Grid-BGC: a Grid-enabled research platform for high-resolution surface weather interpolation and biogeochemical process modeling.

Peter Thornton[1], Henry Tufo[1,2], Nathan Wilhelmi[1], Matthew Woitaszek[2], Craig Hartsough[1], Jason Cope[2]
[1] National Center For Atmospheric Research
P.O. Box 3000
Boulder, CO 80307-3000
[2] University of Colorado at Boulder
Engineering Center, 430 UCB
Boulder, CO 80309-0430

***Abstract* - Over the past three years we have developed a research-quality platform to support scientific users in the configuration, deployment, and analysis of high-resolution simulations of terrestrial biogeochemical processes. The platform consists of a web-based user interface (portal) that organizes and automates the complicated workflow necessary to perform high-resolution, large data volume simulations, a workflow system and database that manages the platform interactions with multiple users and multiple input and output datasets, a Grid Service Interface that manages the deployment of execution requests on remote computational resources and the staging of input and output datasets, a Reliable Job Execution Service that manages the simulation execution on the remote computational resource, and a portable core science library that encapsulates the surface weather interpolation and biogeochemistry simulation science executables.**

INTRODUCTION

Based on the expression of significant demand from the community of researchers exploring terrestrial biogeochemical dynamics through the implementation and evaluation of numerical models, our research team undertook to develop and test a research-quality platform that improves the ability of this community to advance their science without the usual large investments of time, energy, and funds in the technical aspects of the simulation environment. Our system enables experts in the field of biogeochemical modeling and analysis to configure and deploy high-resolution gridded simulations of water, carbon, and nitrogen cycling over large regions of potentially complex terrain (Fig. 1), and provides a manageable stream of output from these simulations for customized post-processing. Previously, it was necessary for each researcher to support their own implementation of a workflow management system (usually quite *ad hoc*), their own intensive computational platform, and their own bulk storage and retrieval system, in order to accomplish the scale of gridded simulation that is typical for regional and larger scales of terrestrial biogeochemistry research. These requirements are generally significant impediments to scientific progress, and our system was designed to eliminate or reduce many of these obstacles.

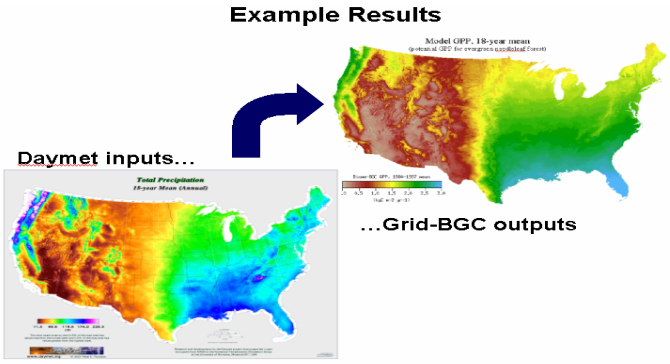

Fig.1. An example of the typical simulation domain targeted by the project, showing schematically the information flow from the Daymet model processing that produces gridded surface weather fields, to the Biome-BGC model processing that ingests these fields and produces estimates of the state and flux variables for carbon, nitrogen, and water cycles.

In the following sections we describe the workflow requirements for the system and the system architecture, with details on the user interface (web portal), workflow management sub-system, Grid service interface, reliable job execution service, and the core science libraries for surface weather interpolation and biogeochemistry simulation. We conclude with some examples of end-to-end testing of the system for biogeochemical simulations over parts of North America.

SYSTEM REQUIREMENTS

The requirements for flow of information and input and output data streams for a typical workflow are illustrated in Fig. 2. Blue boxes show the input parameters and input data streams that constrain a particular simulation workflow. These are provided by the user in convenient formats (either ASCII text files or standard ASCII representations of gridded data). The green ovals represent the core science components: The Daymet model for interpolation and extrapolation of sparse surface weather observations to produce gridded daily fileds of temperature, precipitation, radiation, and humidity [1-3]; the Biome-BGC model of terrestrial biogeochemical

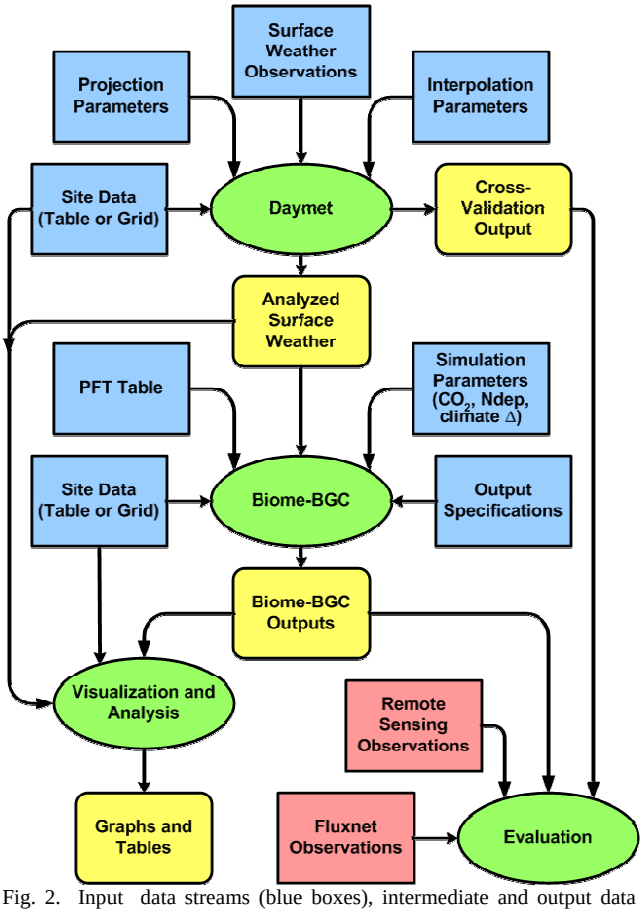

Fig. 2. Input data streams (blue boxes), intermediate and output data streams (yellow boxes), core science components (green ovals), and evaluation data streams (red boxes), combined to show the typical workflow for a biogeochemical modeling project.

dynamics [4-6], which receives input from Daymet as well as from user-specified input streams; and a collection of visualization and analysis tools that operate on the output streams (yellow boxes) from both Daymet and Biome-BGC. The following list summarizes the fundamental design requirements that we established to capture the workflow in Fig. 2:

1. Use emerging Grid-Compute technologies to provide a research-quality platform for terrestrial carbon cycle modeling.
2. Provide a Web Portal user interface to organize the complicated workflow and data object dependencies that are typical of very large gridded ecosystem model implementations.
3. Connect Portal-based simulation definition and control with automated job execution on remote supercomputer platforms, eliminating direct user interaction with the remote computational resources.
4. Provide automated data streaming for very large model input and output datasets between the Portal, remote computational resources, and a remote mass storage facility.

Provide robust analysis and visualization tools through the Portal.

System Architecture

Based on these workflow and design requirements, we organized our system into a small number of functionally distinct sub-systems with well-defined interface requirements (Fig. 3). An important constraint for high-resolution simulations over large regions is the need to handle large volumes of input and output data. We designed our system to interface with the NCAR Mass Storage System (MSS) for long-term storage and retrieval of results, with staging of active project datasets to and from local scratch space.

The system is organized as a web service client that passes requests to a Grid-BGC web service through a grid security infrastructure boundary. System components on the client side include the web portal user interface, through which a user specifies the parameters and input streams for a simulation project, a project management system that manages user input and output data objects and maintains scientific consistency between user requests and core science model requirements, and a job execution interface that generates service requests and tracks their status. System components on the service side include a job execution service and a workflow control service that handle execution requests for the Grid-BGC models and create the desired workflow for these models. The job execution and file transfer management components are implemented using Globus Toolkit services; the Grid Resource and Allocation Management Service (WS-GRAM) for execution management and the Reliable File Transfer Service (RFT) for reliable file transfer

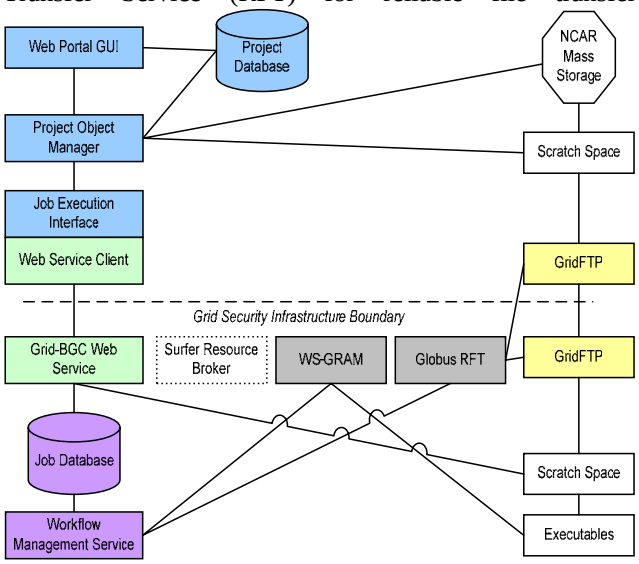

Fig. 3. System architecture, showing the division between operations taking place on the client (top) and service (bottom) sides of the Grid security infrastructure boundary.

management. To more easily perform resource selection as the system expands, a broker service will also be another essential service for this sub-system. As it stands, our system is not a pure service oriented architecture (SOA), although the system has evolved in that direction over the life of the project, in particular with respect to the web service side of the architecture. Movement toward a SOA is motivated in part by the potential for using the current web service architecture to support other earth science applications and workflows. We are currently testing the interface of our web service with an indexing service that registers metadata about multiple services, using the WS-Resources system. This permits clients to search a registry for a particular service and identify public methods and interface requirements.

CLIENT-SIDE COMPONENTS

The client-side component visible to an end-user is the portal interface. By implementing the user interface (UI) as a web-based portal, we can make changes to the central server without having to distribute applications to the end users. The UI is structured using Java Server Pages, and the Struts framework. User authentication is accomplished within the University Corporation for Atmospheric Research (UCAR) Gatekeeper authentication system, which requires that all users of the Grid-BGC system be registered through the UCAR Gatekeeper. The system maintains state for multiple registered users, and allows for the sharing of data and project definitions between users. Core functionality includes the ability to define new Daymet and Biome-BGC projects, to upload data objects and specify model parameters, associate them with previously defined projects, and order project execution. Tools are also provided that allow users to monitor the progress of their simulations as they proceed.

Below the UI is an application logic layer that manages the interdependencies between users, projects, and objects (the project building blocks). An important function of this layer is to maintain consistency between the critical requirements dictated by the core science code and the data objects specified by the user. This provides an important level of quality control in the workflow and removes one of the major operational obstacles to performing these large gridded simulations. Because these gridded simulations are typically quite large and computationally intensive, the client-side application logic layer includes an automated process by which the spatial domain of the simulation is disaggregated into multiple simulation regions, referred to as tiles. The temporal domain is also disaggregated so that individual units of work consist of a single time for a single year of simulation. The tiling algorithm is resolution-dependent, so that the upper limit of job size is maintained at an acceptable level, based on known data volume and computational time requirements.

Once a project and its associated objects are defined, the user can initiate an execution request which is then managed by the client-side job execution interface. This interface uses the Globus Toolkit's MyProxy proxy delegation service to issue grid service requests to either the Daymet or Biome-BGC grid service. A single user request is translated into a list of jobs (tile-years) that are passed as independent service requests to the relevant grid service. Identification and proxy information is passed with each job request that allows the grid service to return information on the job and also to pass the output files back to the client-side through Globus-RFT and Grid-FTP. The client-side job execution interface then passes the status information on each project execution request back to the user through the interface, including details on the number and status of individual tile-year jobs.

Additional client-side functionality includes the ability to visualize various aspects of the user-supplied input data streams, such as the raw surface weather observations that are a main input to the Daymet science core. Output summaries are also tabulated, including the cross-validation statistics and summary output time series information. Visualization components to handle gridded fields are still under development. Users also can elect to share the data objects they own with other individual users, or with all other users. A variety of "wizard" interfaces guide users through the requirements for uploading new datasets. The client-side system includes conversion utilities that translate all user inputs into a standard format (netCDF) for transfer to the core science routines. All output from the core science routines is also returned in netCDF format. Client-side development has relied on several open source application frameworks to reduce the amount of infrastructure coding required. These include Spring – J2EE Application framework, Hibernate – ORM toolkit, and Globus Toolkit 4 and Java Cog Kit for integration with the grid service.

CORE SCIENCE COMPONENTS

Our aim with respect to the core science components (the Daymet and Biome-BGC code bases) has been to make modifications as necessary to facilitate the input/output requirements of the Grid Compute implementation, but to leave the scientific algorithms unchanged from their stand-alone implementations. The core code has been designed from its inception with this approach in mind, so that the science algorithms are all wrapped in a core library that can be called by any number of different "front-end" codes. The advantage of this approach is that we are ensuring that all of the previous and ongoing efforts to evaluate and parameterize the core science algorithms will be applied to our Grid-BGC project seamlessly by using the latest core science library.

The input and output handling for the core science components has all been standardized to operate with netCDF file formats. This has simplified previous problems with cross-platform portability of input and output files that were caused by differing binary

representations. The generation of netCDF output files also makes it possible to publish the output results in an automated way through the use of OpenDAP protocols.

To guarantee integrity of results in the new web service environment and with new core science modifications, we performed a series of simulations using previously defined input datasets, and compared the new results to known good output datasets. The system has passed these tests, and we have proceeded with operational testing of the end-to-end capabilities of the system on a full-scale research problem.

OPERATIONAL TESTING

We have exercised the system on two separate full-scale research applications: a 1km gridded simulation of surface weather over the conterminous United States, and a 10km girded simulation of surface weather and biogeochemical dynamics over Canada (up to 60°N latitude). The U.S. domain consists of 253 2° x 2° tiles, and we performed simulations of surface weather for the years 1998-2003. The Canadian domain consisted of 15 10° x 10° tiles, and we performed simulations for the period 1960-2003.

REFERENCES

[1] P. E. Thornton, S. W. Running, and M. A. White, "Generating surfaces of daily meteorological variables over large regions of complex terrain," *Journal of Hydrology*, vol. 190, pp. 214-251, 1997.

[2] P. E. Thornton and S. W. Running, "An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation," *Agricultural and Forest Meteorology*, vol. 93, pp. 211-228, 1999.

[3] P. E. Thornton, H. Hasenauer, and M. A. White, "Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: an application over complex terrain in Austria," *Agricultural and Forest Meteorology*, vol. 104, pp. 255-271, 2000.

[4] P. E. Thornton, "Regional ecosystem simulation: combining surface- and satellite-based observations to study linkages between terrestrial energy and mass budgets," in *School of Forestry*. Missoula: The University of Montana, 1998, pp. 280.

[5] P. E. Thornton, B. E. Law, H. L. Gholz, K. L. Clark, E. Falge, D. S. Ellsworth, A. H. Goldstein, R. K. Monson, D. Hollinger, M. Falk, J. Chen, and J. P. Sparks, "Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests," *Agricultural and Forest Meteorology*, vol. 113, pp. 185-222, 2002.

[6] B. E. Law, O. J. Sun, J. Campbell, S. Van Tuyl, and P. E. Thornton, "Changes in carbon storage and fluxes in a chronosequence of ponderosa pine," *Global Change Biology*, vol. 9, pp. 510-514, 2003.