Bringing High Performance Climate Modeling into the Classroom

Lan Zhao, Wonjun Lee, Carol X Song
Rosen Center for Advanced Computing
Purdue University
West Lafayette, IN
{lanzhao, wlee, carolxsong}@purdue.edu

Matthew Huber, Aaron Goldner
Earth and Atmospheric Science Department
Purdue University
West Lafayette, IN
{huberm, agoldner}@purdue.edu

ABSTRACT
Climate science educators face great challenges on combining theory with hands-on practices in teaching climate modeling. Typical model runs require large computation and storage resources that may not be available on a campus. Additionally, the training and support required to bring novices up to speed would consume significant class time. The same challenges also exist across many other science and engineering disciplines. The TeraGrid science gateway program is leading the way of a new paradigm in addressing such challenges. As part of the TeraGrid science gateway initiative, The Purdue CCSM portal aims at assisting both research and education users to run Community Climate System Model (CCSM) simulations using the TeraGrid high performance computing resources. It provides a one-stop shop for creating, configuring, running CCSM simulations as well as managing jobs and processing output data. The CCSM portal was used in a Purdue graduate class for students to get hands-on experience with running world class climate simulations and use the results to study climate change impact on political policies. The CCSM portal is based on a service-oriented architecture with multiple interfaces to facilitate training. This paper describes the design of the CCSM portal with the goal of supporting classroom users, the challenges of utilizing the portal in a classroom setting, and the solutions implemented. We present two student projects from the fall 2009 class that successfully used the CCSM portal.

Categories and Subject Descriptors
H.3.5 [**Information Systems**]: Online Information Services – *web-based services.* J.2 [**Computer Applications**]: PHYSICAL SCIENCES AND ENGINEERING – *Earth and atmospheric sciences*

General Terms
Reliability, Performance, Design.

Keywords
Community Climate System Model (CCSM), Science Gateway, TeraGrid, Education Users, User Interfaces.

1. INTRODUCTION
Simulation models are playing an increasingly important role in research and education across many disciplines. Use of large, complex and interacting models has become more and more common in today's society. It is important to help students learn how to build models and use modeling tools to solve complex human and natural system problems by combining the modeling theory with hands-on experience. However, the challenges for educators to integrate computationally and data intensive models into a curriculum can be daunting. Many campuses today have limited high performance computation (HPC) and data resources. Combining that with the classroom need for guaranteed resources during a specific and short period of time for a significant number of users is often an impossible barrier to overcome. Even if an instructor is able to secure the necessary resources, it is a significant challenge and time-consuming task to port, install, configure and validate a complicated model on a specific system. It requires a high level of expertise both with the model itself and with the computer systems in order to select the appropriate parameters, execute a simulation, manage the output data and use post-processing tools to analyze the results. Finally, the training and support required to bring novices up to speed could consume significant class time.

As part of the TeraGrid's effort in broadening participation, the TeraGrid science gateway program leads the way of a new paradigm to address these issues. The Purdue CCSM portal is a TeraGrid science gateway that aims at bringing large parallel climate model simulations into classrooms, into the hands of educators and students who may or may not be familiar with the use of HPC systems.

CCSM (Community Climate System Modeling) is a fully coupled state of the art climate model developed at the National Center for Atmospheric Research (NCAR). It is a well established research tool for studying and predicting the Earth system. CCSM is a large complex software package, with separate land, atmosphere, ocean, and sea ice components, and a central coupler that communicates data amongst them. CCSM simulations require supercomputing resources and a large amount of data storage for the output. According to the CCSM documentation, a typical model run on an IBM "bluesky" system (a set of IBM Power4 8-way nodes), at a dataset resolution of T42_gx1v3, generates 6.5 GB history files and 0.9GB restart files for each model year. It can simulate 7.5 simulation years each day using 104 CPUs.

To provide the broad user community with access to the CCSM climate modeling tool, Purdue TeraGrid resource provider

developed a science gateway enabling students and researchers to set up and run CCSM 3.0 on the TeraGrid Steele cluster. The Purdue CCSM portal made it possible for any student and researcher, regardless of their computing expertise, to run CCSM simulations on TeraGrid resources in a web browser [1]. It is a one-stop shop where users not only can launch a model run, but also monitor the run, access output data and apply post processing methods to obtain results. By removing the barriers of entry for climate modeling, the CCSM portal helps improve and broaden the use of TeraGrid for climate research.

The Purdue CCSM portal has been used for both research and education. For instance, it has been used in a sdy of the impact of wetland drainage on Midwest hydro-climatology. The researchers conducted sensitivity experiments designed for the first-ever study of hydro-climatic changes in Midwest USA due to large scale agricultural drainage carried out in early 20th century [6,7]. The portal was also used in a graduate course (POL 520 / EAS 591) taught by Drs. Matthew Huber and Leigh Raymond during the fall of 2009 at Purdue University. The portal allowed students to gain hands-on experience with running climate models and to explore geoengineering solutions to mitigate the effects of climate change.

In the following sections, we first describe how we designed the CCSM portal to facilitate student learning, and our experience gained by supporting the Purdue POL 520/EAS 591 class in Fall 2009, including user requirements, challenges and issues encountered, as well as solutions implemented. We then analyze the usage data and present the results of two student projects using the CCSM portal. We discuss related work and our future plan at the end of this paper.

2. CCSM PORTAL

The goal of the Purdue CCSM portal is to make a world-class, fully-coupled climate model easier to use and more accessible to a broad user community. For the education user community, our design goal was to help remove barriers and improve the teaching and learning of climate modeling by providing them with access to a research quality CCSM model through an intuitive and user-friendly web interface. The portal front end was developed using the GridSphere portal framework. It provides a visual interface for composing, configuring, and running a CCSM simulation. The interface also allows users to monitor the status of their submitted runs, access input/output data, apply post-processing packages on the output and visualize the results. At the backend, a copy of the CCSM 3.0 model was installed in a shared community software area and validated on the TeraGrid Steele cluster at Purdue. There is also a community data area that holds model output for each user.

The CCSM software comes with a set of scripts that perform basic tasks such as configuration and execution. The portal front end invokes a set of wrapper scripts using GRAM protocol which, in turn, invokes the corresponding CCSM scripts. Internally the portal uses a community account and a TRAC (TeraGrid Resource Allocation Committee) allocation to submit jobs to the TeraGrid system. Based on our experience, using a shared community account instead of individual user's TG account is critical to simplifying access in a classroom environment and ensuring portal adoption by instructors and students.

The CCSM model is quite complex to configure and execute. Many parameters may need to be changed at configuration time. Users often need to upload their own input datasets and modify the appropriate configuration settings accordingly. They also need to understand the details of the job scheduling system in order to configure their job submission scripts appropriately. To help reduce the learning curve of the model, the user interface must be designed to meet the needs of users who have different levels of expertise. For this purpose, we designed two sets of interfaces: a basic user interface and an advanced user interface. The basic interface provides a set of template-like web pages that guide new users through a typical CCSM workflow: create a new case, configure a case, and run a case (Figure 1). A screenshot of the interface for case creation is shown in Figure 2. The goal of this interface is to help beginners quickly learn the basics of setting up a CCSM simulation without the need to learn about the specifics of the backend computation resources. As a tradeoff, there are some restrictions on what the users can do through this interface. For example, users can only submit simulations for predefined periods: 1 month, 1 year, 5 years and 10 years. The model output will be in monthly format. In addition, they cannot choose which PBS queue the job is submitted to. The portal determines how to break down a simulation into segments and submit them to the appropriate computation resources without user intervention.

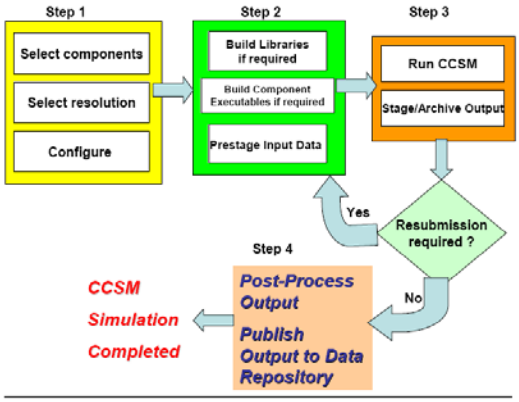


Figure 1. The workflow of the CCSM portal

Once the user has learned the basics of running a CCSM simulation, he can start preparing and running real-world scale simulations using the advanced user interface. The advanced interface still follows the same basic workflow of a CCSM simulation, though with many options to support various operations of the model that are available if run manually via command lines, e.g., modify configuration files, upload customized input data, select a specific computation resource, and restart a simulation using restart pointer files. Figure 3 shows the interface that allows users to easily modify configuration files.

Managing data for CCSM simulations for users is a much demanded capability in the portal. The CCSM portal provides interfaces for users to upload customized input/configuration files as well as downloading output data to their local system. It also has job management support for status monitoring and debugging information.

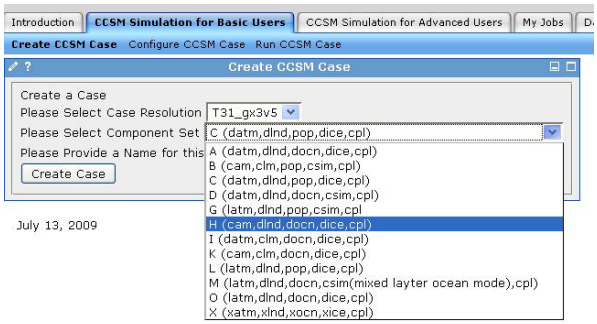


Figure 2. The basic user interface for case creation

Figure 3. The advanced user interface for case configuration

3. Class Use

The CCSM portal was used in a graduate class (POL 520/EAS 591, Purdue) in the fall semester, 2009. In this section we discuss in detail the requirements for use of the CCSM portal in a class, the challenges, solutions, and the lessons learned.

3.1 Class Description

The course POL 520/EAS 591, "Great issues: climate and policy", is designed to better educate students in the use and misuse of modeling across disciplines, so that they will be better builders and consumers of integrated, complex models of coupled human/natural systems in their own careers. More specifically, students will be trained to (1) model patterns of carbon cycling and radiative forcing in order to better understand past changes in climate and better predict future changes; (2) model changes in consumption and production of goods and services across regional, national, and international scales using economic models; and (3) explain and predict changes in public policy influencing these economic and environmental impacts using models of individual choice and collective action. All of these models are increasingly common and indispensable tools in the worlds of research. Yet models remain controversial and are subject to criticism despite their prevalence, and the debate becomes intensified as models become more and more complex, seeking to integrate the output from one complicated system as input to another. Through this class, students will gain the broader range of experiences needed to integrate models more effectively, and to better avoid the criticisms and problems leveled at the modeling process by its critics, instead of just learning the details of modeling within a single discipline.

In this class, students were trained in the theory and use of the models and received hands-on experience running CCSM3 via the TeraGrid. A key part of the class is having the students work in interdisciplinary teams (including at least one liberal arts and one College of Science student in each team) on semester-long projects to generate policy recommendations based on their own analysis of output from a suite of scientific, economic, and political models of climate change impacts. The project includes 3 components:

1) Proposals for potential strategies for responding climate change challenges, including mitigation, adaption for a team's nation, as candidates for suite of class climate model runs. The students design a set of climate scenarios to be explored with the CCSM simulations on the TeraGrid.

2) Review of national contributions to, risks from, and opportunities regarding projected climate change.

3) Recommendations for specific policy options to address climate change, based on this nation's unique contributions and concerns (documented the second paper component), including explicit consideration of political feasibility based on models of political behavior and based on the results from the CCSM simulations.

At the beginning, all students logged in the CCSM portal using their Purdue account during a lab session. They started with the basic user interfaces to learn how to create a simple case with default configurations and how to analyze the model output. They submitted simple simulations that took a few hours to run without resubmission. Later, the students were divided into groups and worked on their main project assignments using different kinds of configurations designed in the projects.

3.2 Class Requirements

The portal has been used by a small group of researchers before the class use. Use of the portal in a classroom setting has posed new requirements, some of which are common for general users as well:

Scalability: Class usage tends to be "bursty" with peaks and spikes during lab sessions and around project due time. The portal needs to support simultaneous use of the entire class while they are all performing various operations on the portal.

Performance: The portal interface should be responsive under stress. In addition, the time it takes to schedule and execute portal jobs need to be within a certain time range so that the students can finish their project assignment in time. This requirement continues to come up when we try to scale up the portal to the next level.

Easy to use: The portal interface needs to be simple and self explanatory for beginners while in the same time provide all the functionalities needed by skilled modelers.

Data access: CCSM simulations generate a large amount of output data. The portal needs to have adequate space to hold the data and a strategy to manage the quota for each user. It also needs to allow users to easily access and transfer the data.

Data sharing: The instructor wishes to be able to create a few cases for use as control cases by the entire class when performing data analysis. This allows the students to focus on their project design and execution instead of spending time and resources repeating the same simulations.

Information access and job management: When students are learning about the portal environment and the climate model, they tend to make mistakes. Therefore it is important for students to be able to cancel jobs that were not set up correctly, access debug information, and restart a job in the middle.

3.3 Feature Enhancement and Lessons Learned

The portal development team worked closely with the instructors and students to design and implement solutions to address the aforementioned requirements as the class moved along in the semester. First, the basic user interface was simplified to further reduce the learning curve of the model. Second, to handle the large output data volume generated by the projects, quota management is implemented to ensure the community data space will not run out of disk space. Older data will either be archived to tape storage or moved to user's local system upon user's request. Third, to facilitate data access, output data are managed by OPeNDAP server allowing remote data access using OPeNDAP protocol [8]. We also implemented online data analysis interfaces to further reduce the need of downloading large volumes of data. Fourth, users have access to detailed status messages after each operation and to PBS log files for their simulations. Finally, the Purdue RP staff tuned the model performance for different configurations which significantly sped up the execution time for the class projects.

In addition to the above feature enhancements, several coding issues were identified and mitigated/fixed. Previously these problems were not easily identifiable. They were first exposed when the portal was used under stress in the class.

1. Multi-user support – During class use, it often happens that multiple students click on the same button on the portal at the same time. A bug was found that caused the portal to randomly overwrite the information for new CCSM cases created by different users. This issue was resolved by generating unique file names for temporary files and separating them into user based directories.

2. Handle leak – During class use, the portal occasionally crashed. After investigation, two types of handle leaks were identified: The first type of leak happened because a MySQL connection object was not properly closed. A nested call orphaned the reference to an existing connection. The second type was caused by a bug in Axis web service which fails to close all the connections it generates. This issue has been reported by many developers and has not been fixed to date. As a workaround, we increased the value for expiration cache size to 60 seconds for some web services. As a result, if the user invokes the same web service within 60 seconds, it does not open a new connection. Instead it uses the existing TCP connection. This only mitigated the issue of increasing open file handles without completely solving it. Some forums suggest that using axis2 might resolve the problem.

To help speed up job scheduling and meet the class deadline, we developed a servlet that displays the current load of the three PBS queues available on the Steele cluster. This guides students to select the least busy resource so that their jobs can be scheduled with minimal delay. Although all the student simulations were run within the timeline defined by the class project, additional work remains to be done in order to scale the portal in serving more users in the future. One solution is to dynamically submit jobs at multiple TG resources based on the availability of the resources and the data transfer time. We plan to start with Queenbee and Ranger in the coming months.

4. RESULTS

Before using the CCSM portal in fall 2009, the same Purdue class had never been able to incorporate real-world scale CCSM simulations in the curriculum. As a result, the final projects were based around literature research. The availability of the CCSM portal not only enables students to gain hands-on experience in running complex computational models using HPC resources, it also allowed those who do not major in Earth and Atmospheric Sciences to set up a climate simulation in a speedy fashion. The students also felt that the advanced user interface allowed them to run cool simulations while in the meantime it helped them better understand the model code itself.

In 2009 fall semester, 25 students registered at the portal. Figure 4 shows the total number of jobs and total processor hours consumed by the CCSM portal in support of this class. Note that the total number of jobs submitted is 1310. This number is much bigger than the number of simulations ran by the students because most of the simulations are broken down into segments and automatically resubmit until they are completed. The total number of processor hours used by the class is about 162,000, with several usage spikes around December – a time when four groups of students worked on their semester long projects of running 100 year simulations. All 100 year simulations were completed successfully. The total size of output data generated is 0.6 TB. For the 100 year simulations, each run uses 56 CPUs and simulates about 15 years per day on TG Steele. At times of resubmission, it took on average 5 hours for a job to be scheduled. All simulation results are compared against a control carbon dioxide case where pCO2 is increased 1% per year. The students evaluated how effective the engineering approaches are when faced with enhanced carbon dioxide. This creates a realistic experiment which explores how future pCO2 changes will affect the mitigation strategies. All of the projects would not have been completed successfully in time before the end of the semester without the TeraGrid CCSM portal.

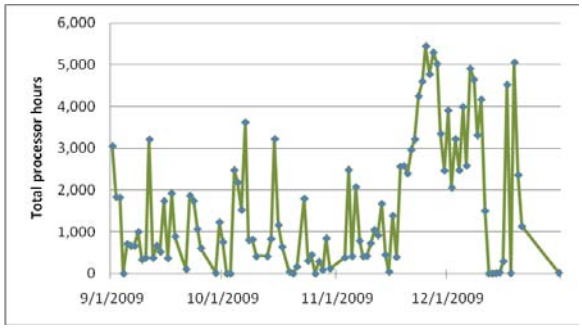


Figure 4(a). Total processor hours used by CCSM portal in support of the Purdue class POL520/EAS591.

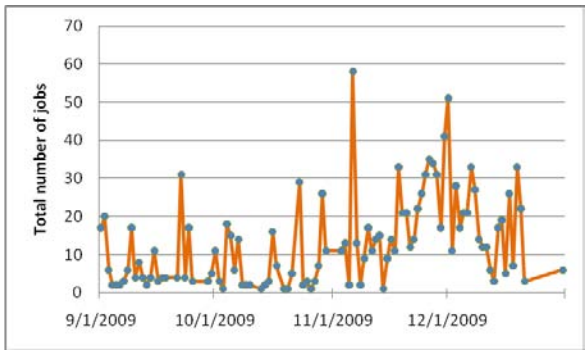
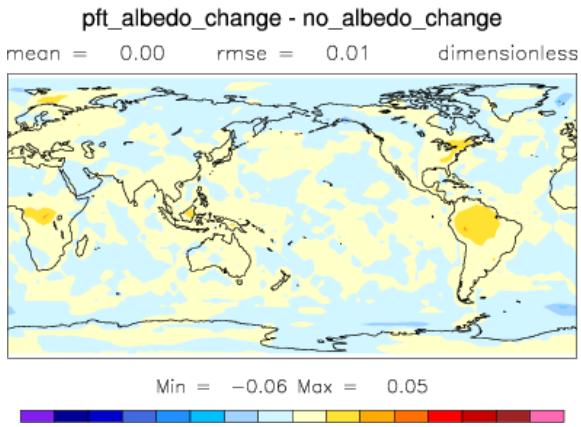


Figure 4(b). Total number of jobs submitted by CCSM portal in support of the Purdue class POL520/EAS591.

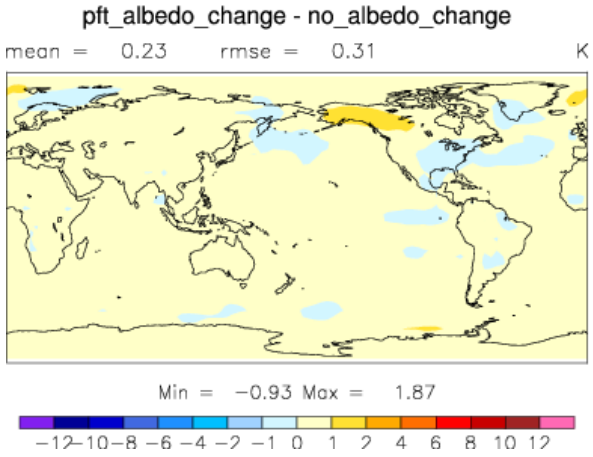
In the following sub sections, two student projects are selected to demonstrate the results of using the CCSM portal in the classroom.

4.1 Student Project 1 - Changing albedo of the earth to reflect incoming solar radiation and reduce global temperatures

Currently, global climate models predict that global temperatures will increase due to anthropogenic carbon dioxide increases. Project 1 focused around changing vegetation and forest albedos within the land model component of CCSM3.0 to look how this would affect global temperature. Albedo can be defined as the reflectivity of the object. The hope is that higher plant albedos will reflect more incoming sunlight thus reducing global temperatures. Plant albedos were doubled in the tropical and mid-latitude regions and Figure 5(a) shows the anomalous changes in global albedo. Results show that plant albedo is increased over the areas where we manually increased albedo. Interestingly, results showed an increase in global temperature. Our albedo changes in Africa and South America caused equatorial temperature gradients to shift due to large decreases in tropical temperatures, resulting in increased poleward heat transport, thus diminishing any reductions in incoming solar radiation and temperature seen in tropics. This project was not able to show that increases in plant albedos decreases global temperature.



5(a)

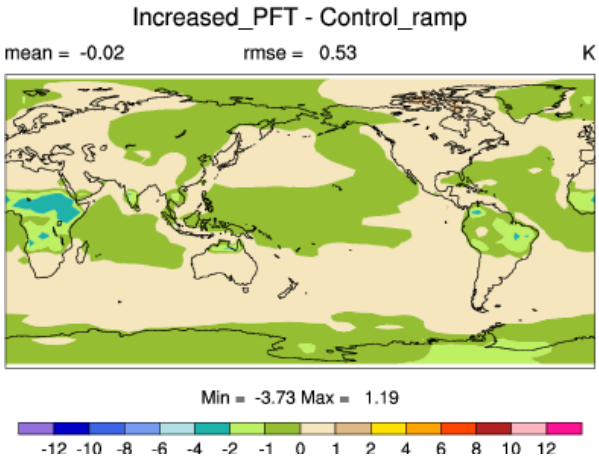


5(b)

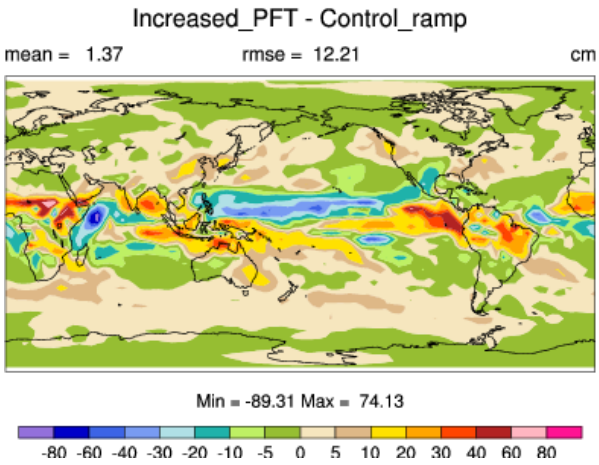
Figure 5. 5(a) shows the increases in albedo due to the increase in plant reflectivity. 5(b) shows the temperature anomaly that is driven by the albedo changes compared with the ramped carbon dioxide cases in Kelvin.

4.2 Student Project 2 - Reforestation of tropical regions to increase precipitation and decrease temperature

Group 2 chose to reforest the tropical regions around 20 North to 20 South in hopes of increasing rainfall due to increased evapotranspiration and cloud cover. Results show a slight decrease in global temperature around .02℃, and large increases in tropical precipitation around 1.37 cm/year. Temperature anomalies are seen in figure 6(a) and precipitation anomalies in cm/year are seen in Figure 6(b). Both results can be seen as fairly positive, especially the increases in precipitation over many tropical regions, which are always in need of precipitation. In addition the small decrease in global temperature forced by reforestation shows that this approach may be a piece of the puzzle in mitigating climate change issues.



6(a)



6(b)

Figure 6. Anomaly for global temperature (Kelvin) compared against the ramped carbon dioxide control run seen in figure 6(a). Cumulative precipitation (cm/year) anomaly seen in figure 6(b).

5. RELATED WORKS

With the emergence of TeraGrid science gateways, many barriers can be overcome and using the high-end computational systems by individual researchers and in classrooms is becoming a reality. Example science gateways include Geographic Information Science Gateway (GISolve), nanoHUB, LEAD, and TeraDRE. GISolve is developed at NCSA with a focus of providing an easy to use web interface for performing geographic information analysis using TeraGrid. It has been used in several college level classes in the University of Iowa and UIUC where the students gained hands on knowledge in areas from the basics of GIS to computing in Statistics (http://sc07.sc-education.org/conference/presentations/GISolve-Handson-SC07.pdf). Nanohub.org (http://nanohub.org/) is a widely used science gateway for online nanotechnology simulations and learning materials, developed by Network for Computational nanotechnology at Purdue. It is based on HUBZero platform and has been used in many classes around the world [5]. It allows students to access online lectures/seminars, build online collaboration as well as run simulations using distributed computation resources at Purdue University, TeraGrid and Open Science Grid [4]. LEAD (Linked Environments for Atmospheric Discovery) project is another cyberinfrastructure that allows meteorologists and scientists to explore meteorological data, forecast models and data visualization tools. It is designed to support a wide range of users including students learning weather modeling and predictions. During the weather camp workshops in 2006 and 2007 students from multiple institutions used the LEAD gateway to learn how to configure, submit, analyze and visualize Weather Research and Forecasting Model (WRF) model simulations using TeraGrid resources in a weather forecast competition [2]. The Distributed Rendering Environment on the TeraGrid (TeraDRE) is another science gateway system that has been used by Purdue students in a number of projects to render 3D animations on a cluster of distributed computers [3].

6. CONCLUSIONS

In this paper, we described our design and implementation of the CCSM portal as an effective means of bringing computation and data intensive climate models into the classroom. It targets both novice student users and expert modelers by providing two sets of interfaces. The portal uses a community account and submits jobs to the TeraGrid Steele cluster. The CCSM portal was successfully used in a Purdue graduate class and, for the first time, the students were able to gain hands-on experience of running 100-year climate simulations on TeraGrid resources via the portal. Several design and implementation challenges were identified and addressed during the class use.

The long term objective for the CCSM portal is to make it routinely used by students and climate researchers. The continued adoption and success depend on the improvement of scalability and performance of the system, such as by dynamically submitting jobs to different TG resources based on availability, and easy access to model input and output by developing utilities that help manage, access and transfer large datasets and couple the data with model execution on the TeraGrid/XD. We are also collaborating with NCAR/NOAA on publishing simulation metadata and data to the Earth System Grid, making them available to a broader research community.

7. ACKNOWLEDGMENTS

This research is sponsored in part by the National Science Foundation under TeraGrid Resource Partners grant OCI-0503992.

8. REFERENCES

[1] Basumallik, A. L. Zhao, X.C. Song, R. L. Sriver and M. Huber. “A Community Climate System Modeling Portal for the TeraGrid”, TeraGrid 2007 Conference, Madison, Wisconsin, June 4-8, 2007.

[2] Clark, R. D., S. Marru, M. Christie, T. Baltzer, K. Droegemeier, E. Joseph, and B. Illston, 2008: The LEAD-WxChallenge Pilot Project: The Potential of Grid-Enabled Learning. TeraGrid 2008, Las Vega, NV, June 9-13.

[3] D. Braun, X.C. Song and L.L. Arns. “Life Beyond the Browser: The TeraDRE, ” The 3rd International Workshop on Grid Computing Environments/Supercomputing’07, Reno, NV, November 2007.

[4] G. Klimeck et al., “nanoHUB.org: Advancing Education and Research in Nanotechnology,” *Computing in Science and Eng., Sept./Oct. 2008, pp. 17-23*

[5] HUBZero – Platform for Scientific Collaboration. http://hubzero.org/

[6] Kumar S., Merwade V., Bain, D. J., Hydro-climatological Impact of Century Long Drainage in Midwestern United States. Eos Transactions, American Geophysical Union, 89(53), Fall Meeting Supplement, Abstract H11F-0828, AGU Fall Meeting, San Francisco, CA, December 2008.

[7] Kumar S., V. Merwade, W. Lee, L. Zhao and X.C. Song. “Hydro-climatological Impact of Century Long Drainage in Midwestern United States: CCSM Sensitivity Experiments”, accepted to publish in the *Journal of Geophysical Research - Atmospheres*, 2010.

[8] OPeNDAP: Open-source Project for a Network Data Access Protocol. http://opendap.org/