FACE-IT: A Science Gateway for Food Security Research

Raffaele Montella[1,2], Alison Brizius[2], Joshua Elliott[2], David Kelly[2], Ravi Madduri[2,3], Ketan Maheshwari[3], Cheryl Porter[4], Peter Vilter[2], Michael Wilde[2,3], Wei Xiong[4], Meng Zhang[4] and Ian Foster[2,3,5]

[1]Department of Applied Science, University of Naples Parthenope, Naples, ITALY
[2]Computation Institute, Argonne National Laboratory and University of Chicago, Illinois, USA
[3]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA
[4]University of Florida, Department of Agricultural and Biological Engineering, Gainesville, Florida, USA
[5]Department of Computer Science, University of Chicago, Illinois, USA

ABSTRACT

Understanding the potential impacts of climate change and the likely effectiveness of adaptation strategies is of crucial importance to the sustainability of both agriculture and natural ecosystems. Improvements in data availability and simulation model fidelity promises to enable significant improvements in knowledge. However, progress is hindered by the challenges inherent in creating and managing increasingly complex data acquisition, processing, simulation, post-processing, and intercomparison pipelines. To address these challenges, we are developing the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT) for crop and climate impact assessments. This integrated geospatial data processing, delivery, and simulation framework enables data ingest from diverse geospatial data archives; data regridding, aggregation, and other relevant processing required prior to simulation; large-scale climate impact simulation using a range of applications, including different agricultural models, and leveraging high-performance and cloud computing; and post-processing to produce aggregated yields and other output variables needed to enable model intercomparison and to connect biophysical model outputs to global and regional economic models and assessments. It leverages the capabilities of the Globus Galaxies platform to enable the capture of both workflows and simulation outputs in well-defined, reusable, and easily comparable forms. We describe FACE-IT and its application to studies within the Agricultural Model Intercomparison and Improvement Project.

1. INTRODUCTION

Many problems facing humankind occur at the intersection of the social, physical, biological, and computational sciences. Issues relating to climate change and food security, for example, require an understanding of interactions between the natural world and human society over long time scales. To this end, researchers seek to characterize vulnerabilities, impacts, mitigation, and adaptation to climate change in human and environmental systems.

Unfortunately, progress on these interdisciplinary problems is hindered by the difficulties that researchers experience when they seek to collaborate around data. The Agricultural Model Intercomparison and Improvement Project (AgMIP: `www.agmip.org`) [10] illustrates some of the challenges. This international project brings together more than 100 agricultural models (and modeling groups) to study climate vulnerabilities and impacts in agriculture and land use, risks to world food security due to climate change, and opportunities for improved adaptation capacity in both the developing and developed world. Yet even conceptually simple tasks, such as driving each of those 100+ models with output from a selection of Coupled Model Intercomparison Project (CMIP) simulations, can become prohibitively complex due to a multiplicity of data formats, inadequate computational tools, difficulty in sharing data and programs, and large data sets. These barriers hinder both research and rapid and effective transmission of new and existing knowledge to policy- and decision-makers [7].

To address these challenges, the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT) project is developing a cloud-based science gateway [14] that provides web-based access to a wide range of data projects, simulation models, and analysis tools. In this extended abstract, we review the technical challenges that the FACE-IT gateway aims to address, outline the FACE-IT approach, introduce the Globus Galaxies platform on which we build, and describe early applications.

2. TECHNICAL CHALLENGES

We expand here on the difficulties that researchers experience when seeking to collaborate around data, using the example of AgMIP investigators needing to link climate model output with many agricultural models.

Multiplicity of data formats. Climate models produce global fields as outputs, in well-structured NetCDF files, while crop models are typically designed for point-based field-scale experiments and expect customized ASCII formats for daily weather data. Climate datasets assume Greenwich Mean Time; crop models typically assume local time. These and many other differences in syntax and

Gateway Computing Environments workshop at SC14 New Orleans, Lousiana USA

semantics can make the task of adapting data for a new purpose challenging and error-prone. In principle, such problems can be addressed by providing translator programs that encapsulate all knowledge required to translate from one format to another. However, lacking mechanisms for sharing and discovering such translators, they are rarely available to researchers when they need them.

Inadequate computational tools. Particularly when working with unfamiliar data, researchers can end up spending a lot of time developing the knowledge and tools required to understand, analyze, aggregate, transform, etc., that data to meet their research needs. Frequently the obstacle to progress is a lack of suitable tools, which forces the researcher to develop their own custom analysis programs—using, for example, tools such as Matlab or R. These activities are not bad in themselves, but could be avoided if the researcher had access to the standard tool suite used by researchers for whom that data is familiar. The required tools typically exist, but the researcher working with unfamiliar data does not know how to find, install, or use them.

Difficulty in sharing data and programs. While some researchers are simply not interested in sharing code and data, many other researchers want to share but find it difficult to do so. We believe that if sharing data and code was as easy as publishing images to Flickr or Facebook, many more researchers would do it [9].

Lack of incentives for pro-social behavior. People's willingness to share can be further enhanced by creating suitable incentives, such as documentation of usage for their contributions [6, 9].

Large data. The increasing size of satellite, climate model, and other data sets requires scalable analysis methods and researchers with the numerical analysis and high-performance computing skills to develop them. Handling the vast bodies of input and output data, making it available in readily usable form, and automating the repetitive tasks of community science to ease the difficulties inherent in collaborations between disciplines and institutions requires completely new, integrated solutions that can be seen as a new branch of information technology.

3. THE FACE-IT APPROACH

We believe that many of these obstacles can be overcome, in part at least, by the judicious use of information technology. To this end, we are creating a Framework to Advance Climate, Economics, and Impact Investigations with Information Technology (FACE-IT) to allow a community to combine in a single (virtual) location—a FACE-IT Instance—the following capabilities:

- A **data store** that collects large quantities of diverse data, organized into scientifically meaningful datasets annotated with rich metadata, type, and format information; with powerful browsing and search capabilities to facilitate discovery of desired data; and with robust access control mechanisms to encourage contributions from people with sensitive data.
- Rich **program collections** for format conversion, analysis, etc., integrated with the data store so that users can easily determine which programs can be applied to which datasets—and then apply programs to selected datasets, easily and efficiently.
- Convenient **data and code ingest mechanisms** for adding data and programs to the data store, with clear record of provenance and automated metadata extraction and synthesis for subsequent discovery.
- Rich **social elements** to incentivize contributions, via recognition of popular datasets and programs.

In this way, we aim to allow researchers across multiple social and earth science disciplines to share not only data but the software tools used to create, manipulate, analyze, and visualize that data. Thus, researchers will be able to develop data manipulation and analysis tools, apply those tools to their own data and to data provided by others, link multiple tools into data analysis pipelines, and share such pipelines with the community. In so doing, we will shorten the time required to complete an analysis, improve information flow, and transform what it means to engage in reproducible research on global change and sustainability. We intend that FACE-IT accelerate discovery both *within* communities such as AgMIP by dramatically reducing barriers to data and code discovery, access, and exchange, and *across* disciplines such as environmental and economic sciences.

We are working to create a full-featured FACE-IT prototype and (in concert with the AgMIP community) an AgMIP-specific FACE-IT Instance that will both advance AgMIP research and enable at-scale evaluation of the FACE-IT approach. Our current prototype supports the following sorts of interactions:

> *AgMIP climate team researchers prepare a set of historical weather data and future climate model projections, plus software tools for generating time-series weather scenarios from these inputs that can be used to drive crop models. They then import these data and utilities into FACE-IT. Within hours, the AgMIP community has access to a pipeline that can be used to drive a suite of agricultural models with the a huge range of data and scenarios. An AgMIP Regional Integrated Assessment (RIA) team member from the University of Ghana prepares survey data for the farms in their region (information on planting dates, crops, cultivars, fertilizer, and irrigation for example) and uploads this data to their FACE-IT workspace. The researcher then chooses the locations, climate models, and scenarios from the AgMIP climate tools and uses these to run the AgMIP RIA workflows with their uploaded survey data. These workflows include multiple models and produce a variety of browser visualizations and publication quality images that can be downloaded directly.*

Data pipelines like these allow researchers to link climate, impacts, and socio-economic models and analyses into the end-to-end assessments of global change vulnerabilities and sustainability that are essential to the ability of science to consistently tackle the major interconnected challenges of our time and our future.

4. ARCHITECTURE, IMPLEMENTATION

FACE-IT builds on the Globus Galaxies platform, which has been developed over the past several years at the University of Chicago, initially in support of the Globus Genomics project [8]. We also benefit from substantial software development undertaken by the communities with whom we will work (see for example Section 5), who have developed most of the domain-specific tools required to populate FACE-IT with useful capabilities.

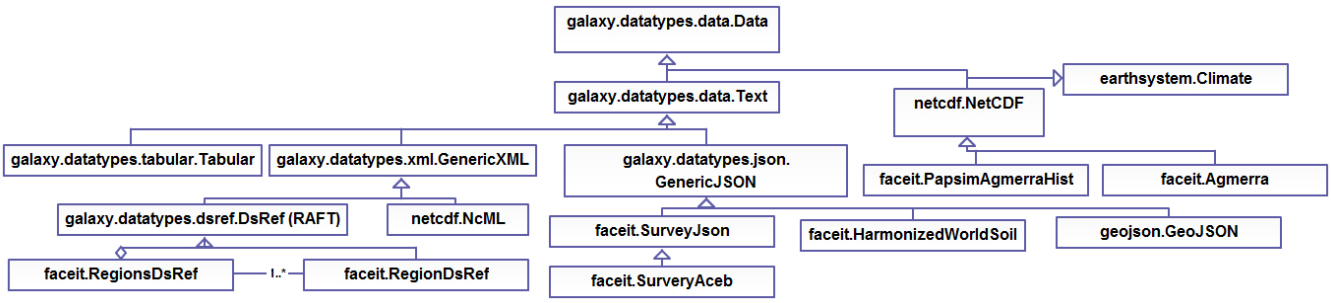

Figure 1: Data types implemented and used in the initial FACE-IT prototype

The Globus Galaxies platform leverages Galaxy [5, 3] for its simple, uniform, and extensible interface for selecting and executing components and workflows; Globus transfer [4] for data movement; Globus Nexus [2] for identify, group, and profile management; Swift [13] for parallel execution of workflow components in large ensemble simulations; and custom elements [8] for elastic, scalable cloud execution. These capabilities together enable the rapid development of cloud-hosted science gateways that support community access to and exchange of complex data and computational tools.

We then define appropriate Galaxy datatypes (see Figure 1) to represent the data types of interest to our target communities, and integrate a suite of data transformation, data analysis, and simulation tools that implement behaviors important to our target communities. We thus leverage native Galaxy facilities to simplify the coupling of external data resources with available analysis tools. This approach is agnostic to the type of data that is returned from a particular data resource, but strongly data typed in order to ensure that different pipeline or workflow steps are correctly coupled: either the input/output data types are the same or an automatic conversion is introduced.

While upload tools easily allow data to be uploaded directly, we also work with earth-system-data providers to make popular data sources (such as the Daymet daily surface weather dataset [11, 12]) FACE-IT-compliant, so that data from these resources can be redirected directly into a user's FACE-IT workspace with a single click. Once data has been accessed by a user and placed into their history, FACE-IT automatically determines data formats and identifies compatible applications. FACE-IT allows users to share and publish data and results as Data Libraries, as rich analysis pipelines (User Histories), as customizable multi-step pipelines (Workflows), or as complete experimental protocols using Galaxy-Pages. Protocols are available to integrate command line analysis tools and to send data sets to external web applications (e.g., for dynamic visualizations).

Because compressed and composite datasets are crucial for many Earth-system applications, we have implemented components in FACE-IT to support compressed and composite dataset management, upload, and automatic type detection. NetCDF, a standard in the Earth-system community for multidimensional data storage of sparse matrices, is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. NetCDF files are binary and self-descriptive. Common Definition Language (CDL) and NetCDF Markup Language (NcML) and both used by FACE-IT to improve data typing and manage conventions. In FACE-IT, we introduce the concept of "NetCDF Schema," implemented as a NcML header representation allowing regular expressions in dimensions and variables definition. New NetCDF-based data types can be introduced in a straightforward way by the creation of the related NetCDF schema.

5. APPLICATION WITHIN AGMIP

AgMIP regional research teams in South Asia and Sub-Saharan Africa use AgMIP protocols to conduct regional integrated assessments to quantify the effects of climate change on food security in their regions. The teams use multiple climate, crop and economic models to answer three key questions: (1) What is the sensitivity of current agricultural production systems to climate change? (2) What is the impact of climate change on future agricultural production systems? (3) What are the benefits of climate change adaptations? The AgMIP Regional Integrated Assessment (RIA) process requires collaboration among a multi-disciplinary team to provide consistent and cohesive inputs at each phase of the process for climate, crop and economic analyses. These processes are described in detail in the AgMIP Regional Integrated Assessments Handbook [1].

This FACE-IT use case focuses only on the crop modeling process, which uses outputs of climate models as input and which generates inputs for the economic modeling phase of the assessments. Many sets of crop modeling simulations are required in order to evaluate current climate and technology conditions, future climate conditions with current technology trends and future climate conditions with adaptation. Each system is simulated for multiple climate models and climate scenarios, multiple crop models, and multiple site-years. These many simulations are evaluated, compared, and used as input to regional economic models.

The data translation tools in the RIAs use site-based AgMIP Crop Experiment (ACE) data and assumed model parameters based on expert knowledge of the cropping systems being modeled (DOME data). These data are converted from spreadsheet templates to comma-delimited format and then to zip archives. Generation of ACE and DOME data which produce good, error-free simulations for multiple models and multiple sites is an iterative process, but for the purposes of this use case, it is assumed that the iterative process to create data for the simulations has been done outside the FACE-IT workflow, using the existing desktop utilities.

Figure 2 shows the FACE-IT workflow for the East Africa demo of the AgMIP Regional Integrated Assessment. ACE and DOME data in comma-separated-value (CSV) format

form the initial input. For the East Africa demo, there are four regions, each simulated for three climate scenarios (baseline and two future climate scenarios), requiring a total of 12 downloads. These data are combined into a single dataset using the "Region to group of Regions" app. QuadUI app converts CSV data to harmonized format, then to DSSAT model format. At this point, additional models (APSIM, EPIC, others) each with unique input requirements can also be added. The DSSAT app calls the model to perform simulations. Harmonization of model outputs is done using the ACMOUI app. Generation of plots from the simulated data are done with the ACMOPlot app.

Figure 3 illustrates three plots generated from the East Africa demo workflow. These plots are typical of the kinds of analyses done by each of the AgMIP teams in Sub-Saharan Africa and South Asia. Figure 3a shows a comparison of harvested yields for two climate scenarios in the Embu region of East Africa. Each box plot is generated from simulated yields from the 320 sites in the region, each simulated for 30 years of weather data for the specified climate scenario. Figure 3b shows the same data in a cumulative probability graph. Figure 3c shows average yields over the 30 sites for each of the 30 years of weather data for the same two climate scenarios. Previously, each AgMIP regional research team generated similar graphs using different methods and software tools. The FACE-IT workflow will allow much greater standardization of analysis and comparability among regions, while still allowing flexibility and customization in aggregation of data by each team.

6. CONCLUSIONS

We have described FACE-IT, a new IT infrastructure designed to accelerate existing disciplinary research and enable information transfer among traditionally separate fields. At present, finding data and processing it into usable form can dominate research efforts. By providing ready access to not only data but also the software tools used to process it for specific uses (e.g., climate impact and economic model inputs), FACE-IT allows researchers to concentrate their efforts on analysis. Lowering barriers to data access allows researchers to stretch in new directions and allows researchers to learn and respond to the needs of other fields. FACE-IT accomplishes these goals by building and integrating a number of powerful web-based software tools to enable researchers to easily develop data manipulation and analysis applications, apply those applications to their own data and to data provided by others, link multiple applications into data analysis pipelines, and share such pipelines with their collaborators and community. Our implementation builds on the Globus Galaxies platform, integrating a variety of data analysis and simulation tools, and can run on both cloud and HPC systems. We described an initial application to food security studies in Africa, a first step towards broad adoption within the international AgMIP community.

7. ACKNOWLEDGMENTS

We thank the Globus Galaxies, Globus, and Galaxy teams for their outstanding work on those systems, and for their assistance with this project. This work was supported by the NSF cyberSEES program award ACI-1331782 and by the DOE under contract DE-AC02-06CH11357.

8. REFERENCES

[1] The AgMIP Regional Integrated Assessments Handbook, `http://www.agmip.org/regional-integrated-assessments-handbook/`.

[2] R. Ananthakrishnan, K. Chard, I. Foster, and S. Tuecke. Globus platform-as-a-service for collaborative science applications. *Concurrency - Practice and Experience*, In press., 2014.

[3] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, pages Unit 19.10.1–21, Jan. 2010.

[4] I. Foster. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*, (May/June):70–73, 2011.

[5] J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, Jan. 2010.

[6] J. Kaye, C. Heeney, N. Hawkins, J. De Vries, and P. Boddington. Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics*, 10(5):331–335, 2009.

[7] J. Lubchenco. Entering the century of the environment: a new social contract for science. *Science*, 279(5350):491–497, 1998.

[8] R. K. Madduri, D. Sulakhe, L. Lacinski, B. Liu, A. Rodriguez, K. Chard, U. J. Dave, and I. T. Foster. Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. *Concurrency - Practice and Experience*, In press, 2014.

[9] J. Porter and J. Callahan. Circumventing a dilemma: historical approaches to data sharing in ecological research. *Environmental information management and analysis: ecosystems to global scales*, pages 193–202, 1994.

[10] C. Rosenzweig, J. Jones, J. Hatfield, A. Ruane, K. Boote, P. Thorburn, J. Antle, G. Nelson, C. Porter, S. Janssen, et al. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology*, 170:166–182, 2013.

[11] 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*, 190(3):214–251, 1997.

[12] P. E. Thornton, M. M. Thornton, B. W. Mayer, N. Wilhelmi, Y. Wei, R. Devarakonda, and R. B. Cook. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. Technical report, Oak Ridge National Laboratory (ORNL), 2014.

[13] M. Wilde, M. Hategan, J. M. Wozniak, B. Clifford, D. S. Katz, and I. Foster. Swift: A language for distributed parallel scripting. *Parallel Computing*, 37(9):633–652, 2011.

[14] N. Wilkins-Diehr. Science gateways – common community interfaces to grid resources. *Concurrency and Computation: Practice and Experience*, 19(6):743–749, 2007.

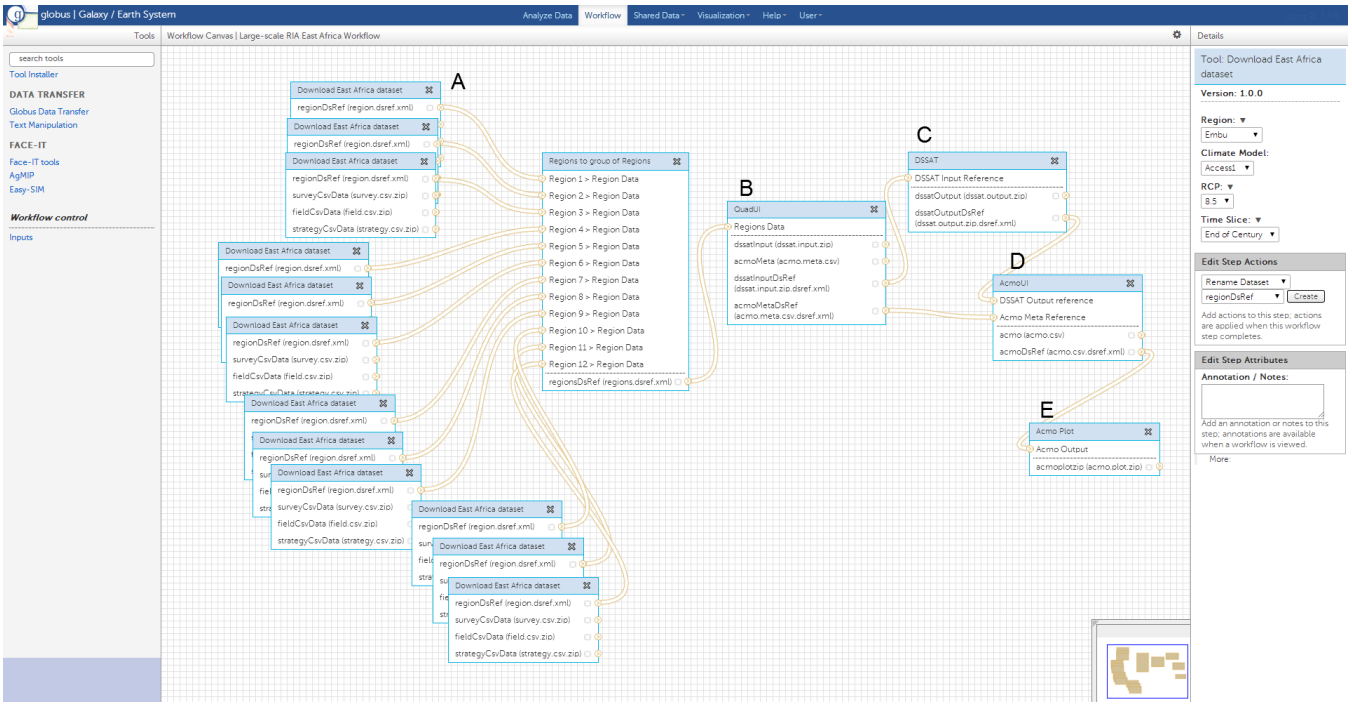

Figure 2: Screenshot of an AgMIP workflow in FACE-IT using the Galaxy workflow canvas and data from the East Africa RIA team: A) data ingest, B) input processing, C) crop/climate impact simulation, D) output processing, E) visualization. This workflow can be modified, published, shared, and reproduced—all on remote resources.

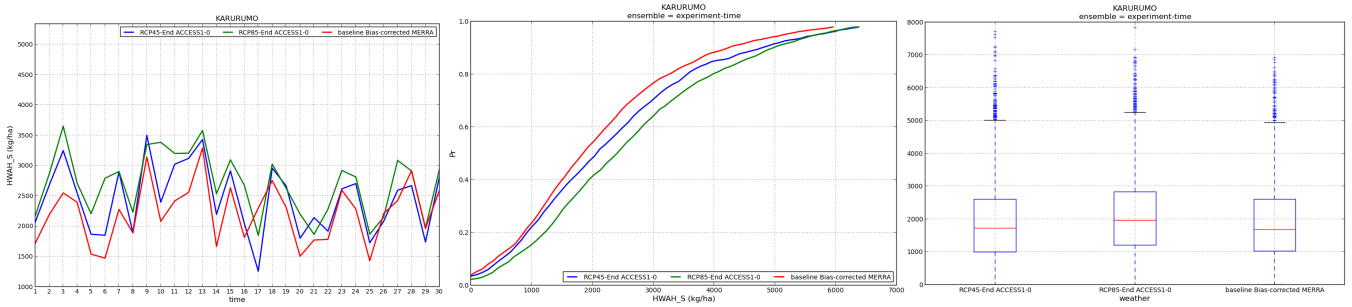

Figure 3: Plots of simulated outputs from the East Africa workflow for two climate scenarios and one region. Each plot includes 9600 simulations representing 320 sites in the Embu region and 30 years of weather data. (a) Boxplot showing comparative distribution of harvested yield. (b) Cumulative probability distribution for biomass production. (c) Average simulated yield for the 320 sides.