Multidisciplinary Education on Big Data + HPC + Atmospheric Sciences

Jianwu Wang[1], Matthias K. Gobbert[2], Zhibo Zhang[3], Aryya Gangopadhyay[1], Glenn G. Page[4]

[1] Department of Information Systems, University of Maryland, Baltimore County (UMBC), Baltimore MD 21250, USA
[2] Department of Mathematics and Statistics, UMBC, Baltimore MD 21250, USA
[3] Department of Physics, UMBC, Baltimore MD 21250, USA
[4] SustainaMetrix, Portland, ME, 04103, USA
{jianwu, gobbert, zzbatmos, gangopad}@umbc.edu, gpage@sustainametrix.com

Abstract. We present a new initiative to create a training program or graduate-level course (cybertraining.umbc.edu) 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 of "Big Data + HPC + Atmospheric Sciences" supported by teaching assistants and 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 and have the opportunity for significant career impact. The paper discusses the challenges, proposed solutions, practical issues of the initiative, and how to integrate high-quality developmental program evaluation into the improvement of the initiative from the start to aid in ongoing development of the program.

Keywords: Big Data, High-Performance Computing, Atmospheric Sciences, Multidisciplinary Education, Developmental Evaluation.

1 Introduction

Next to theory and experimentation, computation has become the third pillar [1] and data-driven science has become the fourth pillar of the scientific discovery process [2] for many disciplines and critical to their research advances, such as bioinformatics, physics, computational chemistry, and mechanical engineering. It demands requirements on a course explaining how data and computation related techniques can help scientific discovery. Yet such a "Data + Computing + X" course is often missing in current curriculum design.

As an NSF-funded CyberTraining initiative to create a nationwide online training program, we are presently designing a "Big Data + HPC + Atmospheric Sciences" graduate-level program/course (cybertraining.umbc.edu) for students in three disciplines (Computing, Mathematics, and Physics) to foster multidisciplinary research

and education using advanced cyberinfrastructure (CI) resources and techniques. The course will teach students how to apply knowledge and skills of high-performance computing (HPC) and Big Data to solve challenges in Atmospheric Sciences. We focus on the application area of atmospheric physics and within it radiative transfer in clouds and global climate modeling, since these topics are important, pose computational challenges, and offer opportunities for big data techniques to demonstrate their impacts.

The participants in the new initiative will be selected competitively to form multidisciplinary teams of three participants with one participant from each area. The material is at the level of an advanced graduate course, and we anticipate most participants to be graduate students, but some can also be post-doctoral researchers or junior faculty. For all three groups, participating can have significant impact on their career in vastly expanding horizons from their own disciplines to two others. After an initial face-to-face course to develop the material, the training will be online with participants working together remotely from anywhere in the nation. In this way, this training can be made available to participants who do not have local access to the material. All work is conducted in an multidisciplinary team with participants from each area, mentored by a faculty and supported by a teaching assistants (TA) from each area. In the first 10 modules consisting of instruction in all three areas, team building is achieved by homework. In the final 5 modules, each team applies the material learned immediately to a small research project, culminating in a technical report and a project presentation. State-of-the-art collaborative and communication tools are used throughout, thus providing deep exposure to skills vital in today's job market.

The rest of the paper is organized as follows. In Section 2, we explain why HPC and Big Data techniques are needed for atmospheric sciences related research. Yet even with the requirements, designing a "Data + Computing + X" course still face many challenges which are explained in Section 3. To deal with the challenges, in Section 4, we design a "Big Data + HPC + Atmospheric Sciences" graduate-level course. Section 5 discusses the benefits, mechanical aspects, and how to integrate program evaluation from the start, and the paper concludes in Section 6.

2 HPC and Big Data Requirements for Atmospheric Sciences

Clouds play an important role in Earth's climate system, particularly its radiative energy budget [3]. On one hand, clouds reflect a significant fraction of incoming solar radiation back to space, which exerts a cooling effect on the climate. On the other hand, same as greenhouse gases, clouds absorb thermal radiation from earth's surface and re-emit a lower temperature, which has a warming effect on the climate. In addition, clouds are also an important chain of the Earth's water cycle and play center role in aerosol-cloud-radiation interactions. As shown in Fig. 1, cloud radiation and energy is also a research topic that can link HPC and Big Data.

Because of the important role of clouds in the climate system, a realistic and accurate representation of clouds in the numerical global climate model (GCM) is critical for simulating the current and future climate. However, at present there is a significant difference among the current generation of GCMs on the prediction of whether and to

what extent the global warming induced cloud changes would accelerate or dampen the warming [4]. The recent Intergovernmental Panel on Climate Change (IPCC) scientific reports have identified the cloud feedback be one of the largest uncertainties in our projection of future climate [5].

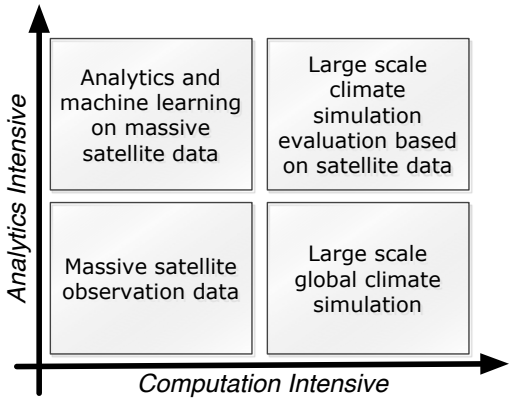

Fig. 1. Categorization of HPC and Big Data related challenges in cloud radiation and energy.

HPC requirements for cloud simulation in GCMs. It is extremely challenging to simulate clouds in GCMs realistically and accurately for two main reasons. First, many cloud-related processes, such as turbulence and convection, cloud droplet activation and growth, and transportation of radiation in clouds, occur at the spatial scale much smaller than the typical grid size of conventional GCMs (~100 km). New techniques, such as cloud super-parameterization embeds cloud-resolving models with resolution around 1 km inside of the conventional GCMs [6], have been developed aiming to solve this problem. However, such new techniques usually come with high computational cost, which more than ever makes HPC an indispensable tool for climate modeling. Another important reason is that many processes are modeled using highly simplified methods even though compressive methods are available to avoid the high computational cost. For instance, in the current paradigm, clouds are simply approximated as "plane-parallel" one-dimensional (1D) column, even though such approximation has been known to cause significant errors in atmospheric radiation and remote sensing computations. Over the past decade, a number of 3D radiative transfer models have been developed [7]. These new models, together with the fast growth of HPC resources, have given rise to emerging opportunities to shift the paradigm from 1D plane-parallel to 3D realistic simulation of the radiative transfer and cloud-radiation interactions [8].

Big Data analytics requirements for evaluation of GCM using multi-decadal satellite observations. The performance and reliability of GCMs are evaluated through comparisons of model simulations with measurements. Traditionally, measurements of the atmosphere made at the weather stations are sparse, especially over oceans, and unevenly distributed. The advances of satellite-based remote sensing techniques have led to a revolutionary change in our way to observe and measure the state of atmosphere. Now, satellite-based measurements of global cloud properties have become an important data

source for evaluating cloud simulations in GCMs [9]. Satellite remote sensing has also led to an astronomically growing amount of data. For example, the measurements from the Moderate Resolution Imaging Spectroradiometer (MODIS) are widely used for GCM evaluation [10]. MODIS takes measurements of the radiation reflected and emitted by earth-atmosphere in 36 spectral bands continuously over a swath width of 2,330 km. Since its launch in 1999, MODIS has made continuous measurements for almost two decades, which are invaluable for understanding climate variability and trend. However, the tremendous volume of data amount (~500 TB raw data, ~PB processed data) has become a difficult obstacle for making full use of MODIS data records.

3 Challenges of "Data + Computing + X" Courses

There are many challenges holding multidisciplinary research training of "Data + Computing + X" courses. We will explain the challenges in the "Big Data + HPC + Atmospheric Sciences" context. Similar challenges also exist in many other disciplines.

First, lower cyberinfrastructure (CI) adoption on advanced data and computing techniques in the current atmospheric sciences/physics curriculum. The traditional curriculum of an atmospheric sciences/physics major usually emphasizes the theoretical and experimental aspects of knowledge, but lacks the basic training on HPC and Big Data. For example at our institution, there is only one graduate course in the Atmospheric Physics Graduate Program, called Computational Physics, that has a numerical computation/simulation component. This course is designed to help the graduate students to build basic programming skills and numerical analysis techniques, as well as the ability to gain insights into physics problems using numerical simulations and models. As an introductory course, this course does not cover any topics related to HPC or Big Data.

Second, lack of training research challenges in applicable domains to apply their knowledge for graduate students in Computing and Applied Mathematics. Graduate programs in Computing (such as Computer Science and Information Systems) and Applied Mathematics teach some advanced Data and Computing techniques, such as programming and distributed computing in Computing, and parallel computing and partial differential equation (PDE) in Mathematics. But they normally do not have basic knowledge in application domains or disciplines that could benefit from what they learn. Instructors often use simple examples or common challenges in daily lives, such as social media analysis, as examples. These examples are often far different from scientific challenges in a discipline like Atmospheric Physics.

Third, lack of customized training for students in different majors. Students in different majors have different knowledge base and learning interests. Currently, graduate students in the Physics Department at UMBC who are interested in HPC or Big Data are usually forced to take advanced courses in other departments, such as Computer Science, Information Systems, or Mathematics. The feedbacks are usually not so positive, because these courses are not tailored for physics students. Meanwhile, students in Computing and Applied Mathematics also do not want to learn too much primary

knowledge in another discipline, especially the theory part, before they know how they could contribute.

Fourth, lack of team-based multidisciplinary training and frontier research projects. Each current curriculums is designed for it own major, which rarely provides opportunities for students in different majors to work together for multidisciplinary research. Many current training programs, such as XSEDE Education and Outreach Services [11] and NCAR/UCAR Education Programs [12] are great resources for trainees within one discipline. While we can leverage all these efforts, there are still lacks of a good list of well-defined frontier multidisciplinary research projects, especially for "Big Data + HPC + Atmospheric Sciences", designed for a team of students with diverse background to work on by applying the knowledge from the training. To conduct multidisciplinary research, many researchers, including us, have to seek collaborations by themselves and learn the knowledge from each other.

4 A Proposed Course for "Big Data + HPC + Atmospheric Sciences"

We propose to design a "Big Data + HPC + Atmospheric Sciences" course addressing the four challenges above through the following innovative approaches: 1) it will teach students in atmospheric Sciences how to implement and run parallel and big data programs at an HPC facility; 2) it will teach students in computing and applied mathematics how to solve atmospheric Sciences challenges by applying their knowledge; 3) it will provide distinctive learning outputs and homework to fit the background and interests of students in different disciplines; 4) it will provide team-based frontier research projects where each team is composed with students in different disciplines so they can collaborate and contribute from their own research interests.

Our proposed 15-module multidisciplinary course includes 1) customized course design for three disciplines with commonalities and differences; 2) data and computing techniques adoption for Atmospheric Sciences (three/four modules each for Data Science, HPC and Atmospheric Sciences); 3) identification of open challenges (including related open data) that can benefit from advanced CI resources and techniques; 4) five weeks long team-based project for frontier research challenges; 5) open source CI software implementation; 6) publications from the designed research projects. If taught during a regular semester, the workload is equivalent to that of a three-credit course.

4.1 Course Structure

Table 1 lists the 15 modules of the course, where it will take around three hours to teach each module. Details of each module are explained below.

Module 1: Introduction of Python/C, Linux and HPC environment. The first module explains the whole structure of the program and required basic knowledge for the program. It briefly goes through a programming language such as Python or C. It also introduces the hardware architecture, available software and basic usage of the UMBC HPCF environment.

Table 1. Modularized structure of the proposed training program.

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	Introduction of Data Science	Know basic tasks and techniques of Data Science
5	Basics of Big Data	Understand the basics of Big Data and demo programs
6	Big Data system: Hadoop/Spark	Write Hadoop/Spark jobs and run them on HPC
7	Basics of Machine Learning	Write a machine learning program using Spark MLlib
8	Basics of earth-atmosphere radiative energy balance and global warming	Understand basic concepts and principles of radiative energy balance and global warming
9	Basics of radiative transfer simulation framework	Understand the basic physics underlying the transport of radiation in atmosphere
10	GCM simulation and satellite observations	Understand the importance of GCM and satellite remote sensing
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 a final presentation from each team

Module 2: Numerical Methods for Partial Differential Equations. This module will explain the basics of partial differential equations, which is commonly used in physical models. It will discuss the use of numerical methods for PDEs, which is one major driving force behind research in many other fields like numerical linear algebra, scientific computing, and the development of parallel computers. It will cover the three basic PDE categories and their mathematical properties with examples. It will discuss two large classes of methods: finite difference and finite element methods.

Module 3: Message Passing Interface (MPI). This module will explain how to write MPI programs which is one of most common approach to build portable and scalable parallel scientific applications. It will cover basic MPI commands such as MPI_Send and MPI_Recv, collective communication commands like MPI_Bcast, MPI_Reduce/MPI_Allreduce, and MPI_Gather/MPI_Scatter. It will also explain how to write MPI programs in both C and Python (through mpi4py).

Module 4: Introduction of Data Science. This module will explain the basic concepts of Data Science, including generic lifecycle and different stages of data analytics, such as acquisition, cleaning/preprocessing, integration/aggregation, analysis/modeling and

interpretation. It will cover basics of descriptive statistics, graphic displays of data summaries, and basics of probability theory (including Bayes' theorem).

Module 5: Basics of Big Data. This module will explain the basics of Big Data, including its 5V characteristics. It starts with the challenges and bottleneck of many applications when dealing with large volume of data. Then it will introduce the basics of distributed file system and why we need them. It will cover Big Data concepts/techniques: data partitioning, data parallelization, key-value pairs, functional programming and MapReduce.

Module 6: Big Data system: Hadoop/Spark. This module will cover how to use two popular Big Data systems namely Hadoop and Spark. It will explain how Hadoop Distributed File System (HDFS) can achieve data partitioning, and fault tolerance and cluster management and job scheduling in Hadoop/Spark. For Spark, it will explain resilient distributed datasets (RDD), RDD transformations (map, join, cogroup, etc.) and actions (count, collection, foreach, etc.), lazy evaluation.

Module 7: Basics of Machine Learning. This module will explain the main lifecycle (training, testing, applying) and main types of machine learning (supervised and unsupervised learning). Major techniques to be covered include inferential statistics, feature selection, regression, correlation, clustering and classification. It will also explain how to construct Big Data machine learning through Spark MLlib.

Module 8: Basics of earth-atmosphere radiative energy balance and global warming. This module will explain the basic concepts and principles that control the radiative energy balance of earth-atmosphere system, and its implications to climate. The module will start with the fundamental physics, such as black-body radiation, followed by zero-order radiative energy balance between incoming solar radiation and outgoing terrestrial longwave radiation. The module will end with discussion of what kinds of roles the greenhouse gases, aerosols and clouds play in the radiative energy budget.

Module 9: Basics of radiative transfer simulation framework. Following previous module, this module will introduce the fundamental physical principles that control the transport of radiation (i.e., visible and infrared light) in our atmosphere. The module will also include the introduction of Monte-Carlo method and it application to radiative transfer.

Module 10: GCM simulation and satellite observations. This module will start with an introduction to the basic concepts and principles of numerical climate simulations, followed by explaining the importance of evaluating climate simulations and why satellite remote sensing products are invaluable for climate model evaluation. Basic concepts and principle underlying satellite remote sensing will also be introduced this module.

Module 11: Project introduction and assignment. This module will explain available research projects to be conducted in the following five weeks (see below for possible projects). For each project, it will cover the required techniques, suggested phases and major tasks, expected outputs, output evaluation metrics and challenges to each discipline. Each team will be assigned one project to work on.

Modules 12-14: Project progress report from each team and feedback from instructors as well. These three modules will be weekly project progress updates and discussions. Since each team has three members, every member will be a presenter for the

reports. All instructors and other teams will discuss the progress, perform peer review, provide feedback and give ratings.

Module 15: Final project presentation. The final module will be the final project presentation and final CI software program and technical report delivery. Each team will give a talk on the problems to be solved, the approaches taken, demonstration of developed software program, the experiments and results, and contributions of each member. All instructors and other teams will provide feedback and give ratings and suggestions for future work.

4.2 Sample Multidisciplinary Research Projects

The sample projects we have designed are listed below. With project assignment, each team will deliver their software program that can utilize CI resources/techniques for the project challenges and a technical report. Every project needs multidisciplinary research, deeper understanding of the topics, and learn necessary new knowledge for the project (such as OpenMP for Project 1 and additional classification models in Project 4). We will make sure each team has a unique project to work on so they can learn from each other and avoid possible plagiarism. Students are also welcome to design their own research projects related with the techniques/knowledge taught, especially for post-docs and junior faculties. Further, we will continue to design new multidisciplinary research projects and plan to maintain them well for future possible usage such as organizing nation or worldwide research competitions based on the projects. We note that the designed projects are closely related with the instructor team's current research interests so it will not be a huge burden on the instructors.

Team Project 1: Tuning of PDE simulations on modern architectures. This team will start with the MPI code for the sample PDE in Module 2 of the training and add OpenMP code to obtain a fully hybrid MPI+OpenMP code. The sample PDE here is the Poisson equation in two dimensions, which is a classical test case for linear solvers [13] and parallel computing. We mention that the team presentations to the whole group provide the ideal platform to share this knowledge with all participants. The team will then proceed to use cutting-edge Intel Xeon Phi KNL processors with 64 or more cores [14]. For best performance, the code needs tuning and the team will evaluate algorithmic and coding changes. With this basis, the PDE in the code will be generalized towards more complex PDEs as they appear in the atmospheric physics modules, such as the linear versions of the equation of radiative transport. This will make connections to the other teams and make the project presentations to all participants profitable for the other teams.

Team Project 2: Monte-Carlo simulation of radiative transfer: serial vs. parallelization. The Monte-Carlo method is a popular method that is widely used for simulating the radiative transfer of light in scattering-absorbing medium, such as cloud and aerosol [15]. It is also used for spectral integration in GCM integration scheme [16]. Because it is conceptually simple and yet highly flexible and relatively easy to be parallelized, the Monte-Carlo method applied in atmospheric radiative transfer is an ideal topic to introduce HPC. The team will first develop a simple serial Monte-Carlo radiative transfer model and use the model to solve some classic problems, such as cloud and aerosol

radiative forcing. At the beginning, problems will be simple with low computational cost that can be handled by serial code. More realistic and complicated problems will be added progressively, which will gradually help students to realize that the traditional serial code could not meet the computational cost for solving real-world problem. Finally, the team will parallelize their serial Monte-Carlo radiative transfer model using the knowledge from the training and then evaluate additional potential improvements of performance by leveraging GPUs.

Team Project 3: Derive regional climate trend and variability from MODIS radiometric measurements. MODIS is a key instrument aboard the Terra satellite launched in 1999 and the Aqua satellite launched in 2002. It takes measurements of the radiation reflected and emitted by earth-atmosphere in 36 spectral bands ranging from near ultraviolet to thermal infrared. The wide spectral coverage of MODIS enables remote sensing of a variety of cloud properties, such as cloud fraction, cloud top height, cloud thermodynamics phase, cloud optical thickness and cloud particle sizes. Since their launch, the two MODIS instruments have been making continuous measurements for almost two decades. It is an invaluable dataset for assessing and understanding the variability and changes of climate on decadal scale. The scientific objective of this project is to derive the variability and trend of direct MODIS radiometric observations over several "climate-sensitive regions", which will include 1) west coast of U.S. (drought), 2) East Asia and North-West Pacific (air pollution), and 3) eastern tropical Pacific (El Niño-Southern Oscillation). The results from this analysis will help the scientific community to understand the climate variability in these regions over the last two decades, and also provide the much needed benchmark for evaluating GCM simulations. The tremendous amount of MODIS data (~100 TB) is a great challenge to this project. The other objective of this project is to create a realistic scenario for the students to apply Big Data analysis techniques provided by Hadoop and Spark. The team will study how to partition data, write data parallel analysis program, and evaluate their scalability on UMBC HPCF.

Team Project 4: Rule based and clustering based classification on large scale MODIS data. Cloud type classification and identification are very important for atmospheric and cloud property retrievals. Many atmospheric applications depend on accurate and automatic cloud detection and classification. Generally, there are two types of approaches for cloud detection and classification based on satellite images. The first type is rule-based classification, which applies a set of thresholds of reflectance, temperature, spatial variances and others [17]. The second type is Machine Learning based classification, such as Bayesian methods [18] and clustering analysis [19]. In this project, students will first parallelize the MODIS cloud mask algorithm in [17] to large scale MODIS data using Hadoop/Spark. Then, they will apply clustering based classification on the same data using Spark MLlib. Last, we will ask students to combine these two types of approaches for further accuracy and execution performance improvement. [19] studied how to use rule-based results as initial classification of the iterative clustering based classification. This project will apply Big Data and parallel machine learning techniques to larger amount of MODIS data on a distributed computation environment.

Team Project 5: Evaluation of GCM through comparison with satellite observation data. The accuracy and reliability of our future climate projection rely on the skills and performance of our numerical climate models. Thus, it is an important task to evaluate the GCMs by carefully comparing the GCM simulations, including simulated cloud properties (e.g., cloud fraction, cloud optical thickness and cloud radiative effects), to observations [20]. A great challenge is that the definition of cloud in GCM is fundamentally different from observations. As a result, it is difficult to make "apple-to-apple" comparisons between GCM and observed cloud fields. To overcome this challenge, a model-to-observation projector, called CFMIP Observation Simulator Package (COSP) [21], is developed to facilitate the evaluation of GCM using satellite cloud remote sensing products. In this project, students will run off-line COSP on the CMIP5 GCM simulations [22]. Both the data generated by the CMIP5 model simulations and collected from satellite observations are in tremendous volume (~100TB) [23]. The team will study how to efficiently process the CMIP5 model simulations and satellite observations utilizing Big Data framework systems like Hadoop and Spark. The team will also evaluate GCM simulation using the COSP and satellite observations, and provide their findings.

5 Discussion and Evaluation

5.1 Benefits to Each Discipline

Table 2 lists the knowledge base and required training for students in different majors based on typical current curricula. It shows while each program offers some basic training, additional training is critical to conduct multidisciplinary research. Our proposed course is designed to fill the gap and take the diverse knowledge background of students in each major into consideration. For each topic, we will identify what needs to be learned for students in each major. For instance, students in computing related majors do not need to understand the physics theories of cloud-related processes, but they need to know their representations in programming models. By this approach, the students will be able to communicate and collaborate, yet they can still focus on their own interests. Further, the courses listed in the second column of the table will be prerequisite courses of our course so that students enrolled have enough preparation to start the course.

We acknowledge that our training program cannot fill all the gaps for "Big Data + HPC + Atmospheric Sciences" based multidisciplinary research. Instead, we select a narrow yet interconnected list of topics from the disciplines. We will focus on one research topic in Atmospheric Sciences, namely cloud radiation and energy. We will also limit the training of data and computing topics to PDEs, MPI, Hadoop, Spark, and three machine learning techniques. Since the instructors will design research projects before the course starts, the designed the projects will help us tailor the necessary knowledge to be taught in the first 10 modules. This will allow students to quickly grasp required knowledge and be prepared to collaborate and then focus on the research challenges. Because the course is designed to be multidisciplinary and research project driven, this

course does not intend to teach complete knowledge for each topic, such as HPC, Big Data and Atmospheric Sciences. Students can take other courses for a more complete knowledge for each specific topic.

Table 2: Knowledge base and required training for students in different majors to conduct "Big Data + HPC + Atmospheric Sciences" based 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

5.2 Teaching Mechanics

We plan to develop the course in three steps. In Year 1, the training material will be developed in a team-taught three-credit face-to-face course held at UMBC during Spring 2018. This is realistic for workload of the instructors and to give enough time for coordination (preparation during Fall 2017 and during the semester itself) among the instructors. This instruction of the first offering will be taped, and these tapes form the basis of the online off-site instruction in the following offerings. Already in the first offering with face-to-face instruction, we will recruit participants from nearby colleges, universities, and government agencies. We wish to create the training not just for graduate students. So, junior faculty at colleges (both public in the University System of Maryland, such as HBCUs Coppin State U. or Bowie State U., as well as private colleges including Notre Dame of Maryland) as well as post-docs / visiting assistant professors at academic institutions or government agencies (such as NASA, NOAA, EPA) who wish to extend the breadth of their professional preparation are part of our target audience. Additionally, graduate student populations exist at many local institutions,

including U. of Maryland College Park, and Johns Hopkins U., who will be interested. All these can already participate in the first offering face-to-face, since they can come to the UMBC campus for one class per week. We plan to offer this class as one meeting per week on Friday afternoons, so that in particular faculty and post-docs from local colleges/universities could avoid time conflicts with their own classes.

Starting with the offering of the training in Year 2, to be offered online in Spring 2019, we anticipate to recruit and admit from around the nation, using recruiting techniques including conferences, professional societies, regional mailing lists, and personal connections, and also the cohort from the first offering as multiplier. Including some post-docs or junior faculty from other institutions in Year 1 might also be very helpful in this outreach. We anticipate currently that the second offering will still use a regular semester schedule using UMBC start and end dates, so that we can continue to fine tune the organization.

Starting with Year 3, we plan to use a summer time slot, i.e., a compressed schedule of about six weeks, to demonstrate the feasibility of completing the training during a period outside of a regular semester, when the participants can focus solely on this training.

The admission of participants will be based on demographic information collected in a web form, a CV, a thorough personal statement, and at least two letters of recommendation. The personal statement needs to address specifically why the participant is interested in interdisciplinary research, how participation will promote his/her career goals, and how he/she can contribute to a team of participants from each discipline.

By teaching the same course at least three years will help us explore and compare different specific teaching mechanics including 1) whether we could setup teams and assign projects earlier so each team has more time to work on their project? 2) should class homework assignment be the same for all students or be linked with their individual projects; 3) whether it is good to have TAs fill in the vacancies if some students drop out in the middle of the course?

6 Evaluation Methods

The assessment of this multidisciplinary training program will follow the American Evaluation Association guidelines for systematic, competent, honest and respectful evaluation that is useful, accurate and conducted with due regard for the welfare of those involved in the evaluation. The methodology of the evaluation will be primarily a developmental, mixed methods approach [24] that focuses on the NSF approved proposal as the project strategy. The methods will include both qualitative [25] and quantitative [26] data collection methods, focusing on the logic and intended emergent outcomes that acknowledges that multidisciplinary training, even in a semester-long training program is uncertain and requires a dynamic approach. The purpose of the evaluation is to foster development of the curriculum over time by building learning and reflection into the process to further iterate the curriculum and training pedagogies and process. The assessment is built around a set of questions collaboratively developed in partnership with the external evaluation team, instructors and other relevant partners to

guide the inquiry. The external evaluation includes the development of success metrics aimed at student learning and quality of faculty collaboration and adaptive learning to further develop the training curriculum as a successful online course.

A key component of our strategy is to engage our outside developmental evaluator at the beginning of the process, one full semester before the first offering of the course. Timely feedback and a systems perspective are key features of the developmental evaluation that reinforces the purpose of further developing the on-line course rather than improve a fixed model. The anticipated innovation of the course is expected to be in its design as an adaptive, context specific course, rather than a static model that does not change. For example, after the first offering, data gathered by the evaluator and faculty will be discussed in a retreat format to further evolve the course consider other applications. This feedback loop will be repeated, with particular emphasis on collecting the experiences of the students, considering quality of their academic outputs and examining quality of faculty collaboration. The model will then shift as an online offering only as the context for the course will change and require further adaptation and development.

7 Conclusion

Both the National Strategic Computing Initiative [27] and the Federal Big Data Research and Development Strategic Plan [28] highlighted the importance of workforce development on HPC and Big Data. Starting with the current curriculum design and to prepare the next generation scientists, we present a new initiative to create a training program or graduate-level course in big data applied to atmospheric sciences as application area and using high-performance computing as indispensable tool. We outline a concrete procedure how to create the course and believe that this approach could also be used to create other courses for the "Computational and Data Science for All" educational ecosystem.

Acknowledgment

This work is supported in part by the NSF Grant OAC-1730250: CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources.

References

1. President's Information Technology Advisory Committee. Computational Science: Ensuring America's Competitiveness [Internet]. Available: https://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf (2005).
2. National Academies of Sciences, Engineering, and Medicine, Division on Engineering and Physical Sciences, Computer Science and Telecommunications Board, Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020. Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020. National Academies Press; (2016).
3. Trenberth KE, Fasullo JT, Kiehl J. Earth's Global Energy Budget. Bull Am Meteorol Soc. 2009(90), 311–323 (2009).

4. Soden BJ, Held IM. An Assessment of Climate Feedbacks in Coupled Ocean–Atmosphere Models. J Clim. 2006 (19), 3354–3360 (2006).
5. Intergovernmental Panel on Climate Change. Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press; (2014).
6. Randall D, Khairoutdinov M, Arakawa A, Grabowski W. Breaking the Cloud Parameterization Deadlock. Bull Am Meteorol Soc. 2003(84), 1547–1564 (2003).
7. Cahalan RF, Oreopoulos L, Marshak A, Franklin Evans K, Davis AB, Pincus R, et al. THE I3RC: Bringing Together the Most Advanced Radiative Transfer Tools for Cloudy Atmospheres. Bull Am Meteorol Soc. 2005 (86), 1275–1293 (20015).
8. Klinger C, Mayer B, Jakub F, Zinner T, Park S, Gentine P. Effects of 3D Thermal Radiation on Cloud Development. Atmos Chem Phys Disc. 1–39 (2009).
9. Kay JE, Hillman BR, Klein SA, Zhang Y, Medeiros B, Pincus R, et al. Exposing Global Cloud Biases in the Community Atmosphere Model (CAM) Using Satellite Observations and Their Corresponding Instrument Simulators. J Clim. 2012(25), 5190–5207 (2012).
10. Pincus R, Platnick S, Ackerman SA, Hemler RS, Patrick Hofmann RJ. Reconciling Simulated and Observed Views of Clouds: MODIS, ISCCP, and the Limits of Instrument Simulators. J Clim. 2012(25), 4699–4720 (2012).
11. XSEDE | Education & Outreach [Internet]. Available: https://www.xsede.org/education-and-outreach
12. NCAR/UCAR Education Programs | UCARConnect [Internet]. Available: https://ucarconnect.ucar.edu/education/programs
13. Watkins DS. Fundamentals of Matrix Computations. third. Wiley; 2010.
14. Jabbie IA, Owen G, Whiteley B, Graf JS, Gobbert MK, Khuvis S. Performance Comparison of Intel Xeon Phi Knights Landing.
15. Iwabuchi H. Efficient Monte Carlo methods for radiative transfer modeling. J Atmos Sci. 2015;72. Available: http://journals.ametsoc.org/doi/abs/10.1175/JAS3755.1, (2015).
16. Pincus R, Barker HW, Morcrette J-J. A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. J Geophys Res. 2003(108), 4376, (2003).
17. Ackerman SA, Strabala KI, Menzel WP, Frey RA, Moeller CC, Gumley LE. Discriminating clear sky from clouds with MODIS. J Geophys Res. 1998(103), 32–141, (1998).
18. Uddstrom MJ, Gray WR, Murphy R, Oien NA, Murray T. A Bayesian Cloud Mask for Sea Surface Temperature Retrieval. J Atmos Ocean Technol. 1999(16), 117–132, (1999).
19. Li J, Menzel WP, Yang Z, Frey RA, Ackerman SA. High-Spatial-Resolution Surface and Cloud-Type Classification from MODIS Multispectral Band Measurements. J Appl Meteorol. 2003(42), 204–226, (2003).
20. Intergovernmental Panel on Climate Change. Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press; (2014).
21. Bodas-Salcedo A, Webb MJ, Bony S, Chepfer H, Dufresne J-L, Klein SA, et al. COSP: Satellite simulation software for model assessment. Bull Am Meteorol Soc. 2011(92), 1023–1043, (2011).
22. Taylor KE, Stouffer RJ, Meehl GA. An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc. American Meteorological Society; 2012(93), 485, (2012).
23. CMIP5 - Home | ESGF-CoG [Internet]. Available: https://pcmdi.llnl.gov/projects/cmip5/
24. Patton MQ. Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. Guilford Press; (2010).
25. Patton MQ. Qualitative Research & Evaluation Methods. SAGE Publications; (2001).

26. Wholey JS, Hatry HP, Newcomer KE. Handbook of Practical Program Evaluation. John Wiley & Sons; (2010).
27. Executive Order -- Creating a National Strategic Computing Initiative. In: whitehouse.gov [Internet]. 29 Jul 2015. Available: https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative, (2015).
28. Nitrd N, Others. The Federal Big Data Research and Development Strategic Plan. nitrd.gov; Available: https://www.nitrd.gov/Publications/PublicationDetail.aspx?pubid=63, (2016).