, and signal analysis. In this section we describe some of our work to extend the state-of-the-art in predictive modeling for climate. The need for predictive insights spans a broad range of climate challenges and phenomena, although hurricanes are a key focus.

Accomplishment Highlights:

One research effort has pursued both applications and theoretical advances in sparsity regularized regression techniques (Lasso, Elastic Net, Group Lasso, and Sparse Group Lasso) to obtain a predictive understanding of complex, dynamic physical phenomena, such as regional precipitation or hurricane intensity and frequency. We have established rates of convergence for sparsity inducing hierarchically regularized regression, which includes Lasso, Group Lasso, and Sparse Group Lasso as special cases. The rate depends only logarithmically in the number of dimensions and hence the methods are applicable in the high-dimensional low-sample regime, which is common in climate sciences. In addition to improvements in prediction accuracy, such techniques have helped identify a previously undiscovered source of Atlantic hurricane inter-annual variability in the Somali region of East Africa. We also used sparsity regularized regression as the basis for a data-driven, semi-automatic approach that produced a plausible phenomenological model of the eastern Sahel seasonal rainfall and quantified key climate drivers of rainfall variability. As a part of our sparsity regularized regression research, we have proposed methods for optimizing the regularization penalty for Lasso regression, detecting and ranking prominent temporal phases for climate variables, and assessing predictor statistical significance to ensure the validity of the discovered causal relationships.

In work that accounts for the hierarchical system-subsystem structure of real-world dynamic systems (e.g., atmosphere-ocean systems), we have developed predictive approaches that take this hierarchical structure into account to improve classification and regression based forecasting of rainfall and hurricane activity. As part of that effort, we developed DETECTOR, a hierarchical method for detecting and correcting prediction errors in extreme event forecasts by employing the whole–part relationships between different systems. We also created FORECASTER, an algorithm that constructs a forecast-oriented, feature elimination-based ensemble of classifiers for robust forecasting of extreme events (e.g., low, normal, or high hurricane activity season). In further work, related to regression, rather than classification, we developed a means to use classification results from FORECASTER methodology to train individual regression models for the subsets of data that belong to distinct classes. Both approaches have yielded significant improvements in prediction performance. In another project concerned with forecasting hurricane intensity, we have developed a new predictive technique for cyclone intensity that uses historical intensity information of the 10 closest similar cyclones (determined by a K-nearest-neighbor algorithm) in the historical database. Overall, 30% to 55% improvement has been achieved compared to the current state of the art.

As part of an ongoing effort to develop predictive approaches to be used for data with spatial autocorrelation, we have created a spatial tree approach that can be used for spatial data. To this end, we created the spatial decision tree (SDT) model, including a spatial information gain "interestingness" measure. We used this model to develop an SDT learning algorithm, and have conducted a case study on a real-world remote sensing dataset to validate the usefulness of the proposed approach.

In a project that addresses the evaluation of predictive models, we are working on an approach to use the WRF Regional Model to extensively assess the relative prediction skill of supervised machine learning methods with respect to hurricane activity over eastern Africa and the Indian Ocean, as well as the Atlantic. We have already used this tool and are constructing a tropical cyclone categorical stratification for input data for the machine learning prediction models.

Individual Project Reports:

1.1 High-Dimensional Statistical Dependencies

Contributors: Banerjee (F-UMN), Chatterjee (F-UMN), Chatterjee (G-UMN), Ganguly (F-NE), Pearson (U-UMN), Steinhaeuser (R-UMN)

Activities:

Understanding complex dependencies between numerous climate covariates in an important step in descriptive analysis, predictive modeling, and statistical downscaling. The NCEP Reanalysis project has provided us with spatiotemporal reanalysis data for multiple variables over the past 60 years. Moreover, the NARR and the NOAA CPC high-resolution climate datasets over the US provide accurate measurements for the last three decades. There has been some work done in the past few years to study the complex dependencies between variables. However, although different correlation measures have been used to analyze dependencies and various descriptive models have been proposed, they cannot yet provide a framework for constructing reliable predictive models. In order to do so, we define 'dependence' between a pair of variables to denote conditional dependence between them, given all other variables under consideration. With this definition, under certain statistical assumptions, a predictive model for a climate variable can be built from the conditional probability distribution of the variable. The key challenge in applying such an approach to prediction problems in climate science is that the dimensionality, i.e., number of possible features or factors potentially affecting a response variable, is usually much larger than the number of samples, and certain classical methods for regression, such as ordinary least squares, cannot be applied in such a setting. New methods of sparsity regularized regression were investigated which can be applied to high-dimensional problems in the small sample regime.

Findings:

Predictive modeling of a climate variable, such as temperature in Peru or precipitation in India, can be formulated as one of estimating the conditional distribution of the variable of interest, given all other variables. Using standard methods, the estimation can be cast as a sparsity regularized regression problem. Motivated by the fact that only a small number of spatial locations will be relevant for the prediction, and only a small number of variables in such locations will be relevant, we consider different structured sparsity regularizers, such as Lasso, Group Lasso, and Sparse Group Lasso. The above mentioned sparsity regularizers are special cases of the more general framework of tree-structured hierarchical regularizers. Under mild statistical assumptions, we prove consistency guarantees for tree-structured norm regularized estimation methods. The key finding of the analysis is that under suitable assumptions on the design matrix, estimators with such hierarchical regularizers have a $O(\log p/n)$ rate of convergence. In other words, the logarithmic dependency on the dimensionality implies that such methods can be used in the high dimensional low sample regime, as is common in climate science.

In order to empirically test the performance of our methods we considered the problem of predicting 2 climate variables (temperature and precipitation) over 9 key land locations, using the knowledge of 6 variables over oceans. A recently published work tried to attack the same problem using a different climate network based formulation, and we compared our results to those reported in the article. We found an improvement of $10\% - 20\%$ in prediction accuracy by our models over the existing method. The proposed method also outperforms standard OLS statisticaly significantly on all 18 problems. Further investigation also revealed that for predictive modeling, covariates in the spatial neighborhood of the

response variable have significant predictive information, while those at distant locations have almost no contribution. Ongoing work is extending the framework to a non-parametric setting for modeling both normal and extreme precipitation in the US, as well as applications in statistical downscaling. We have also applied related techniques for descriptive dependency modeling of multivariate, time series, and possibly dependent data, as discussed in Section 2.4.

1.2 Inference of Phenomenological Physical Models

Contributors: Bello (G-NCSU), Gonzalez (G-NCSU), Harlalka (G-NCSU), Kumar (F-UMN), Pendse (G-NCSU), Samatova (F-NCSU), Semazzi (F-NCSU), Tetteh (G-NCSU), Waniha (P-NCSU)

Activities:

First-principles based predictive understanding of complex, dynamic physical phenomena, such as regional precipitation or hurricane intensity and frequency, is quite limited due to the lack of complete phenomenological models underlying their physics. *This research has proposed a data-driven, semi-automatic approach for causal inference of plausible phenomenological models and applying this approach to:*

a) Understanding the sources of variability of the Sahelian rainfall climate system. Sahel rainfall is an important factor for socio-economic growth and development of this region. Our goal is to use penalized regression methods, particularly the Lasso, to discover cause-effect relationships and thereby quantify the compound effects on rainfall that result from the complex interplay among the key climate variables. As a part of this research, we have proposed methods for optimizing the regularization penalty for Lasso regression, detecting and ranking prominent temporal phases for climate variables, assessing predictor statistical significance to ensure the validity of the discovered causal relationships, and quantifying the impact of local and global climate indices on the Sahelian rainfall. We have also performed a cumulative impact analysis (via the ECI model) of data normalization on model inference.

b) Quantifying the influence of East African climatic conditions in modulating Atlantic hurricane variability. An important source of African Easterly Waves, which constitute one of the largest contributors to tropical cyclogenesis in the Atlantic basin, was recently discovered in the Ethiopian Highlands, in close vicinity to the Indian Ocean. Yet none of the known sources of variability of Atlantic hurricane activity, such as NAO, AO, ENSO, Atlantic Dipole and AMO, can be directly attributed to the Indian Ocean and the adjacent landmass. In this study, we identify a previously undiscovered source of Atlantic hurricane interannual variability in the Somali region of East Africa. Lagged regression with elastic net regularization was used to infer a causal pathway from an SST anomaly source in the Indo-Pacific Ocean to this region. The statistical significance of this causal relationship was estimated using Monte Carlo *p*-value estimation method and part correlation as the test statistic.

Findings:

The culmination of this study is a plausible phenomenological model of the eastern Sahel seasonal rainfall and quantified key climate drivers involved in the rainfall variability at different time lags, as shown in the Figure 1.

This has given us the opportunity for further exploration and elaboration of our proposed hypotheses for the plausible physical processes, especially, one of the most important observations about the antagonistic role generally existing between NAO and the Pacific ENSO-related phenomena. The prominent temporal phases estimated using the completely unsupervised ranking methods have been found to be consistent with the prior knowledge in climatology. Moreover, we have also discovered prominent phases in cases where there is a lack of conclusive domain knowledge. Our study of using different forms of data normalization as a preprocessing step has revealed interesting relationships between climate variables on changing time scales. The analysis of plausible physical mechanisms revealed by each experiment suggests that depending on ambient climatic conditions and forcing factors,

some climate drivers may switch their roles, from enhancement to antagonism and vice versa, or to total dissipation, depending on their phases. This approach has helped us in understanding the changing relationships between the climate variables on monthly and seasonal timescales. One of the important conclusions is that monthly climatic influences persist through seasons, because of longer memories, and thus, to a large extent, are dynamically inseparable. Observations of this kind suggest some clues to dynamical climate modelers for improvement of model physics, data assimilation, and parameterization schemes, especially, those used in climate projection over West Africa and the global tropics, as a whole. This work was published in the SIAM Data Mining Conference.

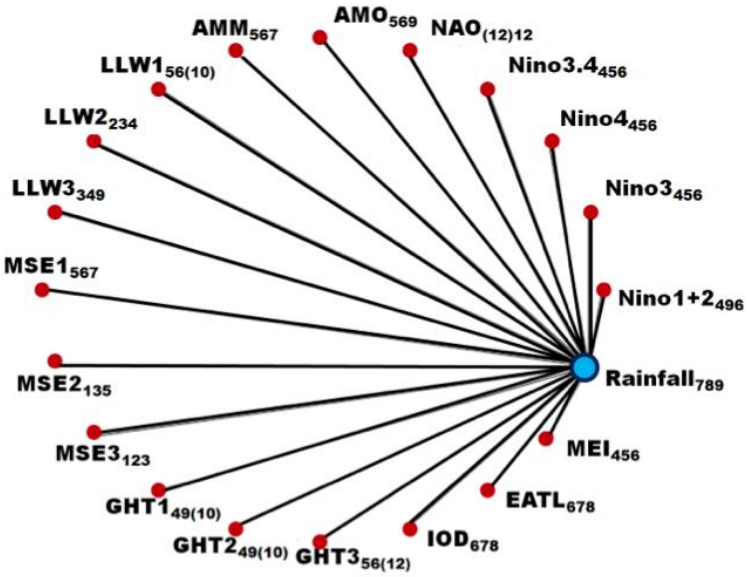


Figure 1: A putative phenomenological model for the eastern Sahel rainfall variability. Each node in the graph describes a significant climate variable, whereas the subscript describes the prominent phases for the corresponding variable.

Using this methodology, we have been able to reveal multiple pathways showing the influence of the North Atlantic Oscillation on the Sub-Saharan African seasonal climates. Previously established North Atlantic Oscillation (NAO) forcing over northern Africa suggested a homogeneous regional-scale climate response, obscuring its actual impacts and pathways at different spatial scales. Empirical evidence is hereby provided to show that multi-year NAO variability impacts the Sahel Region, Western Sahel, and Eastern Sahel boreal summer climates differently. A combined multiscale, multi-phenomena annual cycle impact analysis of global ocean surface temperature and tropospheric features exclusively capture intraseasonal-to-subseasonal NAO variability from winter to mid-spring. This detection is one of the primary sources of variability in the summer climates significantly attributed to different temporal variability unique to each spatial scale, hence separating the dynamics. The corresponding general circulation patterns reveal spatially coherent, but distinct upper level flows that teleconnect to the Himalayas anticyclonic variability. Mechanisms associated with anomalously wet and dry conditions vary along preferential pathways, as the NAO evolves through diverse spatial scales and interacts with distinct oceanic-atmospheric modes.

A relative humidity-based index calculated over the Somali region of East Africa was found to have a high correlation with Atlantic hurricane seasonal count, second only to AMO among well-known climate indices (Figure 2). Furthermore, regression analysis with elastic net regularization shows that while AMO, TNA and AMM have higher total contributions to the explained variance of Atlantic hurricane seasonal count, this East African relative humidity-based index has the highest unique

contribution that cannot be explained by other climate indices (Figure 3). This shows the importance of this region for the comprehensive understanding of the sources of hurricane inter-annual variability in the Atlantic basin. Its omission from predictive models could significantly limit the accuracy of the forecast of Atlantic hurricane seasonal activity.

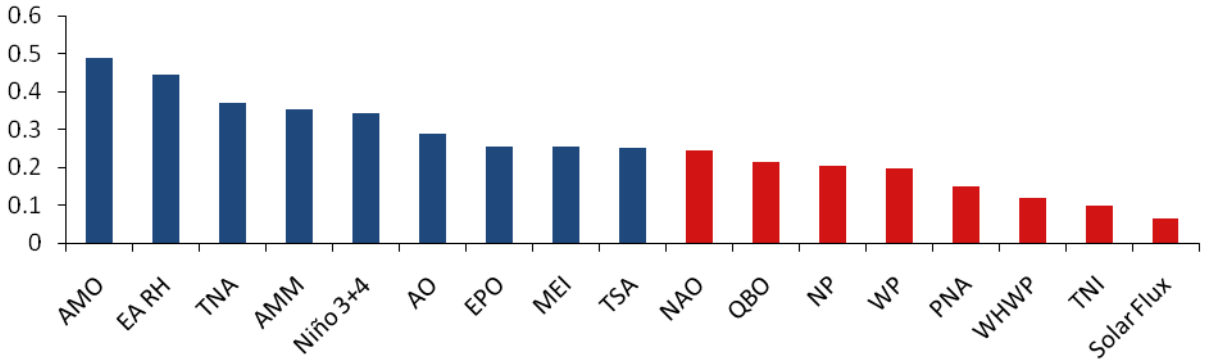


Figure 2: Histogram of absolute values of highest Spearman correlation between climate indices and relative humidity at East Africa (EA RH) in the months of January to June and Atlantic hurricane seasonal count from 1950 to 2011. Blue columns indicate statistically significant correlations and red columns indicate not statistically significant correlations with p=0.05.

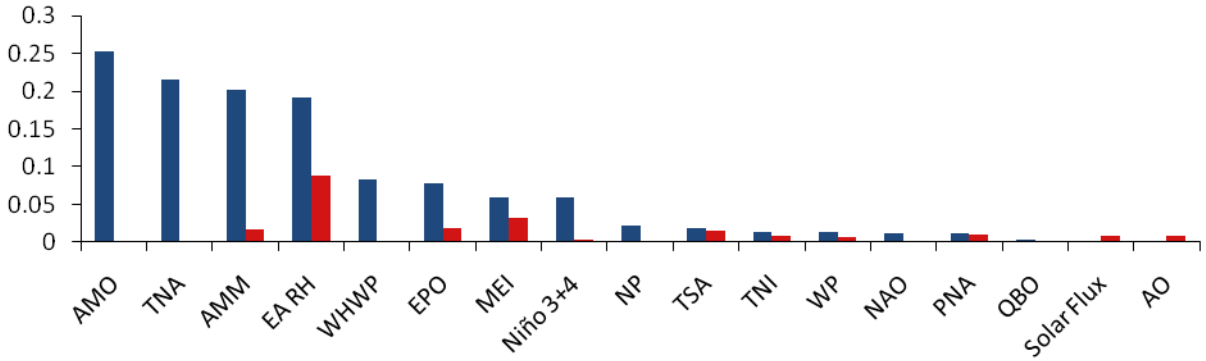


Figure 3: Histogram of total and unique contributions to explained variance of regression model with climate indices and relative humidity at East Africa (EA RH) as predictors and Atlantic hurricane seasonal count from 1950 to 2011 as response. Blue columns indicate total contributions to explained variance and red columns indicate unique contributions to explained variance.

1.3 Forecasting Seasonal Climate Extremes

Contributors: Altaher (U-NCAT), Fetanat (G-NCAT), Gonzalez (G-NCSU), Chen (G-NCSU), Choudhary (F-NWU), Gonzalez (G-NCSU), Hendrix (P-NWU), Homaifar (F-NCAT), Knapp (C-NCDC), Kumar (F-UMN), Lacewell (G-NCAT), Melechko (C-NCSU), Njoku (U-NCAT), Pansombut (G-NCSU), Samatova (F-NCSU), Semazzi (F-NCSU), Sencan (G-NCSU), Tetteh (G-NCSU)

Activities:

The goal of this activity is to significantly improve the predictive skill of the forecasts for seasonal climate extremes and to demonstrate the value of these data-driven methodologies in the context of two use-cases: (1) North Atlantic hurricane activity and (2) Sahel rainfall activity. Two complementary methodologies have been developed:

(a) *Classification-based forecasting the **activity** of the extremes*: Real-world dynamic systems such as physical and atmosphere-ocean systems often exhibit a hierarchical system-subsystem structure.

However, the paradigm of making this hierarchical/modular structure and the rich properties they encode a "first-class citizen" of machine learning algorithms is largely absent from the literature. Traditional data mining approaches focus on designing new classifiers or ensembles of classifiers, while there is a lack of study on detecting and correcting prediction errors of an existing forecasting algorithm. In this work, we developed DETECTOR, a hierarchical method for detecting and correcting prediction errors in extreme event forecasts by employing the whole–part relationships between different systems. We also proposed FORECASTER, an algorithm that constructs a forecast-oriented feature elimination-based ensemble of classifiers for robust forecasting of extreme events (e.g., low, normal, or high hurricane activity season). In contrast to existing classification methods that predict the current system state, FORECASTER predicts the future phase of the system based on the preceding multivariate data, and it is able to handle highly underdetermined problems. Forecaster supports nonlinear system behavior as well.

(b) *Regression-based forecasting the **activity** of the extremes*: Complex dynamic systems rarely act in a linear manner, and have been perceived to behave in phases. Hence, linear regression methodologies are not necessarily suitable to capture the differences in behaviors between different phases, as phase transitions are not specifically accounted, increasing the likelihood for an increased amount of outliers to the linear fit. To advance regression mining techniques in a manner that utilizes all *a priori* knowledge obtained by more robust mining techniques, such as classification methods, we can adapt regression experimentation to be executed in a manner that consistently handles data as pertaining to specific classifications. In doing so, data is studied in a manner that is more adept to following similar behaviors and underlying systematic trends. In this work, we developed a means to obtain classification results from FORECASTER methodology and based on the classified categories, train individual regression models for tuple subsets identified to pertain to each individual class. The methodology creates an ensemble of classifiers in which each member consists of a specific subset of features being utilized and an individual classification result. As such, each member identifies proper features to utilize for the regression methods, and the results of each classification are then fed to the framework to ensure a singular regression method is executed for each class, and repeat this process for each ensemble member. Several methods were then applied to combine the regression ensemble members, such as simple arithmetic mean, median, and Bayesian Model Averaging. To then correlate this end result with existing data and validate the methodology, leave-one-out cross validation was used for our seasonal hurricane activity prediction use case and quarter-out validation for our Sahel rainfall use case for comparability with reference literature. We then performed correlation calculations to effectively compare the captured trends with known data.

(c) *Statistical estimation of the **intensity** of the extremes*: Developing an automated technique to estimate TC intensity and to overcome the existing errors in estimation is still a challenge. The Dvorak technique (DT) is the state-of-the-art method that has been used over three decades for estimating the intensity of a tropical cyclone. The DT subjectively estimates TC intensity based on visible and infrared satellite images. Despite wide usage of the DT for TC analysis, it has some limitations. Especially, DT does not use the valuable historical data mainly because of the computational and human resources. Our approach was inspired by the availability of historical tropical cyclone satellite data. We hypothesize that discovering unknown regularities and abnormalities that may exist in the large group of past observations could help human experts interpret TC intensity changes. Our goal is to provide a data mining tool that increases the ability of human experts to analyze huge amount of historical data for TC intensity estimation. This line of research discovers a set of facts and guidelines with the statistical justification.

Findings:

Experimental results show that DETECTOR successfully detects and corrects more than 25% of prediction errors in the results of three traditional classification methods with an average of 12% accuracy increase on three extreme event prediction tasks, and by combining DETECTOR and FORECASTER, we can further improve the prediction accuracy by up to 6.7% based on the prediction result of

FORECASTER, which has already improved the prediction accuracy of traditional methods by up to 18%. FORECASTER also increases prediction accuracy over traditional ensemble methods by up to 23% on many datasets in other application domains such as biology. We published our results as an NCSU-CSC Technical Report (1840.2/2408) and submitted a paper to the ICDE conference that is under review. We leveraged the advantages offered by FORECASTER to significantly improve the performance of regression methods. We also leveraged FORECASTER and proposed dynamic networks-based methodology for in-advance prediction of the dynamic tracks of emerging extreme events. We applied this methodology to forecasting whether hurricanes will become land-falling 5-15 days after their embryonic formation.

The regression-based methodology was applied to the prediction of rainfall over the Sahel region in Africa, as well as to seasonal hurricane activity prediction. Experimental results show that this methodology, being bound by a classifier accuracy of 85%, can achieve roughly a 8% increase over standalone traditional regression methods by obtaining a 81% correlation when applied on seasonal hurricane activity prediction. For the rainfall prediction, the methodology found a 18.75% increase over previous attained skill. As a proof-of-concept, we found as part of our tests that as the classifier used improved in accuracy, the regressions inheriting these results would also perform better and the trendlines for these would be much closer to the observation.

Our intensity estimation algorithm has two parts: temporal constraints and image feature analysis. This study focuses on the temporal constraints. Temporal information provides *a priori* estimates of storm intensity (in terms of wind speed) prior to using any satellite imagery analysis. Hurricane Satellite data (HURSAT–B1) includes best-track intensity used as a ground truth data. A case study using North Atlantic Hurricane Satellite data from 1978-2006 is considered. The temporal analysis uses the age of the cyclone, 6, 12 and 24 hours prior intensities as predictors of the current intensity. The 10 closest analogs (determined by a K-nearest-neighbor algorithm) are averaged to estimate the intensity. The distribution of intensity estimation errors of the proposed technique, in Figure 4, shows that 50% of the estimates have a mean absolute error less than 2.4 knots, 75% are within 4.4 knots and 90% are within 7.5 knots. Several validation tests were conducted to statistically justify the proposed algorithm using K-Fold Cross-Validation. The resulting overall root mean squared error (RMSE) of our algorithm is approximately 4.6 knots compared to 11.7 knots from the DT on the same dataset. To analyze the effect of noise, a Gaussian noise of zero mean with 5 knot and 6 hour standard deviations are considered for the prior intensities and duration respectively. The results of the noise analysis indicate that the average RMSE is approximately 8.2 knots compared to 11.7 knots from the DT. Overall, 30% to 55% improvement has been achieved compared to the DT. The current analysis has the ability to decrease the DT noise and also has the potential to provide new temporal constraints on satellite analysis.

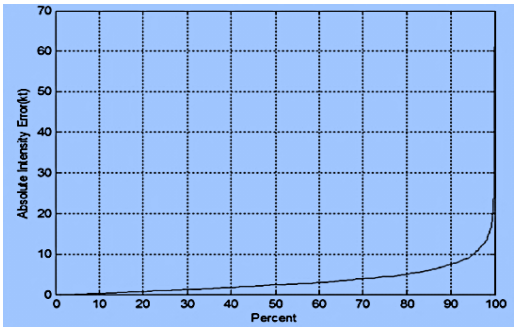


Figure 4: Distribution of proposed technique classification errors (tested for storms during 1997 - 2003) in the Atlantic basin.

1.4 Learning Spatial Decision Trees for Geographical Classification

Contributors: Jiang (G-UMN), Knight (F-UMN), Shekhar (F-UMN)

Activities:

Given learning samples from a spatial raster dataset, the geographical classification problem aims to learn a decision tree classifier that minimizes classification errors as well as "salt-and- pepper" noise; i.e., classification errors that are spatially isolated. The problem is important in many applications, such as land cover classification in remote sensing and lesion classification in medical diagnosis. In the climate science domain, studies show that wetlands contribute to over half of the global emissions of methane, a powerful greenhouse gas; accurate wetland mapping thus helps climate scientists to understand the global

warming process. However, the problem of detecting wetlands in remote sensing data is challenging due to spatial autocorrelation in the data. Existing decision tree learning algorithms, such as ID3, C4.5, and CART, assume that data items are drawn independently from identical distributions, so they produce a lot of salt-and-pepper noise in classification results.

To address the shortcomings of traditional approaches, we propose a new spatial decision tree (SDT) model, including a spatial information gain "interestingness" measure. We use this model to develop an SDT learning algorithm, and we have conducted a case study on a real-world remote sensing dataset to evaluate the proposed approach.

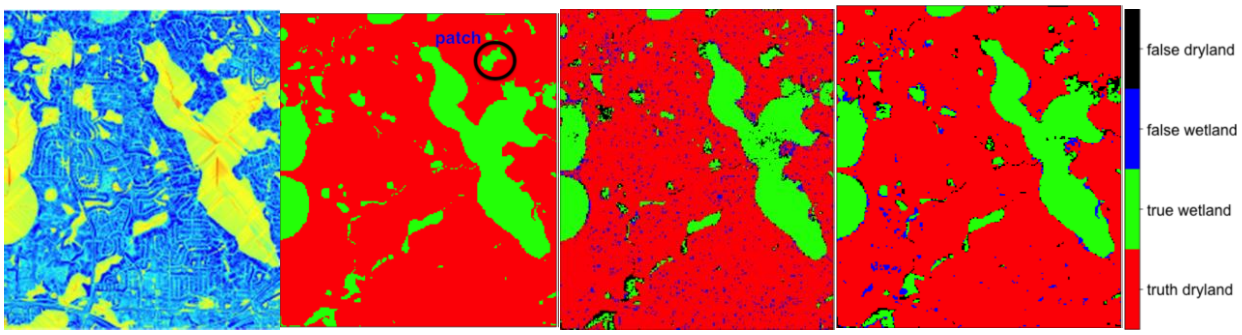


Figure 5. Sample input and output of the geographical classification problem: (a) an explanatory feature, (b) ground truth, (c) output of traditional decision tree, (d) output or spatial decision tree.

Findings:

We tested the spatial decision tree on a real-world high resolution (3m by 3m) remote sensing dataset that was collected from the city of Chanhassen, Minnesota (shown in Figure 5). Explanatory features consisted spectral information over multiple time periods (2003, 2005, and 2008), including visible and near-infrared bands of aerial photos, as well as topographical derivatives such as slope and curvature. Ground truth class labels (wetland and upland cover types) were obtained from a wetland delineation field crew and trained photo interpreters. Figure 5 shows the sample data (a-b), as well as results from traditional decision trees (c) and the proposed approach (d). This case study shows that the proposed spatial decision tree learning algorithm reduced wetland misclassification by more than half relative to a traditional decision tree classifier (21,882 errors vs. 51,907). The salt-and-pepper noise is also reduced, as the output of the proposed method exhibits significantly higher autocorrelation than that of the traditional decision tree. Experiments show that the proposed method is scalable to large datasets.

2. DESCRIPTIVE MODELING

The work in this area studies the interrelationships between different variables or processes within the same subsystem and interactions among multiple subsystems to gain a better understanding of the behavior of the climate system. In particular, climate data, be it observed or model-simulated, has complex dependencies across space as well as time. These dependencies can be local in nature, thus involving spatial and temporal units in a neighborhood, or there may be long range teleconnections and long memory time series effects. One way to capture these relationships and dependencies is with complex networks or descriptive data mining approaches, such as clustering. Another approach is to capture various feature of these dependencies through statistical modeling, estimation, testing, and inference. To this end, Bayesian, empirical Bayes and resampling-based techniques of inference are studied. Regardless of approach, this broad class of research activities will feed into other parts of the project, primarily into predictive modeling.

Accomplishment Highlights:

One project has developed techniques based on complex networks to find climate dipoles, i.e., pairs of regions on the surface of the Earth whose behavior has significant impact on the global climate. Novel aspects of this work are network patterns whose edges involve negative correlations and the use of reciprocal nearest neighbor pruning to eliminate irrelevant edges from the network. More recently, the dipole finding approach was extended to address the issue of the statistical significance of the patterns discovered and to take into account time lags, i.e., the fact that events in one part of the climate system can have a delayed impact on the rest of the climate system. The dipole work has identified potentially new climate indices and has been used to evaluate and compare General Circulation Models (GCMs) in terms of how well they reproduce known climate phenomena. Other work has focused on identifying the structure of climatological data on a decadal scale to identify major climatological shifts (or regime changes) using one or more climate variables. By applying this technique to the output of General Circulation Models (GCMs), we hope to be able to anticipate such regime shifts in the future evolution of the climate system.

Understanding statistical dependencies between climate variables and processses form a key step in climate data analysis. Thus, there has also been research in statistical dimensionality reduction and dependence modeling. One focus in dimensionality reduction is using a Probabilisitic Matrix Factorization method for dimensionality reduction. Within the area of dependency modeling, we are studying the directed relationship between variables as part of statistical detection of causality relations, as well as the robust, multivariate generalizations, and efficient Bayesian modeling of such relations. In additional dependence modeling work, we have developed a time series alignment method and extensions of Granger causality to multivariate quantiles and categorical data.

Another project has created a distance-based ENSO index (S-ENSO for spatial ENSO) that tracks the location of maximum near-tropical Pacific warming anomaly instead of its absolute warming. Our spatial ENSO index correlates better with seasonal tropical cyclone activity than standard ENSO indices, especially with increased lead times where traditional ENSO suffers from a "predictability barrier."

2.1 Dipole Discovery

Contributors: Ganguly (F-NEU), Kawale (G-UMN), Kumar (F-UMN), A. Kumar (G-UMN), Liess (R-UMN), Ormsby (U-UMN), Samatova (F-NCSU), Steinbach (R-UMN), Steinhaeuser (R-UMN), Sumler (U-UMN)

Activities:

Dipoles are defined as a pair of regions such that locations within each region are highly positively correlated with each other and locations across these regions are negatively correlated to each other. Such dipoles have proven important for understanding and explaining the variability in climate in many regions of the world. Scientists have known of the existence of such dipoles for about a century. One of the best known pressure dipoles is the Southern Oscillation (SO). This dipole is represented by the Southern Oscillation Index (SOI), which is a time series defined as the difference in the pressure anomalies at Tahiti and Darwin, Australia.

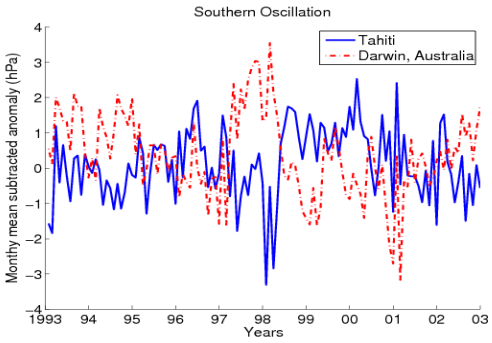


Figure 6. Pressure components of the Southern Oscillation Index (SOI).

SOI captures fluctuations in pressure around the tropical Indo-Pacific region that correspond to the El Niño Southern Oscillation (ENSO) climate phenomenon, see Figure 6. Historically, these dipoles have been discovered by human observation or by using pattern analysis techniques such as the Empirical Orthogonal Function (EOF) over a limited region. However, there are several limitations of the existing methods of finding these relationships, and they required considerable research and insight on the part of

the domain experts involved. Knowledge of these teleconnections and their interactions is particularly important for predicting climate extreme events. For example, while the cold winter over Europe in 2010 could be largely explained by the North Atlantic Oscillation (NAO) which is another teleconnection, and other local indices, the cold winter over North America at the same time is largely due to a combination of NAO and ENSO. Further, the ability to address important questions like the degree of climate change and its potential impacts requires a deeper understanding of the behavior and interactions of these atmospheric processes as well as to capture them precisely. This project aims to provide systematic data guided approaches to find such relationships in spatio-temporal data. Discovery of relationships or dependencies among climate variables involved is extremely challenging due to the nature and massive size of the data. Data guided approaches offer a huge potential for characterizing and discovering unknown relationships and advancing climate science.

Findings:

We present a novel graph based approach to find all the dipoles in a given dataset. We model the climate data as graph with the nodes of the graph represented by the regions on the Earth and the edges represented by the correlation between the anomaly time series of two regions.

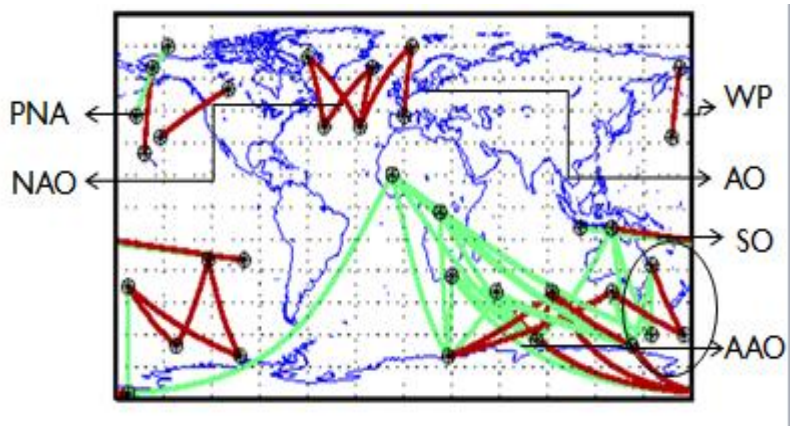


Figure 7. Dipoles declared significant in the NCEP dataset. A set of newly detected dipoles between Australia and the Southern Ocean is marked with a circle.

Our approach based on a novel Shared Reciprocal Nearest Neighbor (SRNN) algorithm, has a number of advantages. The approach allows us to detect all dipoles represented in an individual global dataset within the selected time frame and to determine their individual strengths. It makes it possible to discover new dipoles that may not have been seen. It enables tracking the movements of these dipoles and studying their interactions in a much more systematic way. Another important application of global dipole analysis is its use in the understanding of the skill of various General Circulation Models (GCMs) used for climate prediction. We also present a novel approach for testing the statistical significance of dipoles.

This approach takes into account the autocorrelation, the seasonality and the trend in the time series over a period of time, and is applicable to a wide variety of other problems in spatio-temporal data mining.Figure 7 shows the set of dipoles discovered using our algorithm and the NCEP reanalysis data for the period 1951-2000. Red edges denote dipoles considered significant and green edges denote dipoles considered insignificant by our statistical significance testing method.

Nearly all of the 23 significant dipoles correspond to known phenomena like the NAO, Arctic Oscillation (AO), the Pacific North-America (PNA) pattern, the West Pacific (WP), SO, and the Antarctic Oscillation (AAO). Dipoles considered statistically insignificant dipoles appear to be inconsistent with physical understanding of the climate phenomenon. One significant dipole not yet known is being investigated as a new phenomenon. This newly detected phenomenon can be distinguished from existing ones by correlating the respective index time series with neighboring time series. Fig. 8a shows the correlations between AAO and all time series over land points. Strongest correlations are found over Antarctica, South America, and Australia. In contrast, the correlation with the new dipole (Fig. 8b) is much weaker over Antarctica and shows opposite signs over southern Australia and central South America compared to the AAO. This new finding allows for a much more detailed description of relationships between climate, especially over Antarctica and southern Australia.

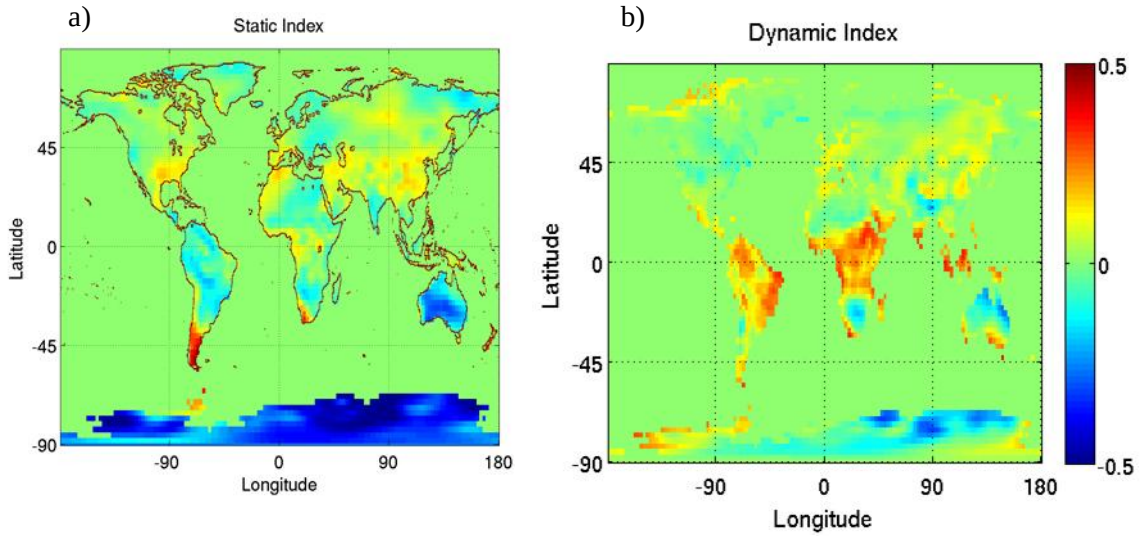


Figure 8. Correlations between dipole time series for a) the AAO and b) the new dipole between Australia and the Southern Ocean.

2.2 Community Dynamics and Analysis of Decadal Trends

Contributors: Agrawal (R-NWU), Jin (G-NWU), Choudhary (F-NWU), Hendrix (P-NWU), Liao (R-NWU)

Activities:

Complex networks constructed from climate data (see Figure 9) have been motivated to understand the dynamics and large-scale processes of the global climate system, for example, correlations and teleconnections across multiple spatio-temporal scales. Our contributions in this area have primarily focused on extracting and analyzing the community structure of such 'climate networks'; an example is given in Figure 9, which shows that the variability of sea level pressure is relatively uniform in the polar regions but varies across ocean basins throughout the tropics. We are developing algorithms for efficiently constructing and analyzing these networks from various observed and model-generated datasets, which can be used both to gain a better understanding of global climate dynamics as well as compare and evaluate climate models in terms of their ability to reproduce the expected patterns and phenomena.

We are continuing to develop our technique for identifying the structure of climatological data on a decadal scale. This technique involves processing one or more temporal climatological features, such as surface air temperature or sea level pressure; dividing the data into overlapping 10-year time windows, and forming a network based on the correlations in the weather within each time window. By clustering the networks and

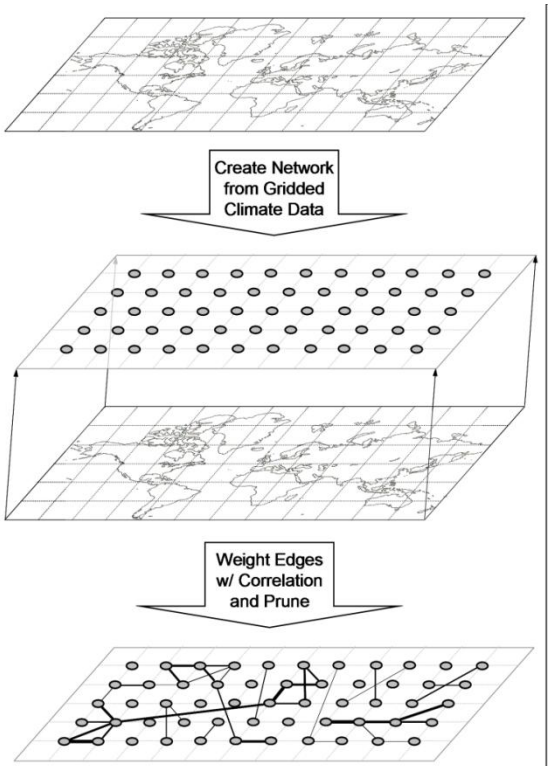


Figure 9: Schematic workflow of constructing climate networks from a gridded dataset.

analyzing of how they evolve over time, we hope to identify major climatological shifts (or *regime changes*) in the data. By applying this technique to the output of General Circulation Models (GCMs), we hope to be able to anticipate such regime shifts in the future evolution of the climate system.

Findings:
We applied our technique to surface air temperature data from the NCAR/NCEP Reanalysis dataset, and by manually scanning the output, we identified dynamics consistent with known climate phenomena, specifically El Niño and the desertification of the Sahel region in Africa. We have published a paper with this result in the ClimKD workshop at ICDM. Since then, we have extended the technique to create networks that represent the evolution of multiple variables across time, and we have begun work on developing algorithms to automatically detect regime shifts and other major events in the evolutionary networks that we generate.

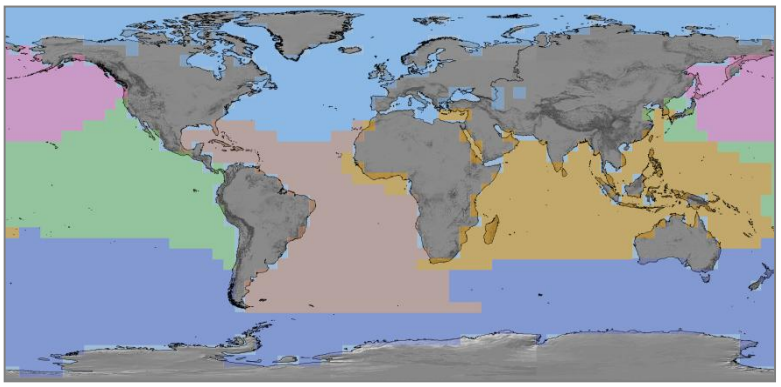


Figure 10: Community structure of the climate network constructed from monthly mean sea level pressure data.

We have also applied these algorithms to a wide range of climate model outputs from the Climate Model Intercomparison Project phase 3 (CMIP3) archive, which formed the basis for much of the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC). Our empirical observations suggest that the dynamics generated by different initial-condition runs of the same model tend to be more similar than those of different models as one might expect; however, the models vary widely in their ability to reproduce the structure obtained from observations (NCEP Reanalysis).

2.3 Predicting Future Tropical Cyclone Activity under Warming Scenarios
Contributors: Faghmous (G-UMN), Haasken (U-UMN), Kumar (F-UMN), Le (U-UMN), Liess (R-UMN), Mesquita (C-BCCR), Semazzi (F-NCSU), Smith (U-UMN), Steinbach (R-UMN)

Activities:
Future tropical cyclone (TC) activity and its response to global warming are some of the most controversial topics in climate change debates. While certain variables that modulate TC activity have been identified, the conditions that cause perturbations to become major storms are poorly understood, effectively making future TC activity predictions challenging. Traditionally, future TC activity has been simulated using high-resolution spatio-temporal climate data generated by physics-based models. Due to the susceptible nature of TCs such high resolution is essential to adequately model TC characteristics. Relying solely on these model outputs, however, raises several challenges: First, reproducibility is challenging the models' dependence on initial conditions and parameterization. Second, given the complex interactions between climate variables at small scales, quantifying the impact any single variable might have on model output is ambiguous. Finally, relying solely on physics-based models must be supplemented given that model output data are too coarse to properly model and track hurricanes. To address some of these limitations we are developing scalable data-driven methods that leverage the increase in high quality observational climate data to build statistical models that adequately capture the climate states that favor cyclogenesis. We detect physical pathways by which each individual index impacts the Atlantic TC activity. In addition to linear correlation techniques, we also employ composite analysis: we subtract the states of the system during extreme negative index periods from the positive extremes to identify the strength of the influence of each index.

Findings:
We began by trying to formalize the relationship between Atlantic sea surface temperature (SST) and Atlantic TC frequency. Previous data analysis studies proposed that increase in Atlantic SST alone could not explain the recent increase in Atlantic TC frequency given that other basins (Pacific, Indian) experienced similar SST upward trends without the associated increase in their TC frequency. Our data-driven research found that warming of the region off the African West coast near 20N during the months preceding TC season, as well as warming of Atlantic between 12-20N and 18-60W provide favorable conditions for African Easterly Waves (AEW) to form into TCs. Thus, increase in Atlantic TC frequency can be attributed to the warming of the above-mentioned regions and favorable AEW activity.

We then investigated the impact of Pacific SSTs on Atlantic TC frequency and activity. The warming anomalies of sea surface temperatures (SSTs) along the near-equatorial Pacific Ocean related to the El Niño-Southern Oscillation (ENSO) are measured over a fixed region (NINO3.4) have well documented global long-range weather teleconnections with North Atlantic TC activity. Traditionally, ENSO's impact on Atlantic TC has been abstracted by monitoring the warming of static regions along the equatorial Pacific Ocean. We posited that the spatial distribution of Pacific Ocean warming might provide better predictive insights into ENSO-Atlantic TC activity than warming anomalies alone. We developed a set of dynamic indices as well as their attributes that track the location of maximum and minimum near-tropical Pacific warming anomaly. Many of these dynamic indices have shown much higher correlations with Atlantic hurricane counts than that is observed with fixed ENSO indices. Our indices include detailed information about the distribution of SST, surface pressure, and outgoing longwave radiation (OLR). OLR is a widely used proxy for high elevation clouds and vertical velocity. As an illustrative example, Figure 11 shows that our dynamic OLR index correlates better with seasonal TC activity than standard ENSO indices. TC activity is measured as potential intensity, which calculates the maximum possible hurricane intensity from temperature, pressure, moisture, and distance from equator. This is especially true for increased lead times where traditional ENSO suffers from a "predictability barrier."

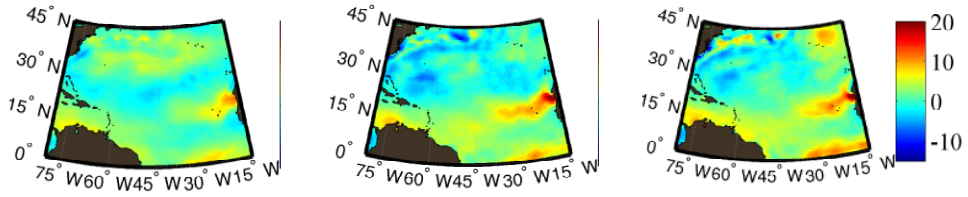


Figure 11: Potential intensity composites for NINO3.4 (left), OLR over the max. SST anomaly (middle), and August-October Atlantic TC counts (right). The OLR index is able to reproduce the large-scale conditions over the Atlantic much approved compared to the NINO3.4 index.

2.4 Statistical dependency modeling and dimensionality reduction
Contributors: Chatterjee (F-UMN), Hyman(G-UMN), Beard (G-UMN), Dietz (G-UMN), Bandyopadhyay (C-Citigroup, India), Bhattacharjee (C-Dundee-Scotland), Klein (C-Census), Lahiri (C-Maryland-College Park), Maiti (C-Michigan State), Mukhopadhyay (C-Virginia Commonweath), Tillinghast (C-Census), Tran (C-Census), Wright (C-Census), Karpatne (G-UMN), Blank (U-UMN), Middleton (U-NCAT), Boriah (R-UMN), Steinhaeuser (R-UMN), Kumar (F-UMN)

Activities: Understanding statistical dependencies between climate variables and processses form a key step in climate data analysis. We are pursuing several research directions within the broad framework of dependency modeling, while considering the multivariate high-dimensional nature of the data, need for robust methods for handling uncertain observations, and long-range temporal and spatial dependencies, among others. In particular, we study directed relationship between variables, as part of statistical detection of causality relations, consider their robust, multivariate counterparts, consider efficient

Bayesian modeling of such settings, consider multivariate ranking, classification and change detection with such data, and also consider Winsorization, benchmarking, risk estimation and prediction bound estimation, and nonparametric modeling of data with multiple sources of variability and complex dependency structures. This research is intimately related to our ongoing research on small area statistics.

Finding relationships between oceanic drivers, such as sea surface temperature anomalies, and naturally occurring land cover disturbances, such as forest fires, is of prime concern, especially for understanding diverse phenomena of Earth's complex systems. Oceanic drivers induce non-linear influences on target events with varying temporal lags and spatial coverage in different geographic and climatic regions, making the problem complex and challenging. As an example, increased sea surface temperatures are known to affect the amount of precipitation received during the wet season in South America, eventually making the vegetation more susceptible to fire. Recent work has shown relationships between Sea Surface Temperature (SST) anomalies in the Atlantic and Pacific Oceans, and Fire Season Severity (FSS) in South America. Specifically, they used Oceanic Niño Index (ONI) and Atlantic Multidecadal Oscillation (AMO) as climate indices responsible for affecting FSS. We investigated the statistical significance of previous findings at varying spatial resolutions and temporal scales, by devising appropriate randomization experiments, and created an interface for exploring our findings in detail. We are further focusing on extending and improving the underlying machinery for finding relationships which honor data characteristics and pattern semantics, such as spatial coherence, temporal consistence, and inherent count data characteristics of FSS.

Traditional principal component and related techniques are very commonly used for studying climate data, although not all is known about the properties of these methods when classical data assumptions do not hold. We are investigating such traditional principal component analysis and its variants from a high-dimensional, dependent data perspective. We are also studying lower dimensional manifold structures associated with minimixing contrast functions. We are also considering the possibility of easily obtainable and interpretable penalized regression estimators, which may not be sparse, but which may then be transformed to a sparse representation. Statistical properties of such regression estimators are under study. In another study, we are focussing on Probabilisitic Matrix Factorization methods for dimensionality reduction.

Findings: Time series data in climate research presents many significant challenges from a statistical standpoint. For both predicting future values and understanding current relationships, it is necessary to be able to model the interdependence of numerous variables (i.e. temperature, precipitation, vegetation, etc.) from many different locations over extended periods of time. Yet the common assumption of independence does not hold in this situation, and assuming everything is dependent on everything else leads to overly complicated models or incorrect covariance structure assumptions. One possible path to overcoming these difficulties is to make an assumption of *sparsity* in the covariance of the time series variables without making explicit the relationships of dependence. This essentially means that we do not decide ahead of time which variables at which times and locations are independent, but in the estimation of the model parameters we ensure that almost all of the variables will be independent. Our research focused on simulations comparing estimates of parameters in vector autoregression (VAR) models based on the assumption of sparsity versus traditional likelihood-based estimates. The sparsity-based methods incorporate an L_1-penalty that is known to increase the number of parameters estimated to be zero compared to the maximum likelihood, conjugate Bayesian and other similar estimates, which inevitably give all parameters some non-zero value, however insignificantly different from zero they might be. A challenge in the sparsity-based approach is that there is a tuning parameter that goes with the L_1-penalty that must be specified, and which determines the level of sparsity forced upon the covariance matrix. This parameter can be adjusted to achieve some form of optimality, but this depends on the true parameters values. The simulations explored this relationship, and also confirmed previous research that has shown that while using sparsity introduces bias into the estimates, when the true covariance matrix is mostly zeroes, it does better in the sense of guessing more zeroes correctly than other approaches. Sparse

estimates of covariance matrices were made using the glasso function in the package of the same name in R.

We have obtained a time series alignment method, and also extensions of Granger causality to multivariate quantiles and on categorical data. We have obtained small area procedures for data-depth based multivariate ranking. A major part of our work is on parametric bootstrap for predictive modeling involving benchmarking, Winsorization and other robust procedures, and largely the computational part of this work is complete. A theoretical machinery has been obtained to use consistent nonparametric bootstrap procedures for similar cases.

Several new methods have been discovered and studied to analyze complex dependent data. A new method called Probabilistic Matrix Addition has been thoroughly studied. Methods for detecting rapid or systematic changes in climate patterns, and of testing for statistical significances of tele-connections are under study. Partial results have been obtained on modeling of climate extremes and their behavior under climate change.

Continuing work in this area will potentially combine the sparse estimation methods with multivariate Granger causality tests for comparing the relationship among several time series to better predict the future behavior of the time series variables. The specific application of these techniques to understanding the climate in the Sahel region just south of the Sahara Desert remains a future research possibility.

3. Change Detection

Climate and environmental data are characterized by various types of changes including periodic patterns, subtle trends and major shifts. For example, positive feedbacks inherent in the climate system are often responsible for prolonged droughts persistent over space and time, and have contributed to large-scale famines and displacement of millions of people who live off the land. Abrupt land cover changes in the spatial context indicate transitions of environment between different ecological zones and thus offer potential for understanding interactions between climate change and ecological systems. Further, identifying patterns of change in the sea surface height has applications in detecting ocean eddies, which are key drivers of marine ecoystems. Hence, the automatic detection of changes in large spatio-temporal datasets, over land, ocean and atmosphere, is important for monitoring and understanding the behavior of the global climate system. Our work in the area of change detection and analysis include wavelet and graphical model based approaches for abrupt change detection, statistical methods for parameter change detection with complex dependency patterns, and detecting change intervals and persistent regimes in time series based representations of non-stationary data.

Accomplishment Highlights:

We developed a technique for detecting major droughts which are persistent over space and time. We formulated the problem of drought detection as a Maximum-a-Posteriori (MAP) inference problem on a Markov Random Field (MRF), and developed an efficient MRF based abrupt change detection algorithm. MAP inference based drought detection has been shown to find almost all major droughts in the past century any many lesser known significant droughts. Of broader significance, the novel KL-ADM method we developed to solve the MAP inference problem has been shown to be highly scalable, i.e., can solve the MAP inference linear program with around 7 million variables in just 15 minutes using just 8 cores.

Multiple projects are ongoing to develop statistical techniques and methods that may be applied to detect changes in data with complex dependency patterns, as with climate data, most immediately the data on Atlantic tropical storms. Theoretical results on change detection methodology have been developed. For example, we have formed and solved a hypothesis testing problem of whether there has been any change in hurricane parameters, e.g., the sustained wind speed and minimum central pressure, over time. In a project where we studied how the duration of each hurricane relates to climate covariates

and whether this relationship has changed, a generalized linear mixed effect change detection methodology is being built and individual hurricane effects are being accounted for. In a third project, where we explore change detection techniques for extreme wind speeds by using top-order statistics, we have made several developments, bootstrap and MCMC methods have been employed extensively, and we have completed developing both the theoretical machinery justifying these, as well as the computational algorithms. In addition, we are building wavelet based techniques to isolate strong low frequency events in temperature or pressure fields that may indicate an abrupt change. Together, these new methodologies are giving insights on the changing behavior of major hurricanes.

For detecting abrupt changes in the land cover, we developed the Abrupt Change Interval Miner (ACIM) and Sub-path Enumeration and Pruning (SEP) for the discovery of interesting spatiotemporal sub-paths/intervals. These algorithms find abrupt changes that occur in geographical space, indicating sharp transitions regions of environment between different ecological zones. We also framed the challenge of monitoring ocean eddies as an unsupervised learning problem and presented a novel change detection algorithm that automatically identifies and monitors eddies in sea surface height data based on heuristics derived from basic eddy properties. We further improved upon the state-of-the art connected component eddy monitoring algorithms to track eddies globally. We are also exploring non-stationary time series clustering to identify persistent climate regimes and construct a dynamical model for each regime. In this work, we are generalizing a finite elements method (FEM-K-trend), a newly developed technique that can find clusters even when there is a time-trend in each regime without any explicit assumptions on data distributions.

Individual Project Reports:

3.1 Finding Regions of Change

Contributors: Banerjee (F-UMN), Fu (G-UMN), Liess (R-UMN), Snyder (F-UMN), Wang (G-UMN)

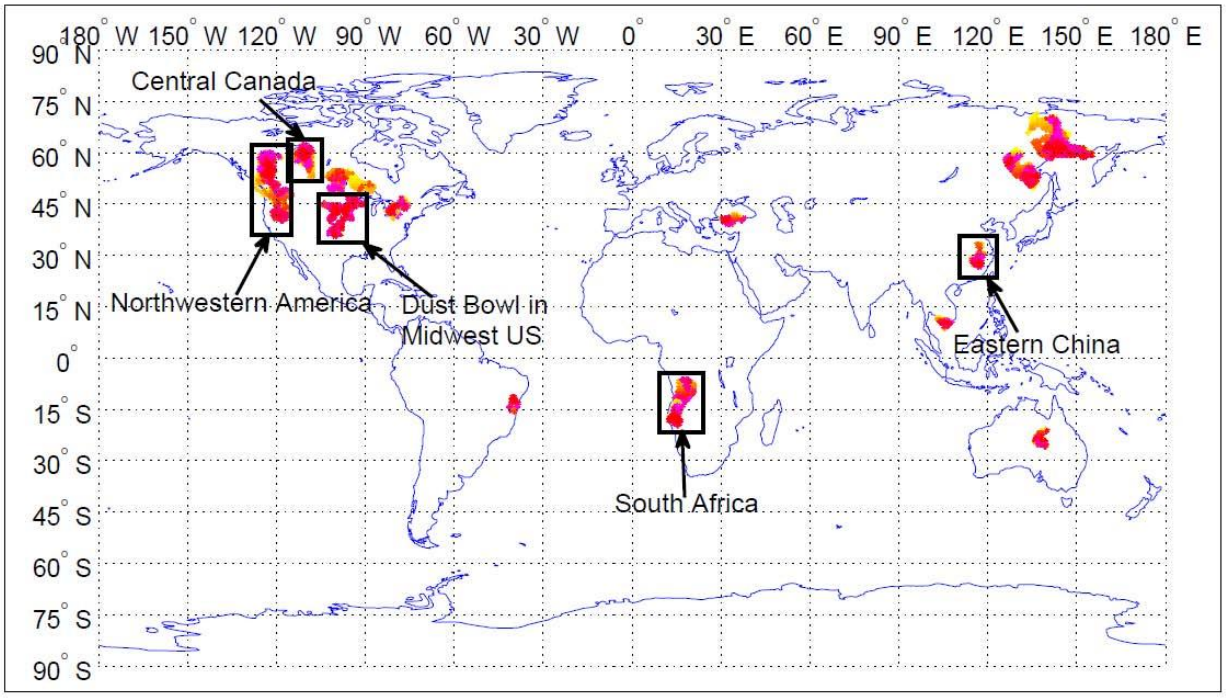
Activities:

Environmental changes such as to our climate, ecosystems, and natural resources, are often abrupt, large, and lead to catastrophic events through nonlinear interactions. Even though detecting and understanding such abrupt changes is of critical importance, limited number of statistical machine learning models have been developed for this purpose. The study of abrupt changes can be divided into two parts—detection, and understanding. We focus on methods for detecting abrupt changes with emphasis on droughts which are persistent over space and time, e.g., the Sahel drought, the dust bowls, etc. We formulate the problem of drought detection as a Maximum-a-Posteriori (MAP) inference problem on a Markov Random Field (MRF) and develop an efficient MRF based drought detection algorithm. In particular, we use a 3-dimensional (latitude × longitude × time) grid graph to model a given spatio-temporal climate dataset, where each node represents a location. Each node is assumed to have a latent 'climate state,' a discrete variable referring to whether the node is 'normal' or 'low' compared to historical normal values. Suitable node and edge potential functions are designed to make sure the inference pays attention to the observed values while maintaining spatio-temporal smoothness of climate states.

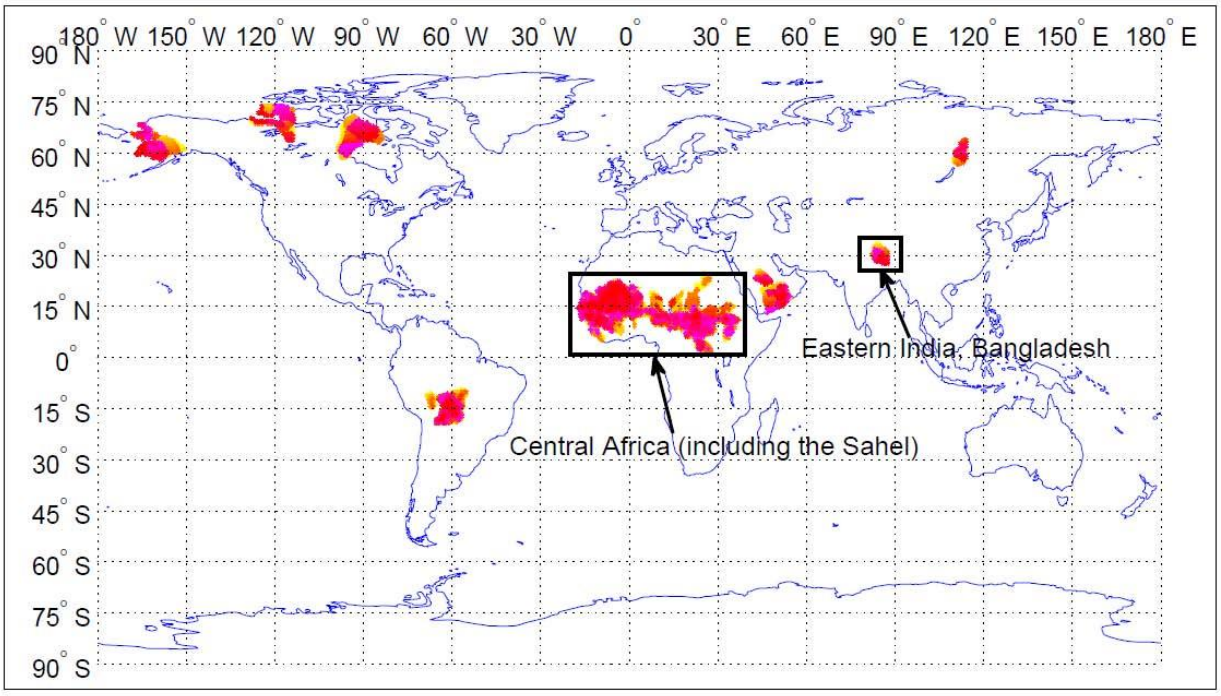
Two approaches have been considered for solving the MAP inference problem. Both consider a graph-structured linear programming (LP) relaxation of the problem, and develop efficient algorithms for solving the LP. The first approach is based on proximal updates where suitable Bregman divergences are used as proximal functions. The second approach is a novel alternating direction method (ADM) which uses KL-divergence based augmentation, and can be easily run in parallel using multiple cores. Both approaches are guaranteed to solve the graph structured linear program, and the solution to the original problem is obtained by suitable rounding techniques.

Findings: The current focus is on using precipitation datasets, in particular the Climate Research Unit (CRU) dataset, for detecting significant droughts. The dataset has high resolution (0.5 × 0.5 degrees) and covers the years 1901-2006. *MAP inference based drought detection has been shown to find almost all*

major droughts in the past century. Figures 12 (a) and (b) show both well-known and less known significant droughts detected by the method. The results have been reported in a paper published in SDM'12, a top-tier data mining conference. In recent work, the novel KL-ADM method has been shown to be highly scalable, since it can run in parallel using multiple cores. When run over 8-10 cores in parallel, the method solves the graph structured linear program with around 7 million variables in just 15 minutes. *The KL-ADM based algorithm has been theoretically shown to have a linear convergence rate, and arguably the fastest and most scalable existing approach to solve graph-structured linear program.* In future work, the method will be applied to climate model outputs for both skill evaluation and projections, as well as paleoclimatic data to understand past mega-droughts.



(a)



(b)

Figure 12: Major droughts detected by the MAP LP approach: (a) Droughts starting in the range 1921-1930, and (b) Droughts starting in the range 1961-1970.

3.2 Detecting change in complex parameters of spatio-temporal data

Contributors: Chatterjee (F-UMN), Lu (G-UMN), Yuan (G-UMN), Kumar (F-UMN), Leiss (R-UMN), Deitz (G-UMN), Snyder (F-UMN), Ganguly (F-NEU), A. Kumar (G-UMN), Fu (G-UMN)

Activities:

Multiple projects are ongoing to develop statistical techniques and methods, that may be applied to detect change in data with complex dependency patterns, as with climate data. Such datasets have features which correspond to complex parameters, like multivariate tail quantiles, latent variable variance components, frequencies and spectra-related quantities, which are typically difficult to study, especially in a change detection context. We concentrate on these parameters since in particular climate contexts they correspond to physically meaningful objects, whose patterns and change in patterns are important for understanding climate change.

In several parts of our work, we concentrate on the data on Atlantic tropical storms for greater understanding of this important climate phenomenon as well. In one project, we study the number of occurrence, the maximum sustained wind speeds and the minimum central pressure of Atlantic hurricanes and detect whether there is a change in these key characteristics.

In a second project, we model the relationship between the number of hurricanes and the maximum sustained wind speeds and build a generalized linear model. Furthermore, we test whether the relationship has changed over time. We research on how the duration of each hurricane relates to climate covariates and whether this relationship has changed. A generalized linear mixed effect change detection methodology is being built and individual hurricane effects are accounted for.

In the third project, we explore change detection techniques for extreme wind speeds by using the top *k*-order statistics. To study the change in extremes, we generalize the traditional approach of looking at maximum statistics. In this project, we study multiple order statistics and see whether the distribution changes over time because multiple order statistics releases more information and is robust to outliers.

Our fourth project departs from the theme of using Atlantic tropical storms, and concentrates on climate model features instead. In this project, using global rainfall temperature, and pressure fields from observations and a subset of IPCC AR4 climate models (A1B scenario) and an ensemble average, we examine both how models capture 20th century abrupt events and how these events may change by 2100 with continued climate change. We use wavelet analysis to isolate strong low frequency events that may indicate an abrupt change. This analysis produces a time-frequency profile that shows the periodicities of the different climate models. Our current wavelet approach is being used to understand how the frequency of these abrupt events is changing with climate change, and to further analyze the physical mechanisms that trigger abrupt changes which can move an environment into a different stable state.

The last project has been undertaken in the second year, while the first three have seen some preliminary studies and developments in the first year, and more comprehensive research in the second year. The effort in the first year was mainly on understanding the data and its features, and on developing the mathematical model to quantify change, and the significance of change. During the second year, we analyzed our mathematical models comprehensively, established theorems relating to the mathematical framework we developed, and performed simulation experiments and data analysis.

Findings:

In the project where we model the number of hurricanes as coming from Poisson distribution, and model the maximum sustained wind speeds and the minimum central pressure as coming from Gumbel distribution, we have formed and solved a hypothesis testing problem of whether there is any change in the parameters of those distributions. The likelihood ratio based approach is implemented and the average run length criteria and p-value criteria are used signal a change. Theoretical results on change detection methodology have been developed.

In the project where we study how the duration of each hurricane relates to climate covariates and whether this relationship has changed, a generalized linear mixed effect change detection methodology is being built and individual hurricane effects are accounted for. Some additional analysis, using climate indices and variables have been studied, along with the effect of the space and time components on the data.

In the third project, where we explore change detection techniques for extreme wind speeds by using top *k*-order statistics, we have several developments. Bootstrap and MCMC methods have been employed extensively, and we have completed developing both the theoretical machinery justifying these, as well as the computational algorithms. We investigate the multinomial change detection model and develop a modified information criteria to select the best change detection model. Binary segmentation procedure is implemented and multiple change detection models are built. From simulated data, the information criteria catch the true model very effectively. This methodology is applied to study the change in proportion for different categories of hurricanes, thus giving insight on the changing behavior of major hurricanes.

Our preliminary findings in the fourth project suggest that models do a reasonably good job of capturing 20th century abrupt events; however, the more subtle events are difficult to extract from the climate record and may indicate either model dampening of abrupt changes or omission of physical mechanisms that contribute to positive feedbacks leading to abrupt changes. Current research is focusing on the mechanisms contributing to abrupt changes and whether or not those processes are well represented in the AR4 suite of models. Figure 13 uses NCEP precipitation data for 1948-2009 and contains all signals between 15.5 and 31 years. In addition to

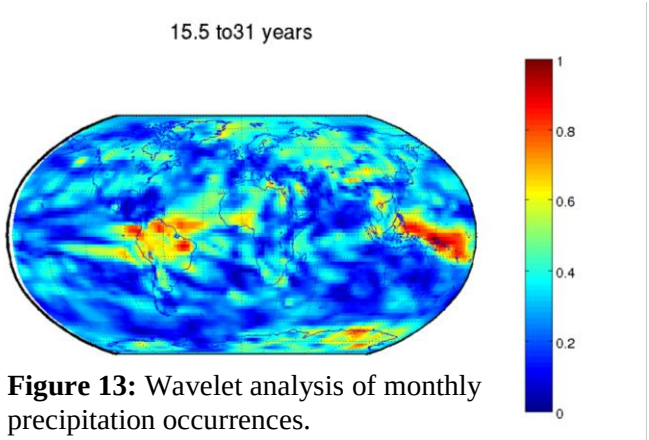


Figure 13: Wavelet analysis of monthly precipitation occurrences.

the well-known African Sahel drought south of the Sahara desert, we see some patterns over Brazil extending southwards into the Atlantic Ocean and another pattern off the coast of Indonesia, Papua New Guinea and over Solomon Islands. These two patterns are known as the South Atlantic Convergence Zone and the South Pacific Convergence Zone, respectively. However, the present results are the first indicating that these zones also exhibit an oscillatory character with periods roughly between 16 and 32 years.

3.3 Mining Intervals of Change Events in Climate Data

Contributors: Liess (R-UMN), Shekhar (F-UMN), Zhou (G-UMN), Jiang (G-UMN), Gebril (R-NCAT), Gorji-Sefidmazgi (G-NCAT), Homaifar (F-NCAT), Neelon (U-NCAT), Faghmous (G-UMN), Styles (U-UMN), Gibson (U-NCAT), Mithal (G-UMN), Boriah (R-UMN), Kumar (F-UMN)

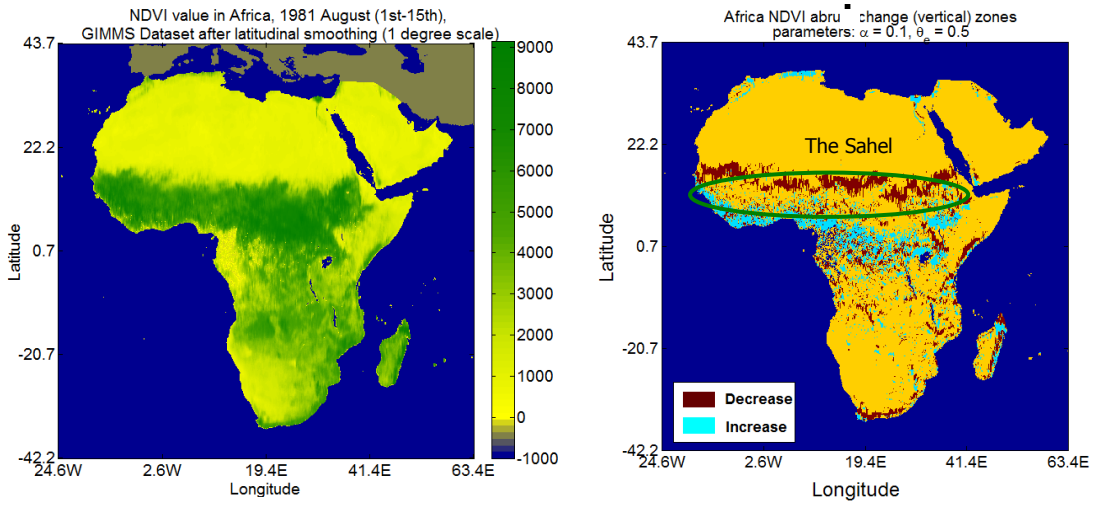
Activities:

Changes that occur in climate data may be persistent in space and/or time and may be representative of a change in the underlying regime or generative process. Changes may further exhibit characteristic interval semantics in their spatial or temporal patterns. Since climate data possesses inherent space-time properties, such changes can potentially be observed by projecting the data to a reduced representation, by either marginalzing over the spatial dimension or the temporal dimension. Such a representation allows for the detection of changes occurring in a particular spatial direction at a given time, or in the temporal dimension at a particular location. Detecting the periods of abrupt changes in such 'time series-like' representations can be useful in many scenarios. For example, abrupt changes in space could indicate sharp transition regions of environment between different ecological zones (a.k.a. ecotones). As these ecotones may be vulnerable in response to climate change, identifying their footprints can help us understand the interactions between climate change and ecological systems. Detecting changes in atmospheric variables, such as temperature and pressure, can help identify local regimes of change and linear time trends in each regime. Further, identifying spatio-temporal patterns in ocean variables, which show persistent changes over time, can be helpful for detecting ocean eddies, which are drivers of marine ecosystem, and in addition to dominating the ocean's kinetic energy, play a significant role in the transport of water, salt, heat, and nutrients.

Findings:

In our work on abrupt land cover change detection, we had developed an interest measure named sameness score for quantifying the consistency of abrupt change, e.g., change direction or change magnitude. We also designed and evaluated a preliminary computational approach, namely, Abrupt Change Interval Miner (ACIM), which enumerated and pruned intervals in a top-down manner exploiting subset relationship among intervals. We further designed and evaluated a Sub-path Enumeration and Pruning (SEP) approach for the discovery of interesting spatiotemporal sub-paths/intervals. We further analyzed spatial autocorrelation level of eco-climate data at multiple resolutions. The proposed approach for abrupt land cover change detection was used to analyze the smoothed Sahel precipitation anomaly data. We discovered a few major abrupt change intervals of precipitation. For example, the interval from 1967 to 1971 is the well-known abrupt decline of precipitation in Sahel. We also identified several abrupt increase intervals (1903-1908, 1944-1953, 1986-1988 and 2008-2010), which are not widely discussed in literature. For example, rainfall in Sahel seems to start recovering since 1986. We discovered a longer interval of abrupt precipitation decrease from 1957 to 1983 which includes the two shorter periods (1968-1971 and 1981-1983) we previously discovered. ACIM was also used to analyze vegetation cover data (see Figure 14). It discovered many ecotones, as shown in Fig. 14(b); an interesting finding is characterization of Sahel of an ecotone, indicated by the green ellipse. Experimental and theoretical analysis showed that both proposed SEP row-wise and top-down algorithms are orders of magnitude

faster than naive algorithm. The top-down design decision has better performance than the row-wise for datasets with longer patterns while the row-wise always has lower memory cost.



(a) (b)

Figure 14. (a) Map of vegetation cover (measured in Normalized Difference of Vegetation Index, NDVI) in Africa, Aug. 1-15, 1981. (b) Abrupt transitions of vegetation detected in Africa (south to north).

We presented a novel change detection algorithm that automatically identifies and monitors eddies in sea surface height data based on heuristics derived from basic eddy properties. To demonstrate its performance we analyzed eddy activity in the Nordic Sea, an area that has received limited attention and has proven to be difficult to analyze using other methods. We further improved upon the state-of-the art connected component eddy monitoring algorithms to track eddies globally. The proposed approach does not pre-process the data and therefore minimizes the risk of wiping out important signals within the data. We employed a physically-consistent convexity requirement on eddies based on theoretical and empirical studies to improve the accuracy and computational complexity of our method from quadratic to linear time in the size of each eddy. We accurately separate eddies that are in close spatial proximity, something existing methods cannot accomplish. We compared our results to those of the state-of-the-art and discuss the impact of our improvements on the difference in results. There is a significant difference in the total number of eddies detected by

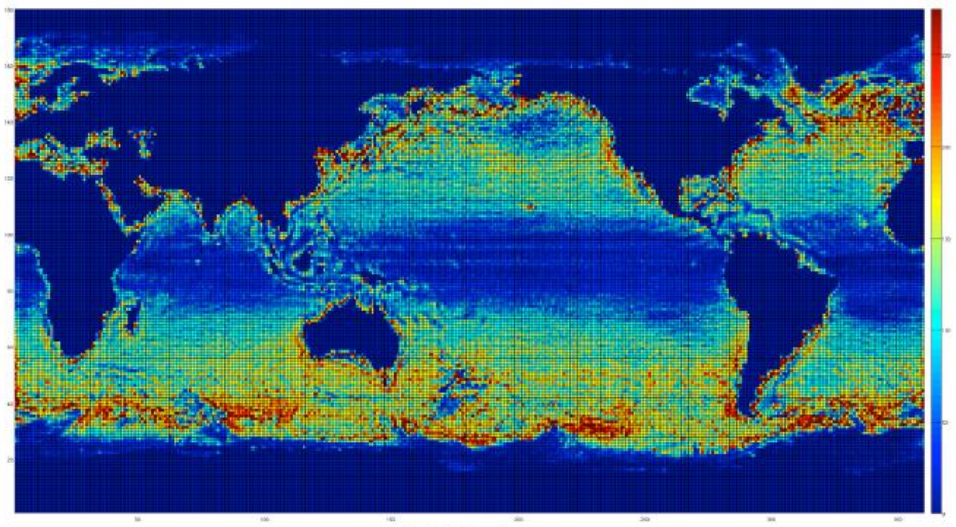


Figure 15: Aggregate counts for eddy centroids that were observed over the period: October 1992 - January 2011

our propsed approach and the previous work most notably at high latitudes and along the equator. Figure 15 shows the aggregated spatial distribution of eddies. It is evident from the eddy centroid distribution that high density regions tend to be along currents (i.e. Gulf Stream (North Atlantic) and Kuroshio Current (North West Pacific)) and in open oceans

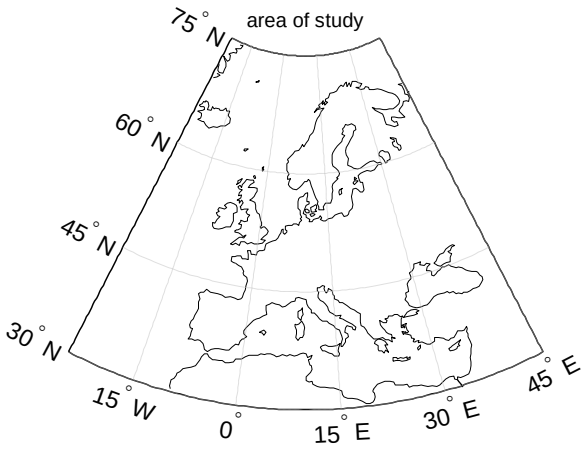


Figure 16. Area of Study

We developed a non-stationary time series clustering approach to identify persistent regimes over multiple atmospheric time series and also construct a dynamical model for each regime. The finite elements method (FEM-K-trend), where K is the number of clusters, is a newly developed technique in the literature. In contrast to FEM methods, most of the non-stationary time series clustering methods assume local stationarity of time series. On the other hand, FEM-K-trend can find clusters even when there is a time-trend in each regime without any explicit assumption about probabilistic models (e.g., Gaussian, etc.). In this method, instead of finding local stationary regimes, the concept of *model distance functional* is defined and a convex linear combination of these functionals is used to present the dynamical model of the time-series. We applied FEM-K-Trend for clustering of skin temperature (temperature of the surface at a radioactive aquarium) which belongs to the area shown in Fig 17 for a time period of 5 years from 1970 to 1975. The resulting time series has 558 dimensions and around 1800 temperature measurements (one datum per day). The simulation results show the switching times between the two clusters and also a dynamical model of temperature in each of two clusters and each of 558 dimensions. The results are shown in Figure 17. FEM-K-trend needs to define a specific type of model distance functional such as Euclidean distance. In meteorological systems, there are many variables that have coupling amongst each other. FEM-K-trend can help us to understand the relation between clusters of these coupled variables.

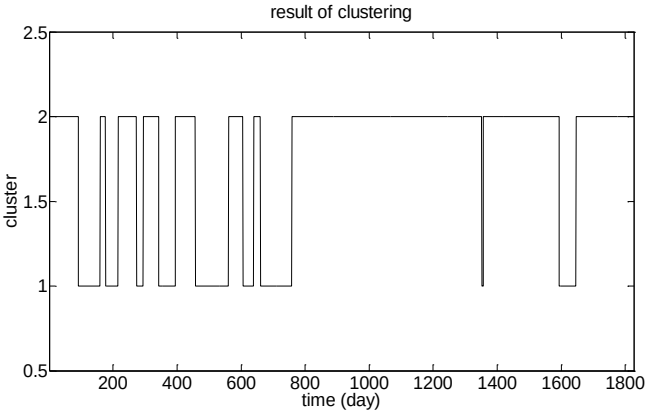


Figure 17. Result of Clustering

4. CLIMATE EXTREMES AND UNCERTAINTY

We define climate extremes inclusively as severe weather or hydrological events as well as relatively large changes in regional hydrometorological patterns which are caused by natural climate variability or climate change with potentially significant consequences on critical infrastructures, key resources or human lives and economy. Our research and education activities span the spectrum of this broadly construed climate extremes area, with a particular focus on generating predictive insights, along with their uncertainties, that are credible at spatiotemporal scales relevant to resource managers and policy makers. The computational methods developed are data-intensive, where data in this context includes weather or climate related observations, archived climate model simulations, as well as information relevant for multi-sector consequences of climate and global change.

Project Highlights:

One project has explored / developed methodologies and techniques in three areas: (a) characterization of extremes and trends, (b) uncertainty characterization and impacts, (c) enhanced predictive modeling. Work in the first area has resulted in a recent (2012) paper published in Nature Climate Change, where we identified significant increasing spatial variability in the trends of rainfall extremes. In the second area, a Bayesian model aiming to combine multiple global climate models for quantifying uncertainty in regional mean temperature and precipitation extremes has been partially developed. In the third area of predictive modeling, our research has shown the applicability of sparse regression to climate extremes. In a recent Statistics and Data Mining conference paper, the value of sparse group lasso regression is shown through its predictive error reduction and via new proofs on consistency under certain assumptions. Our recently published work seeks to infuse physics-based insights into data driven models for predicting extremes and reducing their uncertainty using massive datasets.

Another of our major efforts in this area is multivariate quantiles, which is driven by the application to climate extremes such as hurricane and extreme precipitation. The main idea behind this line of study is that quantiles can be defined as minimzers of convex functions, and this definition lends itself to a generalization to a multivariate, high-dimensional framework. We created a general framework that provides guidelines to specific extreme value models by specifying the objective function to be a certain type. We now have a complete understanding of the scenarios when the tuning parameter is in a certain range and the sampling distribution for the quantiles converges to a Gaussian distribution. Simulations have been conducted to support the existing results. The ideas generalize naturally to multivariate settings, for regression quantiles, other applications, and for inference.

Individual Project Reports:

4.1 Novel Methodologies for Characterizing Extremes and Uncertainty

Contributors: Breidenbach (U-UMN), Chatterjee (G-UMN), Das (R-NEU), Das (U-NEU), Ganguly (F-NEU), Kodra (G-NEU), Kumar (F-UMN), Steinbach (R-UMN), Steinhaeuser (R-UMN), Tolen (G-NEU)

Activities:

The novel methodologies and techniques that have been explored/developed can be roughly categorized into two broad areas from a data sciences perspective: (a) Characterization of extremes and trends , (b) uncertainty characterization and impacts.

Findings:

(a) Characterization of extremes and trends: In our recent (2012) paper published in Nature Climate Change, our research infused new insights to a debate on the trend of 20^{th} century rainfall extremes in India. Although prior literature had disagreed on the nature and direction of the trends, we did not find a homogeneous trend in extremes over India. Rather, we identified significant increasing spatial variability in the trends of rainfall extremes. Although we did discuss physical mechanisms for this insight, we plan on exploring these through a combination of regional climate modeling and statistical experimental design efforts (e.g., Figure 19). This will enhance ongoing collaboration with a team at the University of Nebraska as well as opportunities for new research assistant at Northeastern University. In a manuscript recently submitted to Nature Climate Change, we exemplify the difficulty in characterizing even one class of climate extremes, namely those related to hot weather (Figure 20). (b) Our broad work on uncertainty characterization and impacts feeds into climate extremes as well. In one ongoing work, a Bayesian model aiming to combine multiple global climate models for quantifying uncertainty in regional mean temperature and precipitation extremes has been partially developed. The characterization of uncertainty for impacts continues to be a theme that plays an important role in many works, including a recent

publication in Environmental Research Letters (2012) on multimodel uncertainty in projections of the Indian monsoon as well as ongoing work on uncertainty in fresh water availability.

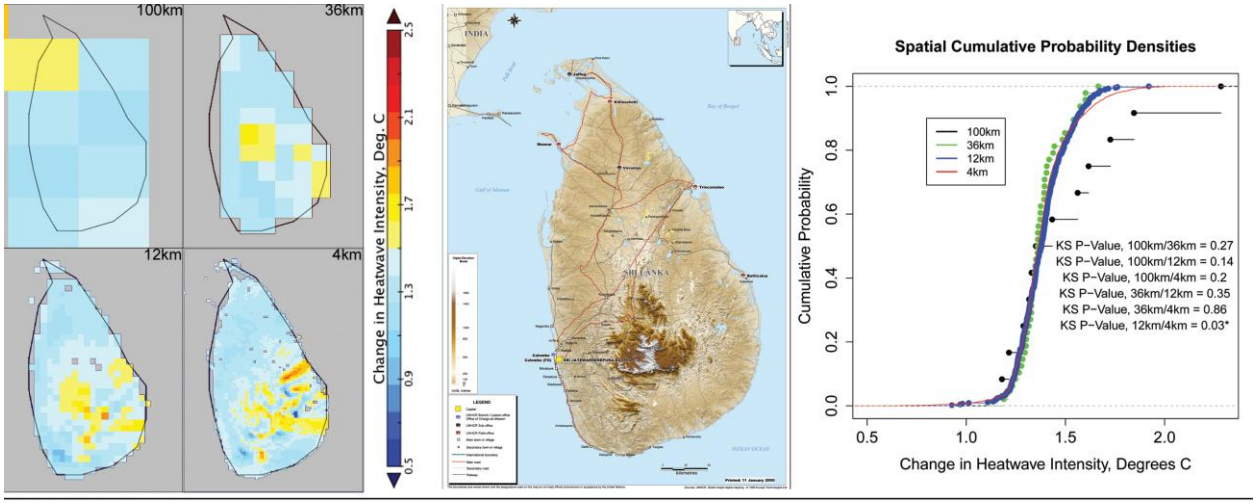


Figure 19: Dynamical downscaling may be crucial for carefully-designed hypothesis testing. Here we show changes in heatwave intensity19 from 2006-2015 to 2056-2065 is calculated using the Community Climate System Model 4.0 (CCSM4) run with RCP 8.5 (i.e., a specific global climate model with a specific greenhouse-gas emissions scenario) dynamically downscaled over Sri Lanka through a regional climate model (RCM). The Weather Research and Forecasting (WRF) model was used as the RCM, which in turn was run at 36km, 12km, and 4km resolutions, respectively, by using the lower-resolution CCSM4 outputs at boundary conditions. A goodness-of-fit comparison, via the Kolmogorov-Smirnov (KS) tests, does not yield substantial evidence for differences in spatial distributions of model runs, which is probably owing to small sample sizes for the 100 and 36km resolution data. However, the effects of topography in the mid-southern Sri Lanka appear more prominent at lower resolutions. The sheer size of the newly-generated dynamically downscaled simulations, as well as the complexity of varying temporal and spatial scales, presents significant challenges and opportunities for collaborative efforts between data scientists and climate modelers. Hypothesis testing with climate modeling and statistical experiment design is an avenue we are actively pursuing.

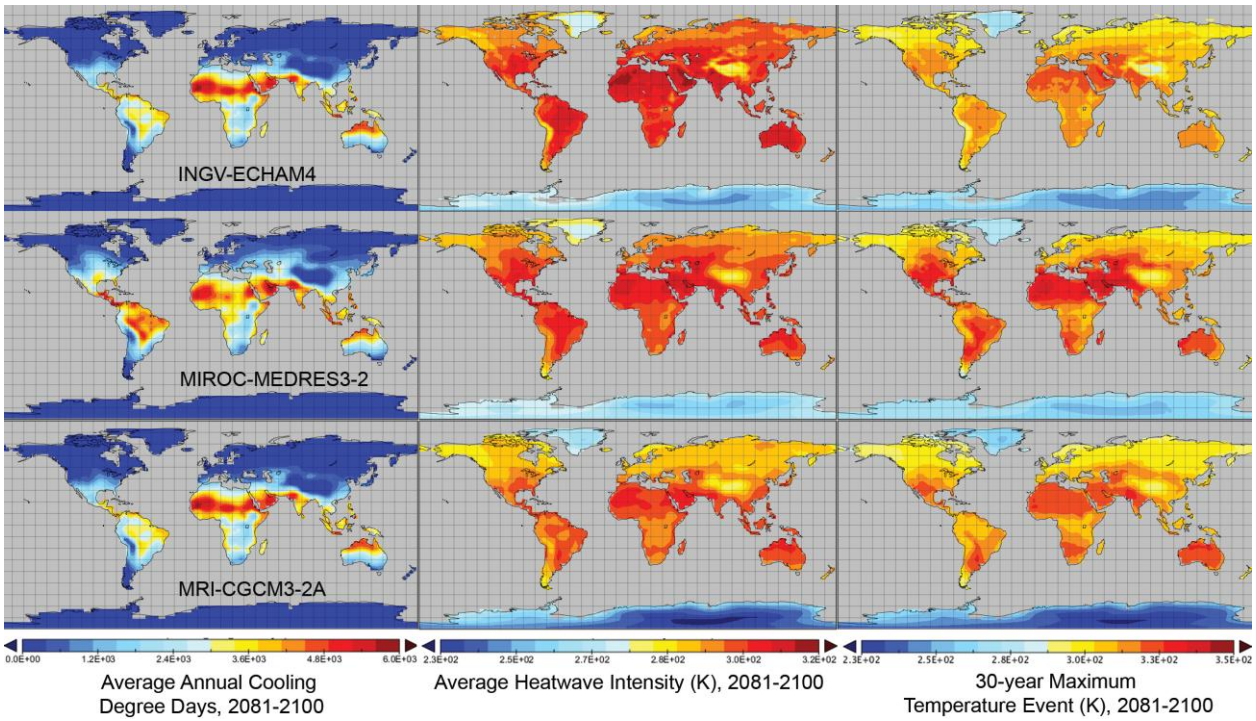


Figure 20: Characterization of extremes requires a summarization of the statistical properties of climate-related observations and model-simulations. A thorough assessment of extremes from massive climate data may become especially challenging when the extremes definitions and indices themselves need to change depending on stakeholder needs. Here we present three different choices: an energy-consumption related metric called Cooling Degree Days26 or CDD (left), an index thought to be relevant for human mortality defined as consecutive nighttime minima events (middle), and a third which happens to be grounded in the statistical theory of extreme values (right). The substantial regional differences suggest the differences in the nature of the insights.

4.2 Physics-guided statistical association models for climate extremes

Activities:

Enhanced predictive modeling research has shown the applicability of sparse regression to climate extremes. In a recent Statistics and Data Mining conference paper and in submission to Nature Climate Change with a variety of collaborators, the value of sparse group lasso regression is shown through its predictive error reduction and via new proofs on consistency under certain assumptions. Three recently accepted conference papers have explored data mining methods, opportunities, and their potential value for extremes. This line of work seeks to infuse physics-based insights into data driven models for predicting extremes and reducing their uncertainty using massive datasets.

Findings:

In Figure 21, we examine spatial traits of extreme precipitation in the United States. Ultimately, we aim to test the hypothesis that combining data driven methods with physical constraints may lend new understanding on how extremes behave with their covariates that may not be captured in current physical climate models. For example, in Figure 22, we present an exemplary visualization that may be helpful for examining dependency among rainfall extremes and variables thought to dictate their behavior. Complex networks have been shown to capture multivariate and multiscale dependence in the climate system and produce predictive insights for land climatology based on teleconnections with ocean-based indices. We

have been exploring the possibility of delineating the improvements possible over physics-based models and domain knowledge. Finally, we are leveraging funding from the Nuclear Regulatory Commission to review the state of the art in stochastic rainfall generators toward quantifying structural risk for flooding. This effort is synergistic with multiple lines of work in the NSF Expedition and also strengthens collaborative relationships with Oak Ridge National Laboratory as well as the University of Tennessee. The focus of this project is on improving the handling of extremes and means of rainfall all in one seamless framework, leveraging information content from various sources of data and multiple co-varying climate variables. For all of the above, high performance may be leveraged to significantly speed up and even enhance insights. For example, in Figure 23, insights obtained from our recent Nature Climate Change paper on the increasing spatial variability Indian rainfall extreme trends. The computational speed up is nearly linear and the insights apparently visually clearer than those obtained in the recent paper. Although this is a simple example of high performance computing applied to a parallel-friendly problem, advances in and applications of HPC will be more critical for complex computational climate problems.

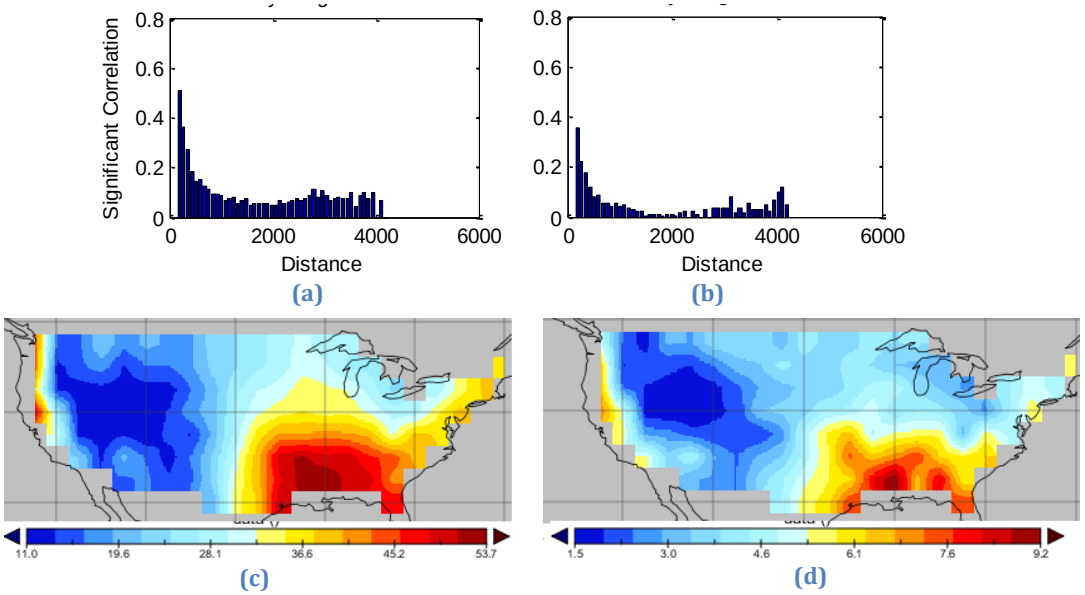
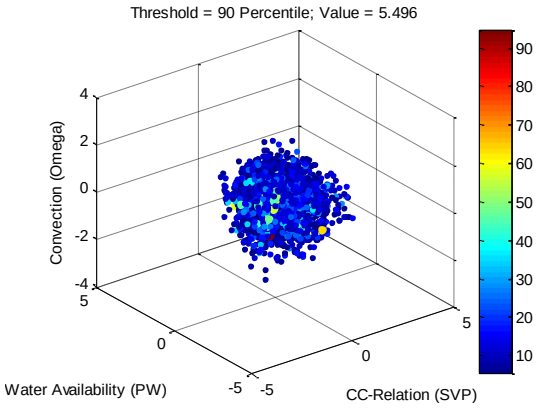


Figure 21: *Distribution of correlation among annual maxima precipitation as a function of distance for (a) reanalysis and (b) direct observations; the spatial distribution of (c) location parameter and (d) corresponding confidence interval for Generalized Extreme Value distributions estimated at each grid-point.*

Figure 22: Plot of observed precipitation extremes (events over the 90th percentile of wet days, gathered from the Climate Prediction Center) with respect to indicator CC-relation, convection and water availability. Colors and size of dots are proportional to the intensity of the extremes.



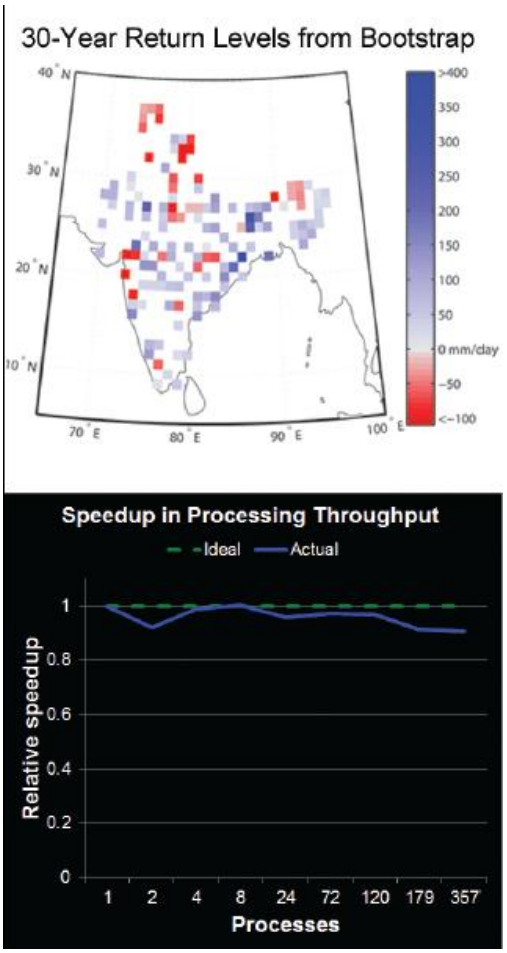


Figure 23: (Top Panel) The simple bootstrap is applied to annual maxima of gridded rainfall extremes over India, and in turn Generalized Extreme Value distributions are estimated for each bootstrap iteration. Bootstrap-mean estimates of 30-year return values are estimate. (Bottom panel) High performance computing allowed for a linear speed up in the bootstrap procedure, allowing thousands of iterations to be performed.

4.3 Multivariate quantiles and convex minimization

Contributors: Chatterjee (F-UMN), Yuan (G-UMN)

Activities:

The main idea behind this line of study is that quantiles can be defined as minimizers of convex functions, and this defintion lends itself to a generalization to a multivariate framework, including high-dimensional framework. We also focussed on studying a general extension work for extreme quantiles. Our starting point is to consider a certain proportion of extreme value instead of the largest or the smallest one alone. This study is driven by the application to climate extremes such as hurricane and extreme precipitation. The research for the top k order statistics is not only about getting more information regarding to the tail behavior of the underlying distribution, but also serve as must for studying the change point related problem. Methodology-wise, we worked on a general framework which can easily provide guidelines to specific extreme value models just by specifying the objective function to be a certain type.

In essence, we consider the function $\varphi(q, \alpha) = EX1\ \varphi(X1 - q, \alpha)$; where α is an indexing parameter, which in case of univariate quantiles may be taken to be a probability. Here, the function φ has certain convexity and other interesting properties, compatible with the kind of parameters we wish to study. In order to consider extreme value cases, we consider a sequence of indexing parameters, αn, and consider a framework where $\alpha n \rightarrow \alpha$, and they are both in [0,1]. Standard assumptions include, for example, that φ is Lipscitz in the second argument.

Findings:
By solving a sequence of convex functions, we obtained a sequence of M -estimators that eventually converge to different distributions according to difference choices for the tuning parameter. We now have a completely understanding of the scenarios when the tuning parameter is in a certain range, and the sampling distribution for the quantiles converges to a Gaussian distribution. Simulations have been documented to support the existing results. Future analysis is primarily motivated by learning what the asymptotic distribution is if the tuning parameter is not in the range where the sampling distribution is not in the domain of attraction of the Normal law. The ideas can generalize naturally to multivariate settings, for regression quantiles and other applications, and for inference.

We have established the following results:

Theorem: *Let us consider the collection of functions $F=\{\varphi n\ (q, \alpha n\ ):$ n is positive integer$\}$. F is almost surely equicontinuous and pointwise bounded, hence by Arzela-Ascoli theorem, for every compact set S in R, almost surely every sequence in F has a further subsequence that converges uniformly on S.*

Theorem: *Under a further set of technical conditions, $\varphi n\ (q, \alpha n\ ) \rightarrow \varphi(q, \alpha)$. Moreover, $\sup q \in S\ |\varphi n\ (q, \alpha n\ ) - \varphi(q, \alpha)| \rightarrow 0$ over any compact set S of R. Also, if the minimizer of $\varphi(q, \alpha)$ is unique, and if $q = +\infty$, then $qn \rightarrow +\infty$.) If q and qn are both bounded, then $qn \rightarrow q$. It is impossible that q is bounded, qn is not bounded once we further assume $\varphi(xi - q, \alpha n\ )$ is convex w.r.t. q.*

These results are being followed up with further theoretical developments to study the properties of quantiles near the tail of a distribution. These are related to the probabilistic and statistical properties of extreme values.

5. MULTI-MODEL ENSEMBLES

Given a limited number of global climate model runs and available observational data, effective utilization of information from both sources is important for better assessing the credibility and uncertainty of model projections as well as for revealing directions toward model improvement.

Accomplishment Highlights:

Much of our work in this area has addressed how researchers should evaluate and weight models to characterize uncertainty in projections. Our work with Bayesian approaches to combining multi-model projections is ongoing, focusing primarily on the infusion of physical constraints for informing model weights. A Bayesian model that combines multiple global climate models for quantifying uncertainty in regional mean temperature and precipitation extremes has been partially developed and we plan a comprehensive evaluation. In additional work, we are developing new approaches for model selection that addresses issues that arise when none of the models capture the actual physical process accurately and when only limited or complex data is available, as in the case of modeling rare events, regional and small area problems.

Another key task is the comparison of GCMs to one another, to the ensemble prediction, to observed data, and to previous versions of GCM models We used an exploratory spatial data mining approach to discover the geospatial footprint of model disagreement and model ensemble divergence on surface air temperature between nine different GCMs' outputs, as well as between an ensemble prediction of the models and with sensor based (reanalysis) data. We found that the GCMs often disagree with each

other and that even when they agree; they may disagree with sensor based data. We also investigated the pervasive notion in climate science that, when using past observed climate as a reference for skill evaluation, equally weighted averages of multiple climate model outputs outperform individual model outputs. In studies of Indian Monsoon rainfall and precipitation in the Western U.S., we demonstrated that while model consensus may be valuable in some regions (e.g., the western United States), overreliance on consensus or multi-model averaging may be misleading because (a) models often disagree strongly, implying limitations for consensus-driven decision-making and (b) multi-model averages may obscure important insights on variability and uncertainty in projections. Finally, our comparison between older and more recent GCMs suggests that multi-model regional variability has not decreased significantly since the development from the CMIP3 to the CMIP5 suite.

Individual Project Reports:

5.1 Climate model selection and ensembles using Bayesian statistics

Contributors: Contributors: Banerjee (F-UMN), Chatterjee (F-UMN), Chatterjee (G-UMN), Ganguly (F-NU), Kodra (G-NU)

Activities:

The aim of this research is to combine global climate model outputs for predicting future climate. The prediction is over a long time horizon, and may be either on a global, a large regional, or a small regional scale. The efficiency, accuracy, precision of various constituent models, and of different techniques for assembling them together is under study. Part of this work is continuation of our earlier and ongoing research on statistical model selection and averaging, online learning and related techniques.

Findings:

Multimodel ensembles are widely used but there is little consensus as to whether and how researchers should evaluate and weight models to characterize uncertainty in projections. We have pointed out the difficulty in evaluating ensembles for low-frequency patterns related to the Indian monsoon. Specifically, in one case a physical explanation was found in literature to support one model's skillful reproduction of temporal rainfall patterns. However, initial conditions were found to have a sizable impact on historical skills of models, raising questions on potential irreducible intrinsic climate uncertainty. Our work with Bayesian approaches to combining multimodel projections is ongoing, focusing primarily on the infusion of physical constraints for informing model weights. Specifically, a Bayesian model aiming to combine multiple global climate models for quantifying uncertainty in regional mean temperature and precipitation extremes has been partially developed. In that same vein, we aim to soon begin work on a comprehensive evaluation of historical model skill and future consensus via multiple metrics with multiple variable including extremes. This may help build a path toward effective model combination and weighting. In Figure 24, which was part of a review paper submitted to a leading journal, we examine the observed dependency between mean temperature and precipitation extremes as it compares to the theoretical Clausius Clapeyron (CC) relation. The degree to which the two variables behave as dictated by the CC is region dependent and also varies by dataset (modeled or reanalysis). Such insights may help inform statistical models for uncertainty quantification by infusing them with physical insight.

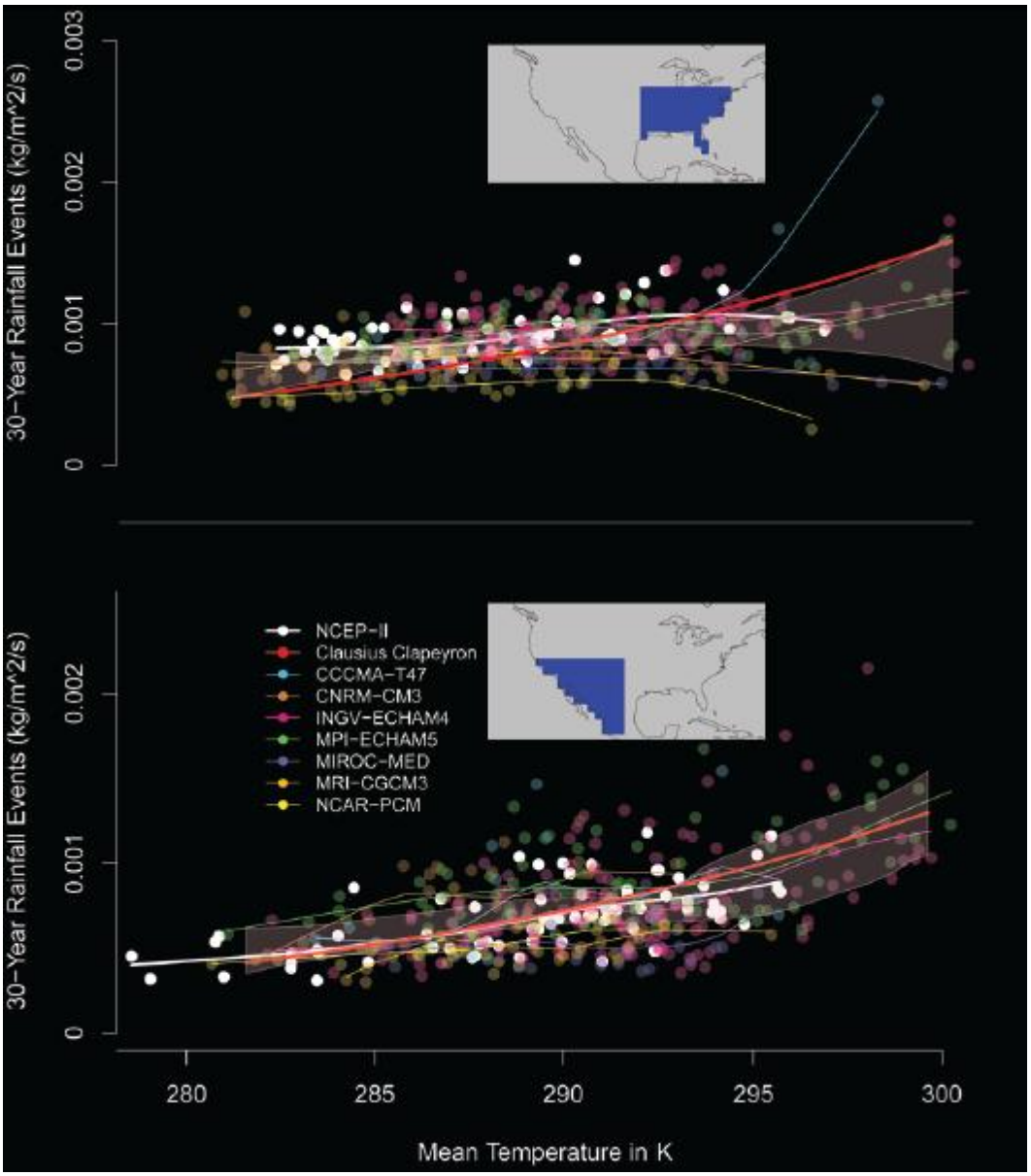


Figure 24: Multimodel ensembles have been used to quantify uncertainty in the structural representation of climate physics; their performance has been evaluated by investigating skills in reproducing historical behavior (skills) and multimodel agreement (convergence) in the future. Here we investigate the uncertainty in precipitation extremes and explore whether physically based relations, like the temperature-dependence of precipitation extremes through the saturation vapor pressure (known as the Clausius-Clapeyron, or CC, relation), may help further inform uncertainty assessments and skill-based model selection. For the southwestern and southeastern United States, a 7-member CMIP3 model ensemble is used for the analysis, with NCEP2 used as a baseline model and the theoretical CC curve shown for comparison. Every point from each model represents a 20-year mean temperature (1980-1999) on the x-axis and a 30-year rainfall (1980-1999, y-axis) with nonlinear regressions fit to each dataset and uncertainty bounds computed using a bootstrap-based resampling procedure. The value of using the multivariate physically-based CC relation in uncertainty quantification is suggested, particularly for extremes (specifically, heavy rainfall) where covariate relations (specifically, temperature-dependence) are known from process physics (e.g., Clausius-Clapeyron).

5.2 Climate model consensus study and regional climate modeling

Contributors: Banerjee (F-UMN), Chatterjee (F-UMN), Chatterjee (G-UMN), Ganguly (F-NEU), Kodra (G-NEU), Mukherjee (Usha Martin Academy)

Activities:

While model evaluation selection is important, one model may not fully capture plausible regional climate change. Thus, it is of interest to effectively synthesize information from multiple models to obtain distributions of change. Bayesian schemes have been developed that attempt to weight climate models based on past climate model skill and future model consensus to obtain distributions of change. One pervasive notion in climate science is that, when using past observed climate as a reference for skill evaluation, equally weighted averages of multiple climate model outputs outperform individual model outputs. Evidence in support of this hypothesis is mostly empirical. We have shown that this is not always the case.

Findings:

A case study revealed that one global climate model output is superior in modeling historical Indian monsoon rainfall periodicity to any other equally weighted average of 7 models. This is the case because this model captures related climate processes well, while the other models do not. In this same paper, however, we show that there is not one clearly superior model for Indian maximum temperature trends. This suggests that climate models must be evaluated and selected on a case-by-case basis, as in some cases a single model may be most appropriate, and in others, it may be useful to use multiple models. In addition, intrinsic system uncertainty was revealed via variability resulting from initial conditions.

In one ongoing work, we explore uncertainty in 21st century United States water availability using two ensembles of global climate models from different development generations as well as two scenarios for 21st century population growth. Results obtained here suggest that while model consensus may be valuable in some regions (e.g., the western United States), overreliance on consensus or multimodel averaging may be dangerous from a policymaking perspective. This is because (a) models often disagree strongly, implying limitations for consensus-driven decision-making and (b) multimodel averages may obscure important insights on variability and uncertainty in projections.

In another ongoing work, we are comparing projections and historical runs for mean temperature and precipitation patterns from the previous generation CMIP3 suite of global climate models to the current CMIP5 suite. Results thus far (Figure 25) suggest that multimodel regional variability has not decreased significantly since the development of the CMIP5 suite, which may emphasize the need for uncertainty reduction methodology outside of solely physical modeling. Initial analysis (Figure 26) implies that even successive generations of model developed at the same institution can vary significantly even if the multimodel mean does not.

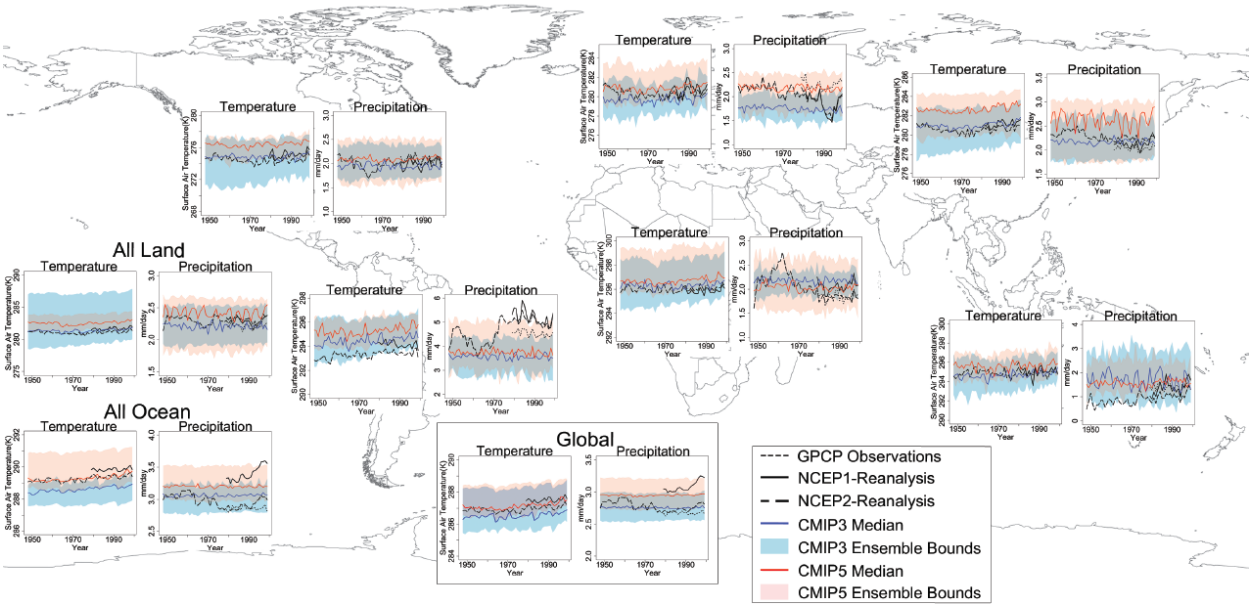


Figure 25: The next- (CMIP5) and current- (CMIP3: IPCC-AR43) generation climate model simulations compared at continental and global scales in terms of average temperature and precipitation over time using reanalysis data (NCEP1 and NCEP2), or surrogate observations, as baselines. Note: CMIP = Climate Modeling Intercomparison Project; NCEP = National Center for Climate Prediction.

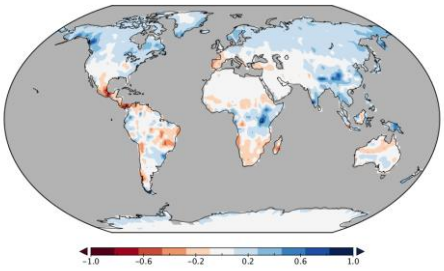
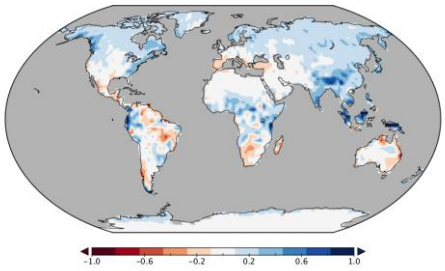
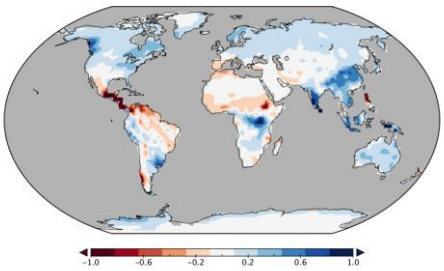
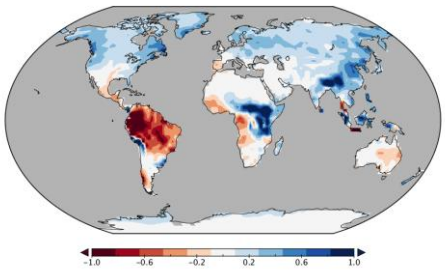
Model ID: CMIP3/ CMIP5	CMIP3 (Future-Past)	CMIP5 (Future-Past)
Multi-model Ensemble		
CCCMA-CGCM3-1-T63/ CanESM2		

Figure 26: (Top Row) Multimodel mean output from previous generation (CMIP3) to current generation (CMIP5) global climate models do not appear to differ substantially. (Bottom row) However, successive generations of models developed from the same institutions exhibit large differences in magnitude and even sign of projected precipitation change.

5.3 Understanding geospatial epistemic uncertainty patterns in global climate models

Contributors: Liess (R-UMN), Semazzi (F-NCSU), Shekhar (F-UMN), Xun (G-UMN), Jiang (G-UMN)

This project ended in Year 1.

Activities:

Given physics-based global climate model (GCM) outputs and sensor-based reanalysis data, the problem of geospatial uncertainty characterization aims to discover interesting, previously unknown but potentially useful spatial patterns of epistemic uncertainty in climate data. This problem is important because it could help to improve GCMs. For example, it could identify the geospatial footprint of model disagreement, model ensemble divergence from sensor-based reanalysis data, or both. Such geo-location may help identify missing regional physics from GCMs and thus spur model refinements. However, this problem is challenging due to the large volume of data, multiple types of uncertainties (e.g., within a single GCM, among multiple GCMs, between GCM ensemble and reanalysis data), as well as spatial and temporal heterogeneity. We employ an exploratory spatial data mining approach to discover spatial footprints of epistemic uncertainty. More specifically, we compute a standard deviation map (with degree by degree geospatial grids) from nine different GCMs' output on surface air temperature. Then we compute the difference map between multi-GCM ensembles and reanalysis data. Two GCM ensemble methods are used, i.e., multiple model averaging and linear regression.

Findings:

1. There is spatial structure in both standard deviation map and difference maps. More specifically, both the standard deviation across multiple GCMs and the difference between GCMs ensemble and reanalysis data show geospatial clustering instead of complete spatial randomness. Values at one location are not independent of those at nearby locations.
2. GCMs show disagreement in many regions as illustrated in Fig. 27. There are high standard deviations of GCMs in polar and tundra area, western coast of North America, north India, as well as southern Latin America. It might imply that more accurate regional models need to be developed in these areas.
3. Even if multiple models agree with each other in some regions, they may be different from sensor-based reanalysis data. For instance, in Fig. 28, we see that there are differences between GCM ensembles and reanalysis data in west coast of Africa, South America and Australia (shown in circles). Those areas are also upwelling areas. This might indicate that these GCMs are not including physics of upwelling phenomenon in certain areas.

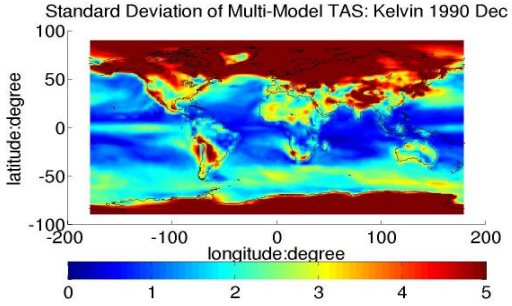


Figure 27: Standard Deviation Map of multi-GCM

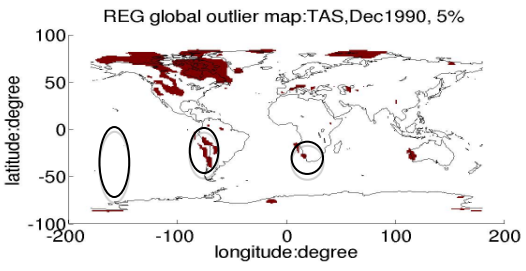


Figure 28: top 5% global outlier in errors of GCM ensemble

5.4 Statistical Model Selection and Averaging

Contributors: Banerjee (F-UMN), Chatterjee (F-UMN), Ganguly (F-NEU), Kodra (G-NEU), Nandy (G-UMN)

Activities:

A number of topics are under study in this broad category. One new approach we are currently studying involves model selection using a bias and variance trade-off. Here we focus on devising a method for the selection of models, including but are not limited to regression model selection, covariate selection, random and mixed effect models, bandwidth selection or non-parametric and semi parametric curve estimation. The problem of model selection is a well-studied problem in both the Frequentist and Bayesian paradigms working reasonably well in regression setup, but not for more complex models that we are studying. Our aim is to extend the notion of model selection from regression models to a wider class of problems that satisfy some regularity conditions. Also the other aim of this study is to conceptualize rigorously the ideas of variability and bias, to be able to use them as criterion for selecting models based on given data. Another aspect of this problem is that sufficient emphasis has been laid on the applications of this novel methodology on climate data, especially problems related to model selection in extreme-value modelling, with application to precipitation.

A different study is ongoing on climate model ensembles, and yet another study on model selection where limited data or complex data is available, as in the case of modelling rare events, regional and small area problems, and where none of the models capture the actual physical process accurately. A component of this study is the relationship between the minimum central pressure and the maximum sustained winds in tropical storms.

Findings:

We have investigated thoroughly the model selection approach in a limited linear regression framework, and have completed a considerable part of the study in multiple linear regression framework. The most general model framework where this methodology might be applicable has been developed, and a branch-off study on model selection in small-area context is now nearly complete.

Prior research suggested that the relationship between minimum pressure and maximum wind speed in a tropical cyclone can be estimated by the physical model:
MaxWindSpeed=δ (μ - MinPressure)^ξ where *δ=200^1/2, μ=1013, and ξ=0.5.* We have examined the shortcomings of this model and improved upon on the above parameter estimates via least squares and maximum likelihood estimation. Additionally, we investigated modeling each variable by extreme value distributions and used a generalized extreme value (GEV) model relating *MaxWindSpeed* and *MinPressure*. The logarithm of both sides of the proposed physical model was taken, and *μ=1013* was assumed while *δ=200^1/2* and *ξ=0.5* were examined. Let Y_i=log(*MaxWindSpeed*) and x_i=log({1013-*MinPressure* }). We tested the linear model $E[Y_i|x_i]=\log(\delta)+\xi$. Diagnostic plots indicated assumptions did not hold. Since assumption violations indicated a higher order term may improve the model, a quadratic term was added which created the new model $E[Y_i|x_i]=\beta_0+ \beta_1 x_i+ \beta_2 x_i^2$. Diagnostic plots indicated that the assumptions of the linear model were violated, however, when compared to nonparametric methods, this model was qualitatively good. Maximum likelihood estimation (MLE) utilized the log scale and assumed that $\log(\{MaxWindSpeed\}) \sim N(\log\{\delta\}+\xi \log\{(\mu- MinPressure)\},\tau^2)$ and that μ=1013, MLE were evaluated and a Likelihood Ratio test (LRT) was done to compare δ and ξ to the initial physical estimates. The physical model estimates were rejected with a p-value of $2.4X10^{-12}$. Dropping the assumption that μ=1013, MLE were evaluated and a LRT was done to compare μ, δ and ξ to the initial physical estimates. The physical model estimates were rejected with a p-value of nearly 0. The wild bootstrap was applied to compare the MLE estimates of ξ and δ (assuming μ=1013). The 2-sided empirical p-values for ξ and δ were 0.0284 and 0.0004 respectively indicating strong evidence for parameters not equal to those suggested by the physical model.

We chose to fit extreme value distributions more reasonable for log({*MaxWindSpeed*}) and log(*MinPressure*). A generalized extreme value (GEV) distribution was most appropriate for modeling the variables. A GEV was fit with a mean modeled by $x_i^T \beta$ where x_i=log(1013- *MinPressure*) and β is unknown. Diagnostic plots indicated that this model fit reasonably except for one extreme outlier. This research has shown the physical model is a good first order approximation of the relationship between the variables, however, due to the heteroskedastic nature of the data, more sophisticated modeling techniques are necessary. Further research will include modeling the dynamics of the bivariate distribution of the log({*MaxWindSpeed*}) and log({*MinPressure* }) via copulae as well as extension of these results including further bootstrapping.

6. High Performance Tools and Methods

Since the worldwide volume of climate data is expected to increase roughly 1000-fold over the next 10-20 years, it will be imperative to make use of high performance computing (HPC) solutions in order to process and analyze this data. In Year 2 of the Expeditions project, we have extended our previous work on high performance data analytics kernels, as well as technologies for compressing and querying huge datasets and for performing similarity searches. We have released software for this work, which is discussed further in the Outreach Activities. Finally, in addition to the published work discussed in this section, we are also actively collaborating within the Expeditions project to develop HPC solutions for work described in other sections of the report, including bootstrapping methods for extreme value prediction and MRF-based abrupt change detection.

Thus, the overarching goal of this HPC activity is to enable

- *Higher spatial or temporal resolution* that is required by (a) precipitation extremes analysis; (b) network-based hurricane prediction; and (c) estimation of spatiotemporal dependencies.
- *Higher data dimensionality* that is critical for (a) Bayesian analysis of multi-model ensembles; (b) sampling-based statistical methods; and (c) multivariate quantile analysis.
- *Greater complexity per data point* that is the key in (a) estimation of complex dependence structures; (b) handling non-stationarity; and (c) multi-resolution analysis.
- *Shorter response time* that is important for interactive "what-if" hypothesis testing.

Accomplishment Highlights:

Efforts in this area are focused on creating algorithms of broad applicability: high performance analytical kernels and spatio-temporal data management tools. For kernels, we have created a library of common data mining tasks that includes k-means clustering, fuzzy k-means, and principal component analysis, implemented on GPUs using CUDA. More recently, we have expanded the kernel library to include highly scalable data clustering algorithms, including a novel k-medoids algorithm, AGORAS. These algorithms have shown speedups of two to three orders of magnitude. With respect to data management, we have developed the ISOBAR code and theoretical performance model to provide a predictive, scalable and power-efficient implementation of this strategy. To support analytics-driven efficient query processing, we developed the ISABELLA and ALACRITY codes. ISABELLA supports indexing of data with lossy compression, whereas ALACRITY supports indexing of data with lossless compression. Both ISABELLA and ALACRITY offer significant data storage reduction and a more than 10-fold speed-up. In another indexing effort, we have developed an image indexing technique based on a new Locality Sensitive Hashing (LSH) scheme. The proposed LSH technique can be used for efficient image search, and an experiment using real data has shown storage and computation improvements of up to 50%. In addition, we have developed a classification algorithm using semi-supervised support vector machines, and probabilistic latent semantic analysis. In another project, we have developed an approach to handle underdetermined problems, i.e., problems with many more features than data points, we developed

BENCH (Biclustering-driven ENsemble of Classifiers), an algorithm to construct an ensemble classifiers through concurrent feature and data point selection guided by unsupervised knowledge obtained from biclustering. We published this work in the premier AI conference, IJCAI 2011.

Individual Project Reports:

6.1 Data Analytics Kernels

Contributors: Agrawal (R-NWU), Choudhary (F-NWU), Hendrix (P-NWU), Liao (R-NWU), Pansombut (G-NCSU), Patwary (P-NWU), Rangel (G-NWU), Samatova (F-NCSU)

Activities:

The goal of this activity is to develop a set of scalable data analysis kernels as part of an effort to provide a library of high performance implementations of standard data mining kernels. We envision this library including a number of functions for several common data mining tasks, such as basic statistics, data clustering, feature extraction, machine learning, and anomaly detection.

Findings:

Scalable data clustering: We have developed high performance (OpenMP and MPI) implementations of density-based and hierarchical clustering algorithms. Our parallel density-based clustering algorithm (DBSCAN) has shown speedups of 30.3 using 40 cores (OpenMP) and a speedup of 5,765 on 8,192 processes (MPI). Our parallel hierarchical clustering algorithm (SLINK) has shown speedups of 19.4 on 36 cores and a speedup of 7,105 on 16,290 processes (MPI).

Additionally, we have developed AGORAS, a clustering algorithm related to k-medoids based on sampling the data. AGORAS is unique in that its runtime is dependent on the distribution of the data (which controls the sample size), but its performance is independent of the full data size. We have performed experiments showing that AGORAS outperforms the state-of-the-art CLARANS algorithm by two orders of magnitude on a dataset of just 64,000 points, and this improvement would only increase with the data size.

Scalable and robust ensembles of classifiers: For some real world problems, the task of creating highly accurate classifiers is complicated by the nature of the data being classified. In climate, data on phenomena such as hurricanes may have thousands of dimensions corresponding to locations on the Earth, so the resulting data contains many more features or grid points than data points or hurricane events. Problems with many more features than data points are called underdetermined or under-constrained problems, and these problems are hard to learn. One reason for this difficulty is that some classifiers scale poorly with the number of features in the data. For example, the time complexity of learning the structure and parameters of Bayesian Belief Networks (BBNs) grows super-exponentially relative to the number of features, and thus makes these methods computationally intractable. Moreover, using too many features for classifier training can prevent the algorithm from identifying the truly relevant features.

We developed BENCH (Biclustering-driven ENsemble of Classifiers), an algorithm to construct an ensemble of classifiers through concurrent feature and data point selection guided by unsupervised knowledge obtained from biclustering. BENCH is designed for underdetermined problems. In our experiments, we use BBN classifiers as base classifiers in the ensemble; however, BENCH can be applied to other classification models as well. We show that, compared to traditional approaches, BENCH runs 2-3 orders of magnitude faster. Moreover, BENCH is able to increase prediction accuracy of a single classifier and traditional ensemble of classifiers by up to 15% on real application datasets using various weighting schemes for combining individual predictions in the ensemble.

6.2 Indexing and Query Processing for Data Analytics

Contributors: Arkatkar (G-NCSU), Buaba (G-NCAT), Gebril (R-NCAT), Homaifar (F-NCAT), Jenkins (G-NCSU), Kihn (C-NOAA-NGDC), Lakshminarasimhan (G-NCSU), Samatova (F-NCSU), Shah (U-NCSU), Shendel (G-NCSU), Sohail (U-NCAT)

Activities:

The data generation process of space-time simulation proceeds from one time step to the next and requires the context of only two time steps, while storing data for only one time step on the disk. In contrast, visual data analytics often requires the full context of the available data, not just a single time step. In fact, simulations that are driven by local space-time relationships are largely performed with the purpose of discovering or explaining non-local and large-scale space-time relationships through interactive "what-if" data exploration. Thus, the fundamental differences in data context and heterogeneity of access patterns demand for analytics-driven data management solutions. This necessitates making data analytics and data reduction the first class citizens of data management design and information query processing.

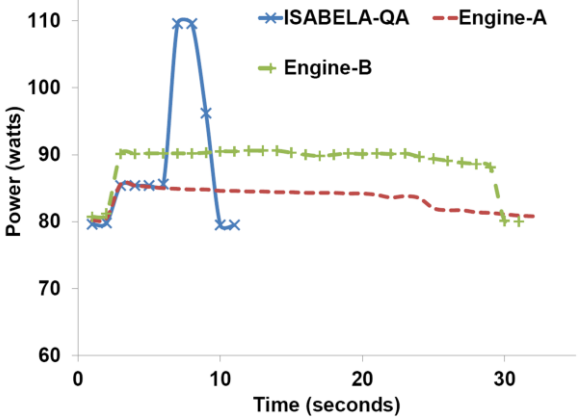


Figure 29: Comparison of energy consumption of ISABELA's query analytics (QA) engine with the state-of-art open source query engines.

Specifically, the problem of image searches in large-scale satellite image data presents a *data-intensive challenge*. A large volume of satellite images with high dimensionality and complexity is being collected, processed and stored periodically by the National Geophysical Data Center (NGDC) in collaboration with National Oceanic and Atmospheric Administration (NOAA). These images come from the Defense Meteorological Satellite Program (DMSP) satellites. The DMSP began in 1991. The visible and infrared sensors collect images across a 3000 km swath, providing global coverage twice per day. Each image is downsized to 363 x 293. Currently, NGDC receives and processes approximately 8.5 GB of satellite imagery data per day from four DMSP satellites.

Table 1. Classification accuracy of sample data from DMSP dataset

Category	SVM	PH+SVM
Aurora	85.6%	89.4%
City light	81.2%	88.7%
Hurricane	89.4%	94.5%

Searching this large dataset to find similar images to a query image using Linear Search (LS) is computationally exhaustive, due to the fact that these images have high dimensionality. In addition, the Linear Search algorithm iteratively compares the query image to all the images in the dataset. Feature selection and classification-based learning offer a promising alternative to traditional indexing for fast searches, especially, searches for similar images.

Data reduction is another key to avoiding I/O bottleneck during data analytics and query processing. Effective lossless compression ensures simulation fidelity, especially at checkpoints. Interleaving lossless compression and parallel I/O is a "secret sauce" for addressing the I/O bottleneck.

Findings:

For indexing and querying multivariate, multi-dimensional time series of scientific floating point data, we have been developing analytics-driven efficient query processing engines, ISABELLA and ALACRITY, that offer a transformative shift from the traditional indexing of data to the indexing of information about data compression and hierarchical data layout in storage. ISABELLA supports indexing of data with lossy compression, whereas ALACRITY supports indexing of data with lossless compression. Compared to

state-of-the-art scientific data management systems (FastBit, SciDB, MonetDB), both ISABELLA and ALACRITY offer significant data storage reduction and more than 10-fold speed-up of per-core processing, and scalable multi-node, multi-core, and multi-GPU performance. Also, because of its light-weight storage footprint and embarrassingly parallel execution model, ISABELLA's query engine offers an 8-fold improvement in energy efficiency compared to state-of-the-art technologies (Figure 29). ISABELLA enables in-situ, query-driven data analytics over compressed data. It supports both precision- and multi-resolution level of detail (LoD).

For indexing satellite images, we focused on the specific instance of searching satellite images and building a structural indexing framework of Defence Metrological Satellite Program (DMSP) imagery based on object recognition and classification. This work started with a thorough benchmark of leading feature extraction approaches such as scale invariant feature transform (SIFT), wavelet feature extraction and texture extraction, both global and local with feature evaluation by correlation. Also, different classification techniques (Naïve Bayes, Support Vector Machine (SVM) and perceptron classifiers) using different set of dissimilarity measure (Euclidean, Manhattan, etc.) have been investigated to find the best optimal image classifier framework for such dataset.

We build a hybrid system of Classification/Searching methods capable of handling the multi-class case with comparable computational complexity both in training and at run time, and has good retrieval performance in practice. The basic idea is to locate nearest neighbors to a query image based on a hashing scheme and to quickly identify similar images; followed by classifying the images based on the distance to its neighbors using a semi-supervised support vector machine (S3VM) that keeps the distance function on the collection of closest neighbors. We developed a Multi-Level Indexing (MLI) framework that can effectively index imagery based on its feature's attributes to reduce the sample size significantly for the Image Matching (IM) comparisons for retrieval. MLI can effectively tackle the drawback of parameter tuning such as the number of hash tables of Locality Sensitive Hashing (LSH) and the computational cost of SVM. Different combinations of features with careful selection of attributes, measures and classification modules provided much better performance. An important stage of the MLI scheme is learning an effective indexing comparison function with very small labelled training examples. While S3VM is capable of exploiting unlabelled data very efficiently, it often suffers from over-fitting. Thus, we developed a probability max-margin technique using probabilistic latent semantic analysis (*p*LSA), a statistical technique capable of deriving a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables.

We extracted a total of 495 features comprising local and global features for each of the 20,000 samples images from the Defense Meteorological Satellite Program (DMSP) imagery dataset. By using MLI to index and retrieve similar images, we achieved an 18.7% improvement in classification rate.

The S3VM baseline is quite good with an average classification accuracy of 85.4%; however by augmenting the S3VM with our proposed *P*LSA technique, we improved the accuracy to an average of 90.9%. In the future, we plan to investigate the fusion of texture and shape features to make our algorithm scalable to larger datasets and to investigate the possibility of representing a query image by a descriptive sketch instead of an actual image.

For image similarity searches, many researchers have become proponents of an approximate nearest neighbor search. The idea being that, in most practical cases, approximate nearest neighbor is almost as good as an exact nearest. Locality Sensitive Hashing (LSH) is the state-of-the-art algorithm for finding approximate nearest neighbors. The LSH is an index-based data structure that allows spatial item retrieval over a large dataset. Our research focuses on developing a new LSH scheme to build a data structure for the DMSP imagery dataset. In the existing LSH scheme, each data sample is randomly projected to *k*-dimensional subspace. This *k*-dimensional subspace is then used to compute the number of hash tables (L) that should be created in order to build the data structure. The parameters, *k* and *L* significantly influence the performance of the LSH data structure. We theoretically demonstrate that there exists a new LSH scheme that outperforms the existing LSH scheme in terms of computational cost,

memory requirement and query runtime. Our scheme is a one-degree-of-freedom design in which the number of hash tables, L is chosen and used to compute the k-dimensional subspace for projecting the data samples. Experiments conducted on 1.6 million texture feature vectors of the DMSP satellite images have shown that our proposed scheme cuts the computational cost down by more than ***50%***; saves more than ***50%*** on the memory requirement; and improves the query runtime by more than a factor of two. In addition, the scheme is highly scalable and is ideal for scientific and industrial application where similar samples are to found in large data archives.

It is shown theoretically that for any query input, our proposed LSH scheme is capable of finding the similar images two times faster than the existing LSH scheme. This means that for a predictive modeling based on historical data, our scheme would be able to report the similar items quickly for inferences to be drawn. In addition, building the data structure under the LSH schemes is computationally expensive. The more the hash tables, the more expensive it is to build the data structure. It is shown that the proposed scheme does fewer computations in order to build the data structure compared to the existing scheme. More specifically, the proposed scheme saves ***50%*** on computational cost. Once the data structure is built, it has to be saved and used for answering queries. The memory requirement to do this could run into few gigabytes if not terabytes. The proposed algorithm is memory-efficient. In our scheme, two optimization techniques are used to reduce the memory requirement further. First, all the data points are sequentially indexed. The indexing is the same across all the L hash tables. This means that each data point is stored once instead of L times. This reduces the memory requirement by a factor of L. Secondly, the projection onto the k-dimensional subspace is not stored explicitly. For each data sample, these projections are quantized into a single integral value; used to point to a specific location on the hash table; and the sequential index of that data sample is stored in that location. Applying these indexing techniques to both schemes, our scheme saves approximately ***50%*** on memory required to store the data structure compared to the existing scheme.

For data reduction, ISOBAR code and theoretical performance model offer a predictive, scalable and power-efficient implementation of this strategy. Using various performance metrics about the system (e.g., network bandwidth and latency, disk read/write throughput) and the application (e.g., compression ratio and throughput, amount of data per core), ISOBAR's theoretical model accurately predicts the optimal balance among the placement of compression (on cores and/or compute/IO/staging nodes), the data movement, and I/O. Using the pre-conditioner, ISOBAR dynamically identifies highly compressible bytes to process by a compression method, while asynchronously writing the remaining, uncompressible bytes to storage with parallel I/O library. This way it effectively hides the cost of compression and I/O synchronization. It operates within the data staging architecture, where various data transformations occur while data is in transit or in situ, from compute nodes to disk. ISOBAR exhibits both read and write performance gains proportional to the degree of data reduction, which ranges as high as 46% on scientific datasets, in addition to reducing the total amount of data that is being stored and accessed. Because ISOBAR applies a pre-conditioner to identify compressible bytes in small chunks, it avoids wasting CPU cycles trying to compress incompressible bytes in the data. By operating on a lower memory footprint and in an embarrassingly parallel manner, this method results in energy-efficiency, while simultaneously offering a high throughput, reduced data movement, and data reduction that collectively translate to a factor of 2-6 reduction in energy consumption. ISOBAR showed superior performance when tested against state-of-the-art lossless compressors.