Best Practices of Student Engagement in Online Teaching of HPC and Big Data Research

Matthias K. Gobbert
Department of Mathematics and Statistics
University of Maryland, Baltimore County
Baltimore, MD, USA
gobbert@umbc.edu

Jianwu Wang
Department of Information Systems
University of Maryland, Baltimore County
Baltimore, MD, USA
jianwu@umbc.edu

***Abstract*—There is a critical nationwide shortage of IT professionals as well as of scientists and engineers with high-performance computing (HPC) and big data related advanced computing skills. Simultaneously, the technology is growing in complexity and sophistication, which has led to the use of multi-disciplinary teams with members from a broad range of home domains everywhere in industry, government, and academia. Moreover, a lot of the vital team collaborations take will place virtually using a variety of software platforms now and in the future. We report here on experiences with preparing undergraduate and graduate students for these career opportunities in several contexts, from regular semester classes, an undergraduate summer research program, to an advanced graduate student CyberTraining program. All these programs are conducted fully online and leveraged concepts of flipped classrooms, recorded lectures, team-based and active learning, regular oral presentations, and more to ensure student engagement and lasting learning.**

***Index Terms*—Online Education, Big Data, High-Performance Computing, Multidisciplinary Education, Team-Based Learning**

I. Introduction

There is a critical nationwide shortage of IT professionals as well as of scientists and engineers with high-performance computing (HPC) and big data related advanced computing skills. This creates immense career opportunities for students in both core departments of computer science, computer engineering, information systems, and related ones such as applied and computational mathematics, and in application departments. For instance, since the use of large parallel computers is necessary for the forecasting of weather, hurricanes, tornadoes, and many other phenomena with vital public importance, domain scientists in physics need to be able to use the computing tools or at least have a deep understanding of them.

This has led to the use of multi-disciplinary teams with members from a broad range of home domains everywhere in industry, government, and academia. The successful communication among team members requires as much knowledge of the IT specialists about the application area and computing knowledge of the domain scientists. The modern university with its own multi-disciplinary teams in many research projects provides in fact experience of us faculty in doing this as well as requires it of our own students who we want to bring into our interdisciplinary research projects. Therefore, this background provides great opportunities for integrating research and education for faculty and their graduate students.

We report here on experiences with preparing undergraduate students, graduate students, and early-career post-docs/faculty for these career opportunities in several contexts, from regular semester classes, an undergraduate summer research program, to an advanced graduate student CyberTraining program. The examples of regular classes range from sophomore level classes with up to 60 students to upper-level electives and graduate classes with fewer students. CyberTraining is an NSF-funded training program in "Data + Computing + X" using multi-disciplinary teams of advanced graduate students, post-doctoral researchers, and junior faculty. The Big Data REU Site is an intensive 8-week summer program to motivate undergraduate students' involvement in research careers. While these examples range from pure teaching to (formally) pure research, it is both necessary as well as vital to include both formal instructional components and an opportunity for research projects in all of them to varying degrees.

Additionally, all these example programs were conducted fully online in research years and leveraged a variety of pedagogical concepts including flipped classrooms, recorded lectures, online comprehension quizzes, uploading and grading of electronic homework, team-based and active learning, regular oral presentations, and more. Besides the purpose to deepen the learning, these techniques are also designed to ensure student engagement. This engagement should vitally include also some personal touch to counteract the isolation of pure online learning. An additional advantage of the pure online is of course also the opportunity for reaching a nationwide audience. This can have impact on inclusivity of the opportunities such as by reaching students in underserved areas of the country or students who cannot travel, e.g., since they might have a disability or be a care-giver for their family.

The following sections collect experiences from the three types of examples, in Section II on the UMBC CyberTraining initiative from 2018–2020, in Section III from the REU Site on Big Data Analytics in Science and Engineering from 2021–2023, and in Section IV from regular undergraduate and graduate classes in recent and current years. Section V collects the overarching conclusions that apply to all types of online learning.

Corresponding Author: Matthias K. Gobbert.

II. The CyberTraining Program on Big Data + HPC + Atmospheric Sciences

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 training explaining how data and computation related techniques can help scientific discovery. Yet such a "Data + Computing + X" training is often missing in current curriculum design.

In 2017, the U.S. National Science Foundation (NSF) published the solicitation "Training-based Workforce Development for Advanced Cyberinfrastructure (CyberTraining)" designed to address this national need. This program continues currently with solicitation number NSF 19-524. Four faculty from three departments across two academic colleges at UMBC joined in response and proposed the UMBC CyberTraining initiative to create the nationwide online team-based training program "Big Data + HPC + Atmospheric Sciences" (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. Figure 1 illustrates graphically the connection between the disciplines.

The CyberTraining program teaches participants how to apply knowledge and skills of high-performance computing (HPC) and big data to solve challenges in Atmospheric Sciences. We focused 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 NSF funded our proposal in the inaugural year 2017 (OAC–1730250) for training programs conducted in 2018, 2019, and 2020 [3].

This program trained 58 participants and we reported on our experiences in conducting such training online and team-based with participants ranging from undergraduates (NSF-funded through an REU Supplement in Year 3), graduate students, post-docs/non-TT faculty, and TT (tenure-track) junior faculty [4]. We specifically describe how to practically create the necessary training material, chiefly the tapings of lectures for later asynchronous online delivery of contents and homework, during Year 1, and how to accomplish this in an institutionally supportive environment, but without the type of resources an institution with an institutional focus on online teaching would have. Thus, we wish to share our experiences to regular faculty, who might want to add aspects of online teaching to their repertoire. Table I summarizes the profile of the participants for our program over the three years. We can see 1) most participants are graduate students since we believe graduate students are still in their early years of their research career and the offering of multidisciplinary education would have bigger impacts on their future career growth; 2) we try to address the under-representation of female researchers in STEM disciplines by having relatively equal number of female participants (27) and male participants (31). Some additional participants not eligible for NSF funding (not graduate students or post-docs/faculty) were included without support. An additional benefit for local participants was the three-credit special-topics graduate course.

Transition from CyberTraining to an REU Site. We reported already in [4] that it is possible to involve undergraduate students in a program designed for advanced graduate students. In Year 3 of the CyberTraining program, we were successful in applying for REU Supplement support for six undergraduate students at our institution from the NSF. We recruited for these positions in August 2019 and admitted two students from each discipline in September 2019. The key to successfully involving the undergraduates was to start the training for these local students during the fall 2019 semester before the actual program in the following spring 2020. Since the undergraduate students had a full course load to start with, the spreading out of material is crucial. The students were grouped by department during fall 2019 with a faculty mentor from that home department. They started by learning about the topics out of the 10 instructional modules that are in their own area, thus when they joined a multi-disciplinary team, they had all something to contribute. We then during winter 2019–2020 leveraged the fact that the lecture videos of the first 10 modules are available for asynchronous delivery. The two teams of undergraduates in fact started on the homework and were able to get a head-start of several weeks of homework submissions before the official start of the program. Using the time thus freed up during several weeks of instructions in Weeks 1 to 10, the undergraduate teams also started on research substantially earlier than Week 11. This concept worked and the undergraduate teams have results on the same level as the more senior teams.

III. REU Site: Online Interdisciplinary Big Data Analytics in Science and Engineering

Based on the experience with online team-based CyberTraining on Big Data, we applied and were awarded a summer undergraduate research program, called an REU (Research Experiences for Undergraduates) Site by the U.S. National Science Foundation (OAC–2050943) for programs in summers 2021, 2022, and 2023. The proposal uses the experience to create a novel online team-based REU Site on the same topic for undergraduate students. We share here lessons learned of how to conduct a summer research program online for undergraduates from around the nation.

Recruiting, participants, projects. We received more than 120 applications of highly qualified and motivated students from around the nation. The 9 participants include 3 females, 1 Native American, 2 African Americans, 3 Asians, 3 Caucasians, 1 Hispanic, and 1 student with a disability. 4 of the 8 NSF-funded participants are from institutions with limited research opportunities according to the Carnegie Classification of Institutions of Higher Education. Each team is highly diverse in terms of race, ethnicity, gender, type of

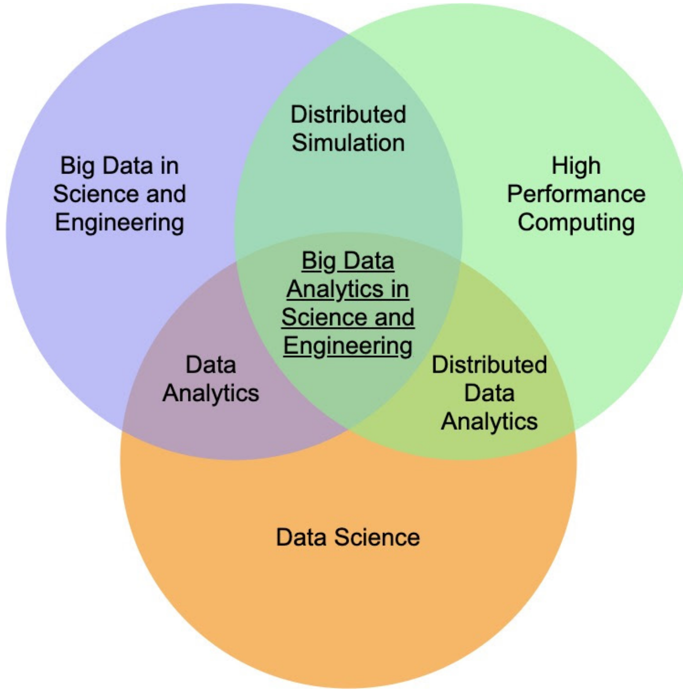

Fig. 1: Illustration of the connections between Big Data, HPC, and their applications in science and engineering.

TABLE I: Profile of participants for our training program.

	under-graduate	graduate	postdocs	faculty	total participants	female participants	total teams
Year 1	0	9	4	3	16	7	5
Year 2	0	14	2	1	17	6	5
Year 3	6	11	4	4	25	14	8
Total	6	34	10	8	58	27	18

home institution, geographic home, and more. This provides challenges, but also incredible opportunities for professional and personal development of the participants.

The 2021 program consists of 2 teams working on the projects (i) Big Data and Machine Learning Techniques for Sea Ice Prediction and (ii) Big Data and Machine Learning Techniques for Medical Image Classification. The projects involve collaborators in the application areas, Yiyi Huang from NASA Langley Research Center and Jerimy Polf from the University of Maryland School of Medicine, respectively. The research uses the CPU, GPU, and Big Data clusters in the UMBC High Performance Computing Facility (HPCF, hpcf.umbc.edu), giving participants a real-life experience on a shared distributed-memory cluster running Linux with batch scheduling of jobs, etc.

Schedule and ancillary activities. Figure 2 provides a schematic of the 8 weeks by distinguishing the three phases of instruction, research, and dissemination, which have distinct purposes, but necessarily need to overlap in time. The first week of the program consists of training on this cluster (Linux OS, slurm scheduler, power of parallel computing) followed by a concise introduction to data science, machine learning, and their software on our cluster (Python, NumPy, Pandas, matplotlib, Tensorflow, Keras, Horovod). The teams are guided by the authors of this note as faculty mentors and by one dedicated graduate assistant for each team; these are our own experienced PhD students/candidates, who also serve as TAs during the training. Simultaneous to this training, each mentor has the students do some background reading on their team's project, and the week ends with presentations by the outside collaborators to *all* teams, followed by social meeting of each team with its collaborator. This sets the stage for weekly progress updates by each team to all participants to widen the communication experience beyond the team.

The bulk of the weeks of the program are dedicated to conducting the research, with a transition from background reading to experimentation with the data to formulating a plan of attack and conducting the simulations. Throughout this, we introduce the need to document and disseminate by adding to the slide deck every week and writing the technical report one section at a time, beginning with background and literature review. The last week of the program culminates with each team giving a full oral presentation and publishing a technical report in the HPCF series. Along with the research training, the participants get exposure to a full set of professional development activities such as presentations by the Dean of the Graduate School about graduate school applications and by staff from the Career Center on services of such an office that

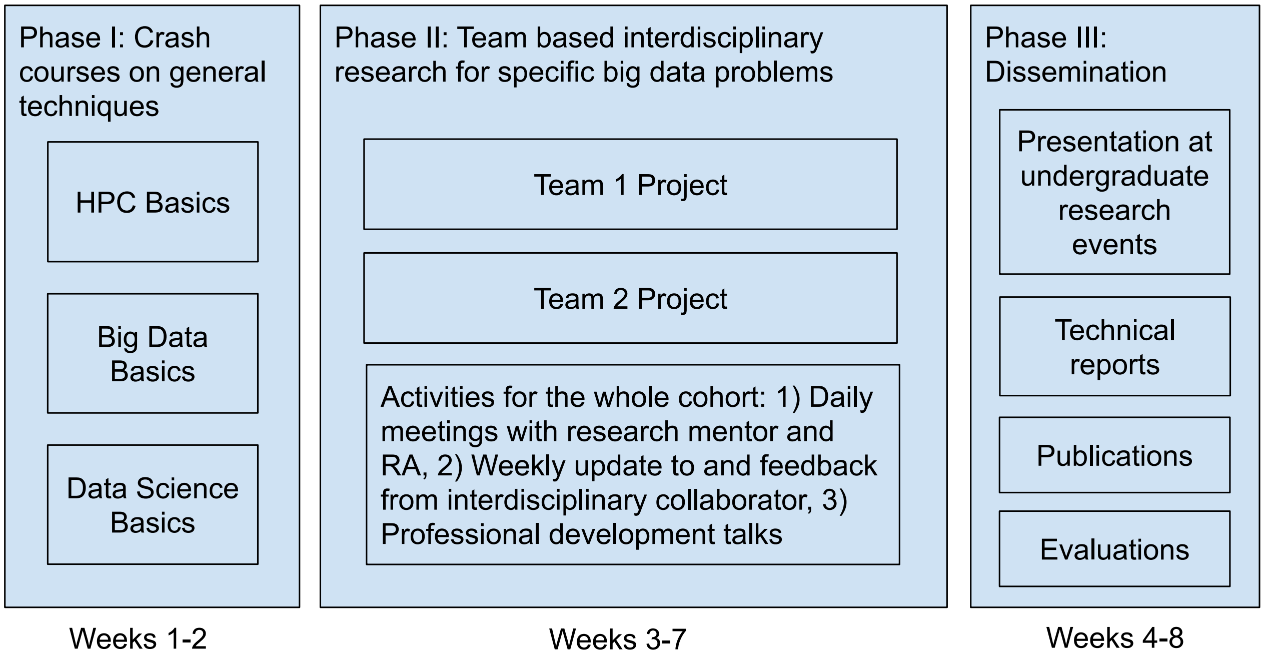

Fig. 2: Structural outline of the 8-week schedule in three overlapping, but distinct phases.

exists also at the participants' home institution. We also invite a wide range of VIPs to the program, from the department chair to the university president to explain the structure of a university more in depth than many know at this point.

An interesting practical aspect of an online program is the need to be mindful of participants from a range of time zones. We ask applicants already to confirm that they will be able to work on an Eastern time zone, where the host institution is located, but we typically schedule formal meetings not earlier than 10 or 11 a.m. Eastern time to accommodate participants on the West Coast. The teams themselves, both individually and in subgroups, definitely work earlier already, depending on where they are located.

Tools and techniques specific to online learning and research. The online format of our program requires the careful and deliberate choice of tools and techniques to facilitate online collaboration and particularly the creation of a community among each team and among all participants [5].

Webex: We choose Webex[1] as the meeting and messaging platform. In this software, running on any operating system including smart phones in a web browser or an app, we create, using Webex jargon, a "Team" for the REU Site 2021, under which we create "Spaces" for Instruction, Team 1, Team 2, and Staff. Each space is set up for a specific set of members (e.g., Team 1 or 2 only, or staff only), comprises a synchronous online meeting space, a chat for messages that persist (not just during one online session), a contents area for files and links, and more. The online meeting area lets all attendees see each others in video feed, allows for screen sharing by any member, provides for recording capability in the cloud, and a connection to the messaging area of that space. For the use of Webex or other online meeting tools, having a computer camera and microphone are vital. We did not make this a prerequisite, but rather we polled admitted participants if they needed such equipment, and we would have provided small-scale funding for it. The need did not arise in 2021.

Google Drive: Our cluster in the UMBC High Performance Computing Facility provides shared storage for data and code, organized by Unix groups, which facilitates shared access to files. But we report that undergraduates are very accustomed to Google Docs (documents, slides, etc.) and thus we leverage this by using Google Drive for each team to store their research files. This readily facilitates the joint creation of slides for the weekly talks on Friday afternoon to all participants on the week's progress, which the teams give by smoothly transitioning from one student to another from section to section. We observe that the number of participants familiar with LaTeX is small, but these all used it through Overleaf, an online implementation for collaborative editing.

LinkedIn: We created a group in LinkedIn for our program and invited all participants to it, as well as established links between all of us. The first author of this note guided an REU Site program from 2010 to 2017 using the same idea. The long-term benefits of this approach are by now paying off: We get regular exciting updates on the educational and career achievements of the alumni of that program, e.g., graduating from top-notch universities, accepting jobs at equally exciting institutions and companies, and generally staying in touch. This setup also allows students from several years to be in

[1] In early 2021, the product formally known as Webex Teams became simply Webex, the new mainline product, while the original Webex was renamed Webex Meetings.

touch with others beyond one year's cohort.

Social program: Research in an online format, where participants never meet each other in person, is difficult enough, and team-building a most substantial challenge, but the historically most significant outcome of a summer REU Site of making live-long friends is even more difficult to replicate! First on the team-building for research, we accomplish this by posing challenging, tightly timed, urgent homework, which helps the team members to learn each others' strengths and weaknesses under pressure. Second for the social aspect, we find that what used to take place casually by joint lunches and evenings at dormitories needs to be created more deliberately. We thus use the daily check-up meeting of one hour to include open-ended discussions, such as on their impressions of attending a PhD defense, chats about career goals, review and feedback of each resume/CV, and finally strictly getting-to-know-each-other icebreakers (e.g., all team members locating one thing in common, then pairwise introductions first within each team, then beyond each team).

Concrete action items: Based on our experience in summer 2021, there are a number of concrete action items that can improve the student engagement significantly. (i) Open the virtual meeting room 15 minutes early and talk to students showing up early as well as encourage them to talk to each other. Light prompts are stories about their location, as visible in the background, or other personal stories to share, like about the previous weekend. (ii) If possible, have the same number of participants in each team, to enable the organization of pairwise meetings across teams in breakout rooms. (iii) Organize pairwise or small group meetings other than the research team, based on something that they have in common, like class standing, racial/ethnic background, etc., to encourage also students to be in touch with people similar to themselves, as opposed to the highly diverse nature of the research team. (iv) In the spirit of "people do not do what you expect, but what you inspect", have one person write a mini-report about each small-group/pairwise meeting, which in just one sentence commits one significant observation to a shared, public Google Document. (v) Conduct one-on-one meetings of each faculty with each participant, both as early as possible in the program to get to know each other, and then again at the program midpoint so as to ask for feedback and to create opportunity for friendly admonition to contribute (if necessary). A good prompt might be to ask explicitly the questions "What would you like me to write in a letter of recommendation? What is your special skill set in general? What did you contribute to your team in particular?" These midpoint meetings go along with the progress review by the program evaluator and shows that the faculty care about people's feedback.

IV. Student Engagement in Regular Online Classes

The previous sections reported on two high-status programs with NSF funding. On the one hand, lessons learned from these programs can support student engagement also in regular classes that are taught online. And on the other hand, many standard techniques for online teaching that were developed for online classes in regular semesters, e.g., [6], [7], also apply to special programs.

Flipped classroom: This pedagogical concept means that the contents delivery does not take place in the synchronous class meetings, but requires asynchronous work ahead of the class meetings by each student. The synchronous class meeting is then fundamentally used to actually do work, which could be specially created assignments or part of regularly assigned homework. This concept applies also to face-to-face classes, and often recorded lectures are used. Having the lectures recorded can be a first step in converting a face-to-face class to online delivery. It is a customary recommendation to implement online automatic multi-choice or true-false quizzes to check on lecture comprehension. Panopto is a tool for taping, but also for hosting lecture recordings, and it provides tracking which user has watched what percentage of each assigned recording.

Team-based learning: One of the things most closely related to the special training or research programs is the use of semester-long assigned teams also in regular classes. As motivated in the Introduction, learning to work in teams is a vital contribution of college education to the typical workplace in industry and government. Many instructors use the Collaborate video conferencing module integrated in the Blackboard or Blackboard Ultra course management software. Based on the experience with Webex in the Big Data REU Site, I will switch to Webex also in regular classes, since team-work during synchronous meetings is much smoother in Webex by pre-assigned teams.

Frequent oral presentations: A key advantage of Webex over Collaborate is that, while screen sharing, Webex still displays the the webcam of the speaker (as well as a number of others). To improve student engagement in online classes, it is vital that each student speak to the class as often as possible. In online teaching, this needs to be purposefully scheduled to ensure that no student is forgotten (whether they like this or not). To this end, I will assign one student to summarizing the asynchronous lecture at the beginning of the synchronous meeting. In-class team-based quizzes end with one member of each team reporting to class. And after working on homework problems before class, a student can be assigned to present their solution to class. These approaches take inspiration from research training, in which it is vital to be ready to show one's solution for checking and feedback by others. The teaching of HPC or programming projects in general is well-suited to this approach, since there can be many correct solutions and thus it is useful for students to see others' solutions.

Project-based and active learning: It is a given that in the STEM fields, but in fact in all fields, learning is deepest and is best retained only, if it is actively used. In particular in programming contexts, it is customary to have one or several substantial projects during the semester. I would like to suggest that programming projects should be embedded into formal reports, since it is naturally expected both in research and in industry that the result of work is not just output, but an interpretation and documentation of the code and its output. This can be accomplished by setting clear expectations for reports, with a given formal outline, by assigning reading in that form, and by deliberately starting with small projects and short reports that grow to multi-week projects with a full-scale report during the semester. It is effective for learning to provide precise feedback on this report and to grade repeated revisions.

This list of ideas was developed over some years, but it turns out that many choices are backed up by educational research. For instance, [8] lists as ideas for teaching Gen Z students including Applied Learning, Intrapersonal Learning, Hybrid Classrooms, Flipped Learning, Learning by Example, Video-based Learning.

These techniques are applicable in any online class, in our cases ranging from sophomore level classes with 60 students to upper-level electives with 10 to 20 students to small graduate classes with 5 or 6 students. By using team-based learning, even a class with 60 students can provide the opportunity for active communication to students. By assigning students for the presentations, I can ensure that each gets a chance to talk. The upper-level and graduate classes tend to involve more programming in the context of HPC, while some other classes are more general in nature.

V. Conclusions

The previous sections showed that related techniques for online work can be used from regular classes to training and research programs, in particular by leveraging the same software tools for all of them and by using the same strategies for student engagement. This is also an example of virtuous synergy between the activities of research and teaching for faculty, and it is natural to involve for instance your own graduate students as teaching assistants in classes and/or research assistants in special programs, which in turn is useful to them, when the same tools are used in research meetings. Such meetings are common for multi-institutional projects, but we suspect that with more faculty having experience with them from recent online teaching, more office hours, faculty meetings, and local research meetings will be conducted virtually, as well.

Acknowledgment

This work is supported by the grant "REU Site: Online Interdisciplinary Big Data Analytics in Science and Engineering" from the National Science Foundation (grant no. OAC–2050943). 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). The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See https://hpcf.umbc.edu for more information on HPCF and the projects using its resources.

References

[1] "Computational Science: Ensuring America's Competitiveness," https://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf, 2005.

[2] "National Academies of Sciences, Engineering, and Medicine and others, future directions for NSF advanced computing infrastructure to support US science and engineering in 2017–2020," 2016.

[3] J. Wang, M. K. Gobbert, Z. Zhang, A. Gangopadhyay, and G. G. Page, "Multidisciplinary education on Big Data + HPC + Atmospheric Sciences," in *Proceedings of the Workshop on Education for High-Performance Computing (EduHPC-17)*, 2017, p. 8 pages.

[4] J. Wang, M. K. Gobbert, Z. Zhang, and A. Gangopadhyay, "Team-based online multidisciplinary education on big data + high-performance computing + atmospheric sciences," in *The 16th International Conference on Frontiers in Education: Computer Science & Computer Engineering (FECS'20)*, in press (2021). [Online]. Available: http://userpages.umbc.edu/~gobbert/papers/CyberTraining_FECS2020.pdf

[5] M. K. Gobbert and J. Wang, "Lessons from an online multidisciplinary undergraduate summer research program," in *The 17th International Conference on Frontiers in Education: Computer Science & Computer Engineering (FECS'21)*, accepted (2021). [Online]. Available: http://userpages.umbc.edu/~gobbert/papers/BigDataREU_FECS2021.pdf

[6] L. B. Nilson and L. A. Goodson, *Online Teaching at Its Best: Merging Instructional Design with Teaching and Learning Research*. Jossey-Bass, 2017.

[7] F. Darby and J. M. Lang, *Small Teaching Online: Applying Learning Science in Online Classes*. Jossey-Bass, 2019.

[8] C. Seemiller and M. Grace, *Generation Z: A Century in the Making*. Routledge, 2019.