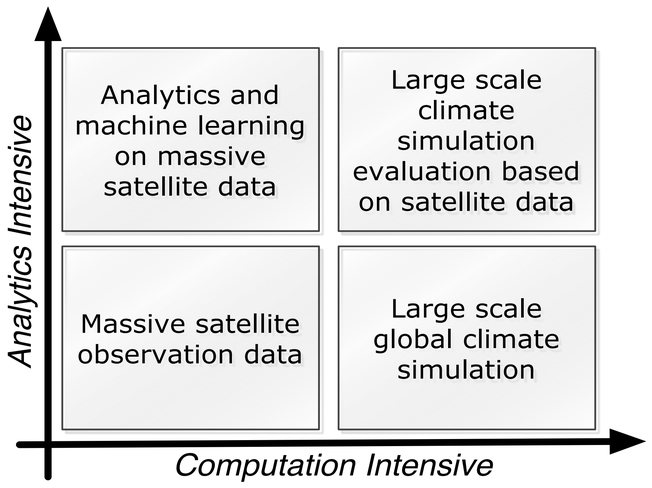
Multidisciplinary Research and Education on Big Data + High-Performance Computing + Atmospheric Sciences

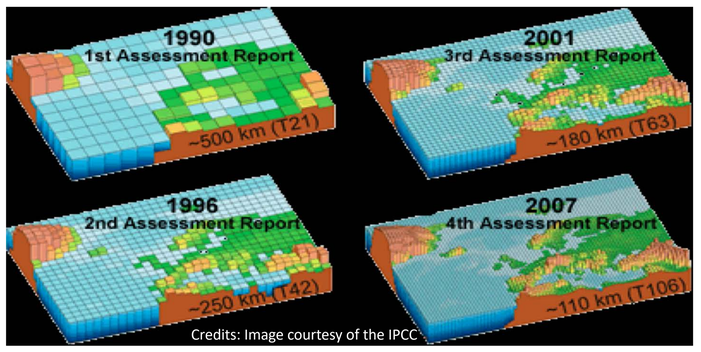
Jianwu Wang[1], Matthias K. Gobbert[2], Zhibo Zhang[3], Aryya Gangopadhyay[1]
[1] Department of Information Systems, [2] Department of Mathematics and Statistics, [3] Department of Physics
University of Maryland, Baltimore County, Baltimore (UMBC)
cybertraining.umbc.edu

HPC and Big Data Requirements for Atmospheric Sciences



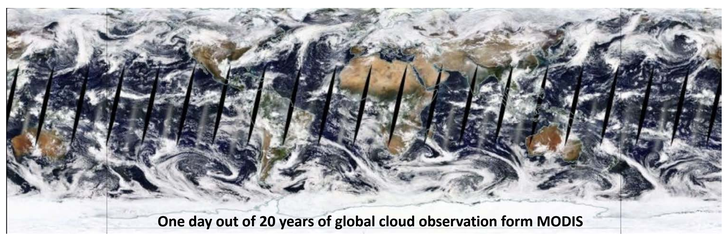
HPC Requirements for Cloud Simulation in numerical global climate models (GCMs)

- GCMs need high temporal-spatial resolution for simulation of climate
- One of most computationally expansive process in GCM is cloud and radiation simulation
- Approximations often have to made to alleviate computational burden



Big Data analytics requirements for evaluation of GCM using multi-decadal satellite observations

- Satellite-based measurements of global cloud properties have become an important source for evaluating cloud simulations in GCMs
- Satellite remote sensing has also led to an astronomically growing amount of data. Since its launch in 1999, the current MODIS satellite observation data has now about 500 TB raw data and PB processed data


One day out of 20 years of global cloud observation form MODIS

Challenges for Courses on "Data + Computing + X"

- Lower cyberinfrastructure (CI) adoption on advanced data and computing techniques in the current atmospheric sciences/physics curriculum
- Lack of training research challenges in applicable domains to apply their knowledge for graduate students in Computing and Applied Mathematics
- Lack of customized training for students in different majors
- Lack of team-based multidisciplinary training and frontier research projects

CyberTraining in "Big Data + HPC + Atmospheric Sciences"

- CyberTraining in big data applied to atmospheric sciences as application area and using high-performance computing as indispensable tool
- The training consists of instruction in all three areas, followed by faculty-guided project research in a multidisciplinary team of participants from each area
- Participating graduate students, post-docs, and junior faculty from around the nation will be exposed to multidisciplinary research
- Thorough program evaluation from the start

Syllabus

Module	Topic	Goal
1	Introduction of Python/C, Linux and HPC environment	Running their own jobs on HPC
2	Numerical methods for Partial Differential Equations (PDE)	Model as PDE and solve them using numerical methods
3	Message Passing Interface (MPI)	Write MPI jobs and performance studies
4	Basics of earth-atmosphere radiative energy balance and global warming	Understand basic concepts and principles of radiative energy balance and global warming
5	Basics of radiative transfer simulation framework	Understand the basic physics underlying the transport of radiation in atmosphere
6	GCM simulation and satellite observations	Understand the importance of GCM and satellite remote sensing
7	Basics of data science and machine learning	Understand the basics of data science and different types of machine learning tasks.
8	Introduction of and Big Data	Understand the basics of Big Data and demo programs
9	Big Data system: Spark	Write Spark jobs and run them on HPC
10	Big Data Machine learning	Write a machine learning program using Spark MLlib
11	Project introduction and assignment	Each interdisciplinary team will be assigned one project
12-14	Project progress report from each team and feedback	20 minutes report from each team + Q&A + rating
15	Final project presentation	Report, software, and final presentation from each team

Filling Knowledge Gap for Multidisciplinary Research

Graduate program	Existing courses can be leveraged	Other main courses offered	Additionally required knowledge
Information systems	• Programming • Data Mining and Machine Learning • Distributed Systems • Introduction to Data Science	• Databases • Artificial Intelligence • Decision Making • System Analysis and Design	• Computational Physics • Parallel Computing • Partial Differential Equations • Big Data Techniques and Systems
Applied Mathematics	• Partial Differential Equations • Computational Mathematics and Programming • Introduction to Parallel Computing	• Ordinary Differential Equations • Optimization Techniques • Combinatorics and Graph Theory • Linear Algebra	• Computational Physics • Data Mining and Machine Learning • Big Data Techniques and Systems
Atmospheric Physics	• Computational Physics	• Atmospheric Physics • Atmospheric Dynamics • Atmospheric Radiative Transfer • Atmospheric Remote Sensing • Quantum Mechanics	• Parallel Computing • Partial Differential Equations • Data Mining and Machine Learning • Big Data Techniques and Systems

Training Mechanics

- All work in multi-disciplinary teams of three (one from each discipline), 15 participants in total each year
- Team-building through TA-supported homework
- Team-based research with one faculty mentor, supported by one TA on that discipline => support is for disciplinary questions throughout training, not for one team
- Years 2 and 3: all work online with participants from around the nation, including taped lectures, team-based work, support, and all presentations

Year 1 Research Projects

1. Numerical Methods for Parallel Simulation of Diffusive Pollutant Transport from a Point Source
2. Mineral Dust Detection Using Satellite Data
3. Parallel Machine Learning Approaches to Cloud Type Classification
4. Spatio-Temporal Sensing Data Causality Analytics-An analysis of ENSO's global impacts
5. The Impacts of 3D Radiative Transfer Effects on Cloud Radiative Property Simulations and Retrievals

How to Get Involved?

- Welcome to apply for graduate students, post-docs, and junior faculty in US
- Advertise the training opportunity to broader community. We plan to have the second training in Spring 2019 and the third one in Spring 2020
- Serve on project mentors and guest speakers
- Help to make the efforts sustainable

Acknowledgments: This work is supported by the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources from the National Science Foundation (grant no. OAC–1730250).

NSF

UMBC
AN HONORS UNIVERSITY IN MARYLAND