CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3540

SPECIAL ISSUE PAPER

FACE-IT: A science gateway for food security research

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

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

SUMMARY

Progress in sustainability science is hindered by challenges in creating and managing complex data acquisition, processing, simulation, post-processing, and intercomparison pipelines. To address these challenges, we developed the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT) for crop and climate impact assessments. This integrated data processing and simulation framework enables data ingest from geospatial archives; data regridding, aggregation, and other processing prior to simulation; large-scale climate impact simulations with agricultural and other models, leveraging high-performance and cloud computing; and post-processing to produce aggregated yields and ensemble variables needed for statistics, for model intercomparison, and to connect biophysical models to global and regional economic models. FACE-IT leverages the capabilities of the Globus Galaxies platform to enable the capture of workflows and outputs in well-defined, reusable, and comparable forms. We describe FACE-IT and applications within the Agricultural Model Intercomparison and Improvement Project and the Center for Robust Decision-making on Climate and Energy Policy. Copyright © 2015 John Wiley & Sons, Ltd.

Received 20 January 2015; Revised 27 March 2015; Accepted 15 April 2015

KEY WORDS: science gateways; climate impacts and food security; globus galaxies

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: agmip.org) [1] illustrates key 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

*Correspondence to: Joshua Elliott, Computation Institute, Argonne National Laboratory and University of Chicago, Illinois, USA.
†E-mail: jelliott@ci.uchicago.edu

capacity in both the developing and developed world. Research groups in AgMIP have widely varying backgrounds, motivations, and access to technical resources such as computation, information technology (IT) expertise, and network connectivity. For most groups, even conceptually simple tasks, such as driving a set of models with output from several Coupled Model Intercomparison Project simulations, can become prohibitively complex because of a multiplicity of data formats, inadequate computational tools, difficulty in sharing data and programs, and large data sets. These barriers hinder both research and the rapid and effective transmission of new and existing knowledge to policy-makers and decision-makers [2].

Similarly, participants in the center for Robust Decision-making in Climate and Energy Policy (RDCEP: rdcep.org) work to evaluate the impacts of global change on physical and socio-economic systems at different spatial and temporal scales. RDCEP members developed the parallel System for Integrating Impacts Models and Sectors (pSIMS [3]) to perform high-resolution, massively parallel climate impact simulations. pSIMS executes large ensemble simulations, often driven by dozens of data sets from diverse sources (climate, soil, management, etc.) and using multiple impacts models. A single ensemble can produce Terabytes (TBs) of output. pSIMS requires large-scale compute, storage, and support resources that are not generally available to research groups in the environmental and agricultural sciences, especially in the developing world where many target users are located.

To address such challenges, the Framework to Advance Climate, Economic, and Impact Investigations with IT (FACE-IT) project has developed, and continues to develop, a cloud-based science gateway [4] that provides web-based access to a range of data projects, simulation models, and analysis tools [5]. In this paper, we review the technical challenges that the FACE-IT gateway aims to address, outline the FACE-IT approach, describe its components and architecture, give examples of early applications, and discuss lessons learned and future directions for our team and users.

2. TECHNICAL CHALLENGES

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

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 require completely new, integrated solutions that can be seen as a new branch of IT.

Inadequate software tools: Particularly when working with unfamiliar data, researchers can end up spending more time developing the knowledge and tools required to understand, analyze, aggregate, transform, and so on 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 are familiar. The required tools typically exist, but the researcher working with unfamiliar data does not know how to find, install, or use them.

Limited access to computational resources: In many cases, inability to access the high performance and high-throughput computing required to efficiently process and analyze data can provide a significant barrier to progress. Many users do not have access to existing clusters or the resources to purchase and maintain their own. Even once obtained, users must spend time learning to use the system and customizing their software to run in those specific environments. Use remote computing resources when required for more computationally intensive analysis.

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

DOI: 10.1002/cpe

and expect daily weather data in customized ASCII formats. Climate datasets assume Greenwich Mean Time; crop models typically assume local time. Such differences in syntax and semantics can make adapting data for a new purpose challenging and error-prone. Such problems can be addressed by providing translator programs that encapsulate all knowledge required to translate from one format to another and by typing datasets to permit discovery of when and what translators are needed. However, a lack of mechanisms for sharing and discovering translators means that researchers rarely use them.

Inadequate visualization tools: Larger datasets require more sophisticated visualization tools, capable of filtering, extracting, and presenting data in useable and meaningful ways. Users need to be able to configure and run processing pipelines that can partition and select a data subset, verify data correctness, preview results before computationally intensive analysis, and produce publication-quality figures to summarize results. In some cases, the graphic rendering of big data can itself be a computationally intensive process, requiring GPU-powered machines [6].

Difficulty in sharing data and programs: While some researchers are simply not interested in sharing code and data, many others 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 share [7].

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 [7, 8].

3. THE FRAMEWORK TO ADVANCE CLIMATE, ECONOMIC, AND IMPACT INVESTIGATIONS WITH INFORMATION TECHNOLOGY APPROACH

We believe that many of the challenges just listed can be overcome, in part at least, by the judicious use of IT. To this end, we aim with FACE-IT to provide a science gateway framework that will allow a community to combine in a single (virtual) location—a FACE-IT Instance—the following capabilities (Figure 1):

- A **data store** that collects large quantities of diverse data, organized into scientifically meaningful datasets annotated with rich metadata, type, and format information; powerful browsing and search capabilities to facilitate discovery of desired data; and robust access control mechanisms to encourage contributions from people with sensitive data.

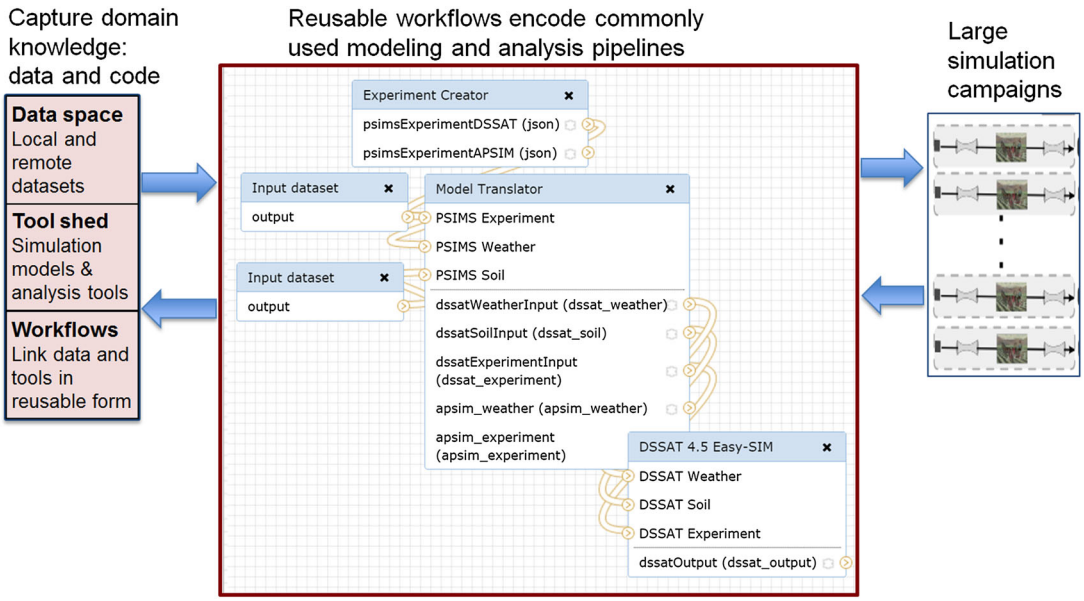

Figure 1. Elements of FACE-IT. Left: Capturing domain knowledge in the form of code and data. Center: reusable workflows encode commonly used pipelines. Right: running large simulation campaigns.

DOI: 10.1002/cpe

- Rich **program collections** for format conversion, analysis, visualization, and so on, 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.
- Access to powerful **cloud computing** and other computing resources to run FACE-IT tools.
- Convenient **data and code ingest mechanisms** for adding data and programs to the data store, with clear records of provenance and automated metadata extraction and synthesis for ingested objects, to simplify subsequent discovery.
- Rich **social elements** to incentivize contributions, via recognition of popular datasets and programs.

We thus aim to allow researchers across multiple social and earth science disciplines to share not only data but also the software tools used to create, manipulate, analyze, and visualize that data. We intend that they 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 also allow users to share and publish data and results as Data Libraries, rich analysis pipelines (User Histories), customizable multi-step pipelines (Workflows), or 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).

4. ARCHITECTURE, IMPLEMENTATION

FACE-IT builds on the Globus Galaxies platform [9], which has been developed over the past several years at the University of Chicago, initially in support of the Globus Genomics project [10]. We also benefit from substantial work by partner communities (see §5.2), who have developed the many domain-specific tools that populate FACE-IT.

4.1. Architecture overview

The Globus Galaxies platform leverages Galaxy [11, 12] for its simple, uniform, and extensible workflow authoring and execution interface; Globus services for data movement [13] and for identity, group, and profile management [14]; Swift [15] for parallel execution of multiple pipelines in ensemble simulations; and custom elements [10] for elastic, scalable cloud execution and for execution on grids and parallel computers. These capabilities enable the rapid development of cloud-hosted science gateways that support community access to, and exchange of, complex data and computational tools.

Galaxy [11, 12], developed by the Nekrutenko lab in the Center for Comparative Genomics and Bioinformatics at Penn State, the Taylor lab at Emory, and other contributors, is widely used in genomic science, but has not previously been applied in the social and earth sciences. Galaxy allows for simple data upload, via Web tools, to user and community libraries; the selection of tools from extensible 'toolsheds'; the linking of tools, based on data types, in pipelines; and the sharing of data, tools, and pipelines with others.

The Globus Galaxies platform on which FACE-IT is based was originally developed to run on a collection of Amazon EC2 virtual machine instances ('nodes'). It supports diverse configurations to match the varying needs of specific use cases and applications. In its normal configuration, it runs Network File System (NFS) across the 'head node' (used to run the Galaxy server, Globus Connect server, and NFS server) and the (dynamically scaled) pool of worker node used to perform computations—although simple demo instances can run everything, including applications, on a single node. The Globus Connect Server supports file transfers from and to external world, and the NFS deployment permits data exchange between the head node and workers. FACE-IT is able to use these various features of the Globus Galaxies platform 'out of the box', with all management services on a head-node AWS on-demand instance and spot instances used for worker nodes.

DOI: 10.1002/cpe

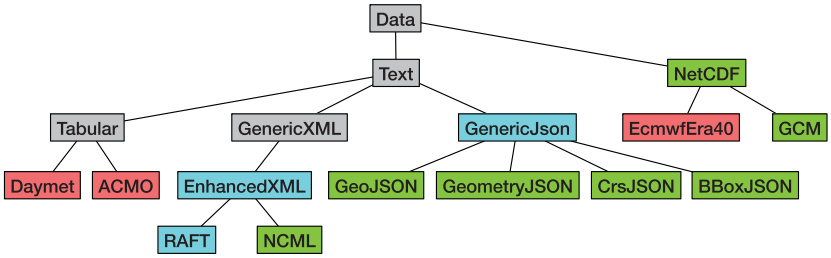

Figure 2. Some of the new data types implemented for the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT). Nodes higher in the tree are base types. Grey boxes are regular Galaxy data types; blue are enhanced data types (§4.2.1); green are new geospatial data types (§4.2.2), and red are new FACE-IT data types (§4.2.3).

The Globus Galaxies platform can also be configured to run applications on high-throughput computing infrastructures such as Open Science Grid (OSG) and on classical high-performance computing clusters and supercomputers. In those environments, the use of NFS for file access is replaced with file staging managed by Swift.

4.2. New and expanded datatypes

An important part of our work in developing FACE-IT was to extend Galaxy with the new datatypes depicted in Figure 2, in order to be able to represent data of interest to our target communities. We also integrate a suite of data transformation, data analysis, and simulation tools that implement important actions on those data types.

In the Galaxy platform, every piece of data has an associated datatype. When new data are added to Galaxy, the Galaxy software attempts to automatically detect the datatype through a process called sniffing. Datatype definitions include information about how to identify if a new file matches the datatype. If the sniffing function fails to correctly identify a file, a Galaxy user may manually change its associated datatype.

Every tool in Galaxy specifies the datatype(s) of the file(s) that it accepts as input(s), and the datatype(s) of the file(s) that it creates as output(s). This use of strongly typed data not only allows Galaxy to understand how tools can be chained together into workflows but also permits Galaxy to perform some amount of automated type conversion—a feature that we exploit, for example, when passing Daymet-format files to tools that require NetCDF inputs (see the following).

We describe here the three primary areas in which we introduced new or modified datatypes.

4.2.1. Extensions of general datatypes. We first describe four enhancements to existing Galaxy datatypes that we introduced in order to meet user needs.

First, we added an enhanced XML datatype to provide full support for XML Schemas and XML transforms. Support for XML Schemas allows us to make datatype sniffing completely automatic. The development of new XML-based datatypes thus becomes straightforward, requiring only the addition of schema and transform files for the new subtype. We also implemented a standard data peek and data display using two type-specific XML transform configurations to be provided by the developer. This approach allows data to be rendered in HTML directly without Python additions.

Second, agricultural researchers often use the Javascript Object Notation (JSON) file format to support model setups and outputs that are completely unstructured or that are defined differently by different subcommunities. We thus developed a JSON datatype supporting the relatively new JSON Schema standard, so that we can handle JSON datatypes similarly to XML datatypes. We implemented a version of JSONT (a transform protocol for JSON similar to XML Transform) for data display and data peek as done with XML. Thus, we have made creating a JSON-based datatype straightforward. Essentially, no Python code is required, just schema and transform definitions.

DOI: 10.1002/cpe

Third, compressed and composite datasets are crucial for many of our applications. For example, an ESRI shapefile is a composite dataset, comprising multiple compressed files. Galaxy already supports composite datatypes and some compressed datatypes. We extended this support to encompass composite-compressed datatypes, implementing component to support compressed and composite dataset management, upload, and automatic type detection.

Fourth, many FACE-IT applications need to be able to consume or produce multiple datasets, especially when running within ensembles. Thus, we introduced the Reference Aggregation File Type (RAFT) datatype, an enhanced XML type whose elements are references to other datasets. Constituent datasets can be of the same or different datatypes. The RAFT datatype is well suited for datasets whose constituent elements may be stored at different locations, for example, in the cloud or at different institutional repositories. The references to those elements can then be URLs.

4.2.2. Geospatial and temporal data types. We introduced new Galaxy data types to support data in the NetCDF and GeoJSON formats, both of which have geospatial and temporal dimensions.

The NetCDF standard is widely used in the earth-system community for the multidimensional storage of dense matrices. This self-describing, binary, machine-independent data format is supported by a set of libraries for the creation, access, and sharing of array-oriented scientific data. In FACE-IT, we use both the Common Definition Language (CDL) and NetCDF Markup Language (NcML) metadata conventions, and introduce the concept of 'NetCDF Schema', implemented as an NcML header representation that allows regular expressions in dimensions and variable definitions. Thus new NetCDF-based data types can be introduced simply by creating the appropriate NetCDF schema.

GeoJSON [16] is a JSON format for encoding a variety of geographic data structures, that is, data for which geo-referenced information is available for each record. A GeoJSON object may represent geometry, a feature, or a collection of features. Features in GeoJSON contain a geometry object and additional properties, and a feature collection represents a list of features. GeoJSON supports Point, LineString, Polygon, MultiPoint, MultiLineString, MultiPolygon, and GeometryCollection geometry types.

Having introduced these two new Galaxy data types, we implemented map-based visualizers as client-side visualizations that interact with a Galaxy data provider. Thus, FACE-IT users can create interactive data output, at least in situations in which the amount of data to be transferred from the server to the client are compatible with the HTTP-based user experience and technical requirements/constraints. Thus, for example, if a NetCDF file's schema is recognized as plottable on a map, then FACE-IT creates an interactive map viewer within the Galaxy UI. This viewer is based on the World Map Service (WMS) and leverages a customized version of the open source Sci-WMS service [17].

4.2.3. New datatypes for FACE-IT applications. We also added several new Galaxy datatypes for data formats that are specific to various FACE-IT applications. For example, we have created a new Daymet datatype to represent the climate files available in the Daymet daily surface weather dataset [18, 19]). We incorporated into this text-based datatype a built-in auto-converter to NetCDF, so that any tool that accepts NetCDF climate data may also use Daymet data without modifications.

The EcmwfEra40 datatype is an example of NetCDF subclassing, using NetCDF Schema and WMS map rendering. This data represent surface data from the ECMWF 40 years re-analysis [20].

A third example of a new FACE-IT-specific datatype is ACMO, used to capture data stored in the AgMIP Crop Model Output (ACMO) format. ACMO is used in AgMIP as a way of harmonizing the output of crop simulations. FACE-IT provides tools for visualizing ACMO files, which makes it easy to visually compare the results of different simulation scenarios or see how different crop models can vary in their results.

4.2.4. Importing data from external sites. We have worked with earth-system data providers to make them FACE-IT-compliant, meaning that their data can be transferred into a user's FACE-IT

DOI: 10.1002/cpe

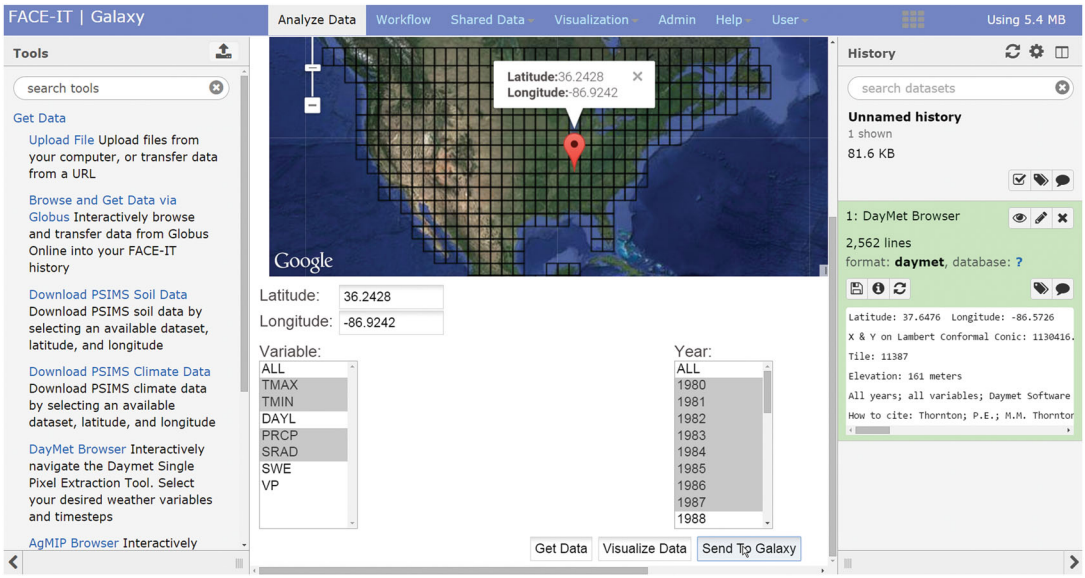

Figure 3. Screenshot of the Framework to Advance Climate, Economic, and Impact Investigations with Information Technology (FACE-IT)/Galaxy interface for the Daymet data browser. Users interact with this interface to select variables, years, and location, and then send the resulting data directly to FACE-IT/Galaxy where it appears in the history as a data object and can be immediately used in applications and workflows.

workspace with a single click. External data sources such as Daymet can be modified to make file transfer easier with the addition of a few lines of Javascript to submit an HTTP POST request. This modification is simple to perform and has no effect on the data interface, unless users browse to the data source through the FACE-IT platform.

Figure 3 shows an example of the FACE-IT interface for the Daymet data browser. Users interact with the familiar Daymet interface (center panel) to select the desired variables, years, and spatial location and then select the 'Send to Galaxy' button to send the resulting data directly to FACE-IT. When the data are transferred, FACE-IT automatically determines the datatype through Galaxy's sniffing process. It then appears in the Galaxy history (right panel) as a data object and can be immediately used by tools and workflows.

4.3. Loosely coupled parallelization

Some individual FACE-IT applications are computationally intensive, while others must be run many thousands or even millions of times. Thus, as explained earlier, we leverage Globus Galaxies support for dynamic scaling of the pool of worker instances. When a FACE-IT application is launched, a job is submitted to an HTCondor scheduler. A custom python daemon runs in the background and monitors the HTCondor queue. The daemon looks for jobs containing a specific ClassAd, which indicates the job should request additional EC2 resources. When a job with this ClassAd is found, a new EC2 spot instance is started where the application will run.

For finer-grained control of parallelism, FACE-IT applications may use the Swift parallel scripting language. In this case, the HTCondor job(s) started will not be the FACE-IT application itself, but rather, Swift will launch a pilot job that can then launch any other number of tasks. In situations where hundreds or thousands of tasks will be running, this can greatly reduce the wait times and overhead associated with a traditional job scheduler like HTCondor.

As previously stated, FACE-IT works well on AWS but is not limited to IaaS clouds. A deployment on Open Science Grid or similarly structured resources is possible by extending Galaxy's job runner script to handle job submission, stage-in, and stage-out [21]. A similar approach could be used with a remote scheduler running on a dedicated high-performance computing cluster not connected to the FACE-IT Galaxy head-node at NFS level.

DOI: 10.1002/cpe

5. APPLICATIONS

We intend that FACE-IT accelerate discovery both *within* communities, such as AgMIP and pSIMS, and *across* disciplines such as environmental and economic sciences. All tools presented in §5.1 and §5.2 are available in the FACE-IT instance and can be used interactively upon registration as a Globus user. In the next stages of FACE-IT development, we will wrap all the tools and the core itself in a Toolshed to make them freely available to interested users.

Application, datatype, and pipeline development within each of our two test communities are guided by established use-cases. These use-cases reflect common tasks and pipelines that are executed frequently and repeatedly by researchers in the community and are amenable to automation in FACE-IT. We describe the objectives of these communities, example use-cases, and the applications and pipelines used in FACE-IT in §5.1 and §5.2.

Other environmental scientists communities are planning to use, or are already using, FACE-IT for their applications; these applications have been developed from scratch by those user groups. For example, the Predictive Ecosystem Analyzer group at the University of Illinois at Urbana-Champaign [22] is working on applications in FACE-IT that will allow them to encode their current workflows for ecological model calibration and bayesian uncertainty estimation. Another researcher is evaluating FACE-IT to build a production workflow for extreme weather event simulations. In §5.3, we present findings from our work with the AgMIP and pSIMS user groups that we believe are generalizable to the wider FACE-IT community and to other science gateway applications.

5.1. Parallel system for integrating impacts models and sectors

Through a set of tools, datatypes, and workflows developed as a toolshed in FACE-IT, we have implemented a single-field, FACE-IT-based version of the pSIMS platform called Easy-SIMS. Easy-SIMS serves as an entry point for users who may not be familiar with either crop modeling or the FACE-IT platform. New users can explore models and data common to agricultural impacts research, and advanced users can rapidly prototype a suite of scenarios in order to run pSIMS for their local regions (support for regional pSIMS simulations at river basin, state, country, or similar scale is being added in the next version of FACE-IT). Easy-SIMS also functions as an educational platform and has been used in student research projects by interns at the University of Chicago.

We provide data selection, translation, experiment construction, and model execution tools. The Easy-SIMS toolshed supports a fixed, simple set of simulation configurations, allowing users to select from among a small set of key environmental, technological, and management options for driving the simulations, The Easy-SIMS workspace currently supports the popular DSSAT and APSIM families of crop models and enables simulations for six globally important crops (maize, soy, wheat, rice, millet, and sorghum). Easy-SIMS provides access to a wide array of global and regional soil and climate datasets and applications. Each application has been wrapped into a tool in Galaxy and placed in an Easy-SIMS Toolshed. Those supported with the first version of Easy-SIMS include the following:

Download soil data from a remote location into your history, given a specified dataset, latitude, and longitude.

Download weather data from a remote location into your history, given specified dataset, latitude, and longitude.

Delta shift generates a climate scenario from a pSIMS weather file by applying a mean and/or variance correction calculated from the output of one or more climate models.

Create experiment produces a model-agnostic experiment file containing information about the simulation to be run. This file includes information such as the model to be used, the crop to be simulated, start and stop times, irrigation options, and fertilizer application configurations.

Translate: The *Download Weather Data*, *Download Soil Data*, and *Create Experiment* tools create data files formatted to work with the pSIMS framework. Because each model has its own unique data formats, the *Translate* tool converts weather, soil, and experiment data from pSIMS format into model-specific input formats.

DOI: 10.1002/cpe

Run DSSAT 4.5: The popular Decision Support System for Agrotechnology Transfer (DSSAT) family of crop models can be used to simulate the impact of environment and management on crops. It accepts as input DSSAT formatted weather, soil, and experiment files. An optional debug flag will cause DSSAT to bring the WARNINGS.OUT log file into your FACE-IT history.

Run APSIM 7.5: The Agricultural Production Systems Simulator is a farming systems model that simulates the effects of environmental variables and management decisions on crop yield, profits, and ecological outcomes.

Plot output plots a variable contained in an Easy-SIM created NetCDF file.

5.2. Agricultural model intercomparison and improvement project

During the first phase of FACE-IT development, we have worked closely with the AgMIP community to create, test, and distribute applications, data, and pipelines that will both advance AgMIP research and enable at-scale evaluation of the FACE-IT approach. The core use-case that has driven application development and testing thus far can be described as follows:

> *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 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 these 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.*

Many crop modeling simulations are required in order to evaluate current climate and technology conditions, future climate conditions with current technology, and future climate conditions with adaptation. Each system is simulated for multiple climate models and climate scenarios, multiple crop models, and multiple sites-years. These simulations are evaluated, compared, and used as input to regional economic models. For example, in a Regional Integrated Assessment study for the Nioro region in Senegal, we conducted five sets of simulations using maize survey data:

- Historical: The conditions that were surveyed, for only the year (2007) of survey conditions for the maize crop; conducted once for each crop model.
- Current: 30 years using current climate and production management system; conducted once for each crop model.
- Future: 30 years using future climate with current production management system. Five scenarios represent future climates generated by different GCMs and for different time slices.
- Current + Adaptation: 30 years of current climate with a climate-adapted management system. This simulation set represents the same simulations as in the Current analyses, but may include multiple adaptation scenarios. In the Nioro cas,e the adaptation management system required modifications to soil profiles as a means to simulate a drought resistant cultivar.
- Future + Adaptation: 30 years of future climate with a climate-adapted management system. These sets represent the same simulations as in the Future analyses, but may also include multiple adaptation scenarios.

As shown in Table I, users must upload 35 compressed files (zip) to perform these crop modeling analyses. (Some files, e.g., for climate and survey data, can be reused for different sets of simulations. Data are categorized into four groups: Survey, Field_Overlay, Seasonal_Strategy, and Linkage. Survey data consist of AgMIP site-based data from farm surveys, yield trials, variety trials, and detailed field experiments. Field Overlay and Seasonal Strategy are both AgMIP Data Overlay

DOI: 10.1002/cpe

Table I. List of files used in demo workflow for Nioro, Senegal maize analysis.

Filenames	Historical	Current	Future	Current + Adaptation	Future + Adaptation
Survey_data-Nioro-MAZ.zip	X	X	X		
Survey_data-Nioro-MAZ.zip	X	X	X		
Weather-Nioro-0XFX.zip	X	X		X	
Linkage-Nioro-MAZ-historical.csv	X				
Seasonal_strategy-Nioro-MAZ-0XFX.zip		X			
Linkage-Nioro-MAZ-0XFX.csv		X			
Weather -Nioro-IxFA.zip			X		X
Seasonal_strategy-Nioro-MAZ-IxFA.zip			X		
Linkage-Nioro-MAZ-IxFA.csv			X		
Survey_data-Nioro-MAZ-Ax.zip				X	X
Field_Overlay-Nioro-MAZ-Ax.zip				X	X
Seasonal_strategy-Nioro-MAZ-0XFX-Ax.zip				X	
Linkage-Nioro-MAZ 0XFX-Ax.csv				X	
Seasonal_strategy-Nioro-MAZ-IxFA-Ax.zip					X
Linkage-Nioro-MAZ-IxFA-Ax.csv					X

for Multi-model Export (DOME) data, which allow modelers to apply assumptions uniformly to multiple models. Field Overlay DOMEs are used to provide data that are required by crop models, but not supplied in survey data. Seasonal Strategy DOMEs allow multi-year simulations for current and future climate conditions, using planting date rules and other imposed management regimens. Linkage files are used to associate Field Overlay and Seasonal Strategy DOMEs to each location in the survey data, which allow Survey data and Field Overlay data to be used for multiple scenarios. In the Nioro case, all data are converted from spreadsheet templates to comma-delimited format and then to zip archives.

In order to accomplish the goals of the AgMIP regional integrated assessment teams, the AgMIP IT team has developed a number of applications for FACE-IT, ranging from the simple to the sophisticated, including the following:

- **QuadUI** translates user data into AgMIP Crop Experiment (ACE) format using JSON data structures.
- **Data combinator** combines and transforms the different formats of data into a single zip archive and/or a unified ACE Binary format that is used for QuadUI translation.
- **AcmoUI** converts crop-model specific output formats into ACMO format for standardized comparison, analysis, and visualization.
- **Plot AcmoOutput relationship**, written in R, enables the user to explore the correlation between different ACMO output variables.
- **Plot climate scenarios**, also written in R, plots climate anomalies for single or multiple-climate file inputs. Figure 4 shows one example of its output: boxplots of monthly and seasonal climate anomalies from dozens of climate models for a collection of AgMIP RIA test sites in Nioro, Senegal.
- **Standardplots_ria** contains R scripts that generate box and CDF plots for single or multiple ACMO output files. Users can interactively define the type, name, variables, and colors of the plot.
- **DSSAT 4.5 plus** includes version 4.5.1.23 of the DSSAT family of crop models (see 5.1 for details) plus data translation from the AgMIP harmonized ACE (JSON) format to DSSAT input formats, and data translation from DSSAT output formats to ACMO format.
- **APSIM 7.5 plus** includes the APSIM model (see §5.1 for details) plus data translation from the AgMIP harmonized ACE (JSON) format to APSIM input formats, and data translation from APSIM output formats to ACMO format.

Figure 5 shows the FACE-IT demo workflow for the Nioro, Senegal maize analysis. For facilitating the reuse of weather data, we use a data translator to combine the uploaded survey data and weather data, generating specific survey data inputs for different sets of simulation. For the Nioro

DOI: 10.1002/cpe

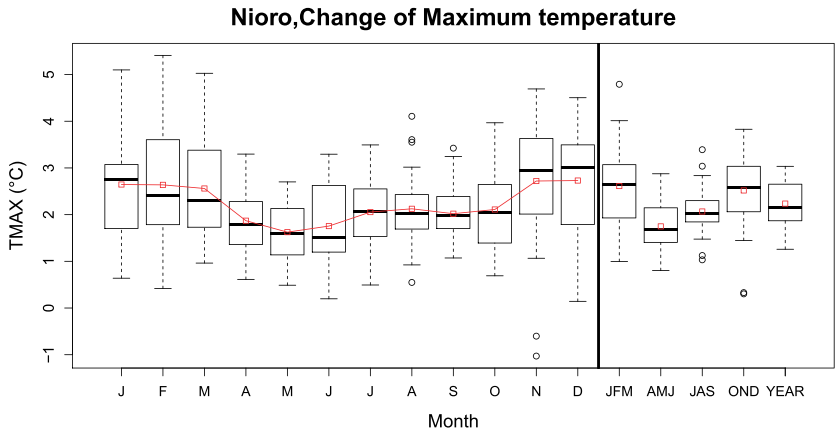

Figure 4. Example of the output of 'plot climate scenarios' app: boxplots of the monthly and seasonal climate anomalies from dozens of climate models for a collection of the Agricultural Model Intercomparison and Improvement Project Regional Integrated Assessment test sites in Nioro, Senegal.

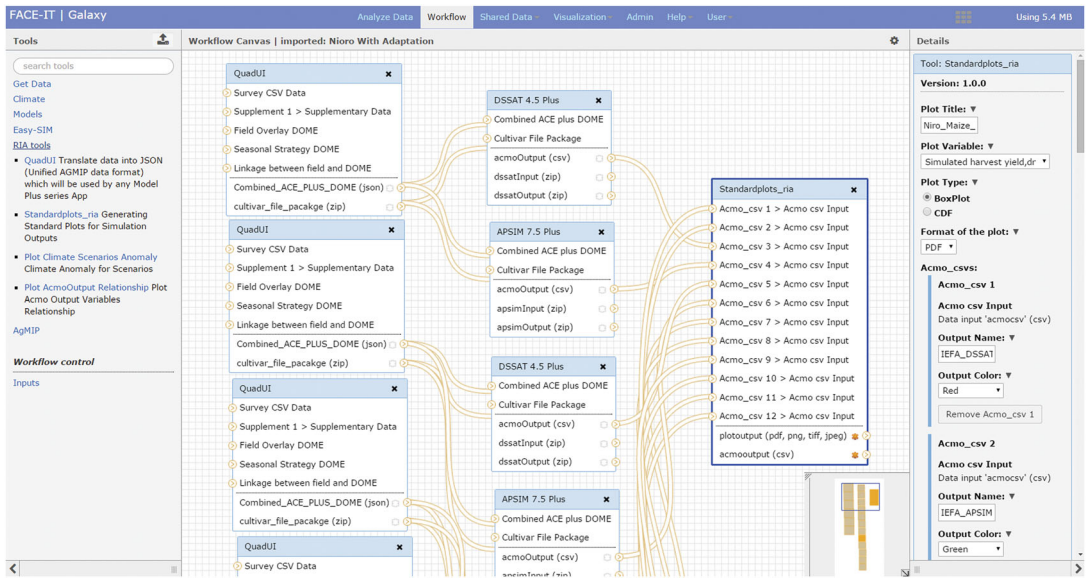

Figure 5. Screenshot of an Agricultural Model Intercomparison and Improvement Project workflow using the Galaxy workflow canvas and data from the East Africa Regional Integrated Assessment team: A) data ingest, B) input processing, C) crop/climate impact simulation, D) output processing, E) visualization. This workflow can be modified, published, shared, and reproduced on remote resources.

demo, there are 100 sites, each simulated for seven scenarios (2007, current, and five future climate scenarios), and two management scenarios (without and with adaptation). QuadUI converts CSV-formatted data to AgMIP harmonized format, then to DSSAT model format. At this point, additional models (e.g., APSIM), each with unique input formats, can also be added. DSSAT calls the model to perform simulations. Harmonization of model outputs is made using AcmoUI app. Generation of plots from the simulated data is carried out with the Standardplots_ria app.

Figure 6 illustrates two plots generated from the Nioro demo workflow. These plots are typical of the kinds of analyses performed by AgMIP teams in Sub-Saharan Africa and South Asia and are directly publishable in the teams reports and summaries. Figure 6a shows a boxplot of harvested yields for partial sets of simulation, in which each plot is generated from simulated yields from 100 sites and 30 years (for historic only one year). Figure 6b shows the same data in a cumulative probability graph. Other output variables can be interactively selected and regenerated

DOI: 10.1002/cpe

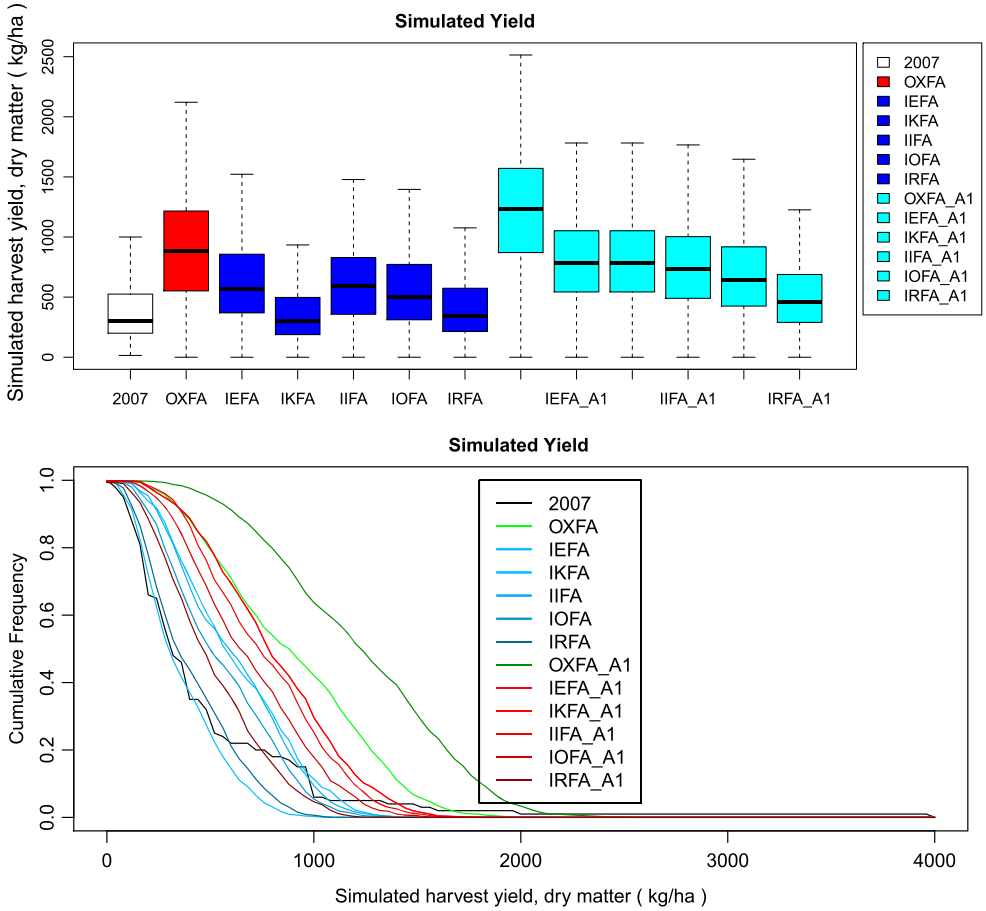

Figure 6. Outputs from the Nioro workflow for different climate scenarios and management regimes. Each plot includes multiple simulations representing 100 sites in Nioro and 30 years of weather data (one year for 2007 simulation). Top: Boxplot showing simulated harvested yield. Bottom: Cumulative probability distribution for harvested yield. Simulations marked A1 include climate change adaptation.

through the interface of the `Standarplots_ria` app. In the past, each AgMIP regional research team generated similar but not directly comparable graphs using different methods and tools. The FACE-IT workflow allows much greater standardization of analysis and comparability among regions and models while still allowing flexibility and customization in aggregation of data by each team.

5.3. Expanding to new use cases

The promise of standardization and reproducibility is an incentive for large-scale distributed projects like AgMIP and RDCEP to use FACE-IT to encode frequently used applications and pipelines. Central access to data, tools, and pipelines that are common to these communities (as well as remote resources to run these tools at scale) then becomes a powerful incentive for community members to engage the platform for their own research. Diverse groups of users elaborate existing data and tools to create research ecosystems that benefit everyone.

One important aspect of the platform design that has emerged from our early work with user communities is that users share technical expertise through the platform. The general user of FACE-IT is a field scientist deeply involved in their environmental and social research. These users create simulation experiments as FACE-IT workflows, selecting tools from a palette and making connections among tools describing the data flow. They generally are not concerned with the complexities of tool and datatype development. They use FACE-IT because the tools and resources are

DOI: 10.1002/cpe

there; each action performed in their data analysis is reproducible; and data, metadata, information pages, maps, charts, data sources, and libraries are sharable with the whole community or a group of colleagues.

FACE-IT power users, on the other hand, may be field scientists, computer scientists, or others with a strong background in computing, data management, or high-performance computing. Power users are the core of the FACE-IT approach: they extend and customize the FACE-IT environment in order to permit regular users within their research groups or in related subcommunities to increase their productivity. Power users develop new FACE-IT tools and create new datatypes by wrapping data formats ingested or produced by tools.

This hybrid-expertise model allows FACE-IT communities to become self-sustaining quickly. And by allowing individuals to focus on the parts of their science or technical applications in which their primary skills exist, it permits research groups within the community to work more efficiently.

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 forms can dominate research efforts. By providing ready access not only to 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 extend their work in new directions and 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 integration of the Easy-SIMS workspace for climate impacts assessment by RDCEP researchers. FACE-IT will make high-resolution regional vulnerability and impact and adaptation assessments available to any researcher with a web connection by leveraging high-performance and high-throughput computing resources provided by (but not limited to) Amazon Web Services and best-available datasets from around the web, including a large number of data options curated by RDCEP. By working to build the capacities of pSIMS into the FACE-IT platform, RDCEP members hope to enable a wider user community to access data, develop and deploy pSIMS workflows, and publish the results.

Using FACE-IT, members of the AgMIP network can now create powerful tools and workflows to streamline the execution of tens of thousands of crop model simulations for regional integrated assessment studies being conducted by dozens of researchers in Sub-Saharan Africa and South Asia. Initial applications to food security studies in Africa represent a first step towards broad adoption within the international AgMIP community.

We continue to enhance and expand FACE-IT. We are adding, for example, the ability to provide live WMS maps rendered from NetCDF files using an external service with a REST web API. We also continue to expand the methods used to describe and type data, for example by exploiting our NetCDF Schema and dataset reference datatype approaches.

During the initial phase of development with the user AgMIP and RDCEP user communities, we have learned many useful lessons about how highly diverse user communities approach a technical platform like FACE-IT. For example, a well designed interface with social tools for sharing and commenting enables a hierarchy of technical skill levels to emerge which we have embraced to provide improved usability and access to our target communities. We have identified at least two types of users that help to create an appealing and robust environment: power users extend and customize the FACE-IT environment using technical skills and experience which regular users create their experiments as FACE-IT workflows selecting tools from a palette and making connections among tools describing the data flow.

DOI: 10.1002/cpe

ACKNOWLEDGEMENTS

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, the NSF Decision Making Under Uncertainty program award 0951576, and the DOE under contract DE-AC02-06CH11357. EC2 resources have been generously provided by Amazon.

REFERENCES

1. Rosenzweig C, Jones J, Hatfield J, Ruane A, Boote K, Thorburn P, Antle J, Nelson G, Porter C, Janssen S, et al. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology* 2013; **170**:166–182.
2. Lubchenco J. Entering the century of the environment: a new social contract for science. *Science* 1998; **279**(5350):491–497.
3. Elliott J, Kelly D, Chryssanthacopoulos J, Glotter M, Jhunjhnuwala K, Best N, Wilde M, Foster I. The parallel system for integrating impact models and sectors (pSIMS). *Environmental Modelling & Software* 2014; **62**:509–516.
4. Wilkins-Diehr N. Science gateways – common community interfaces to grid resources. *Concurrency and Computation: Practice and Experience* 2007; **19**(6):743–749.
5. Montella R, Brizius A, Elliott J, Kelly D, Madduri R, Maheshwari K, Porter C, Vilter P, Wilde M, Xiong W, Zhang M, Foster I. Face-it: A science gateway for food security research. In *Proceedings of the 9th Gateway Computing Environments Workshop, GCE '14*, Piscataway, NJ, USA, 2014; 42–46. IEEE Press.
6. Giunta G, Montella R, Agrillo G, Coviello G. A GPGPU transparent virtualization component for high performance computing clouds. In *Euro-Par 2010-Parallel Processing*. Springer: Berlin Heidelberg, 2010; 379–391.
7. Porter JH, Callahan JT. Circumventing a dilemma: historical approaches to data sharing in ecological research. In *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Michener WK, Brunt JW, Stafford SG (eds). Taylor & Francis: London, England, 1994; 193–202.
8. Kaye J, Heeney C, Hawkins N, De Vries J, Boddington P. Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics* 2009; **10**(5):331–335.
9. Madduri R, Chard K, Chard R, Kelly D, Dave U, Foster I. The Globus Galaxies platform: delivering science gateways as a service. *Concurrency and Computation – Practice and Experience* 2015. DOI: 10.1002/cpe.3486.
10. Madduri RK, Sulakhe D, Lacinski L, Liu B, Rodriguez A, Chard K, Dave UJ, Foster IT. Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. *Concurrency and Computation – Practice and Experience* 2014; **26**(13):2266–2279.
11. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. *Current Protocols in Molecular Biology* 2010; **10**:1–21.
12. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 2010; **11**(8):R86.
13. Foster I. Globus Online: accelerating and democratizing science through cloud-based services. *IEEE Internet Computing* 2011; **15**(3):70–73.
14. Ananthakrishnan R, Chard K, Foster I, Tuecke S. Globus platform-as-a-service for collaborative science applications. *Concurrency and Computation – Practice and Experience* 2015; **27**:290–305.
15. Wilde M, Hategan M, Wozniak JM, Clifford B, Katz DS, Foster I. Swift: a language for distributed parallel scripting. *Parallel Computing* 2011; **37**(9):633–652.
16. Butler H, Daly M, Doyle A, Gillies S, Schaub T, Schmidt C. *The GeoJSON format specification*, 2008. (Available from: http://geojson.org/geojson-spec.html) [accessed on January 2013].
17. Howlett E, Signell RP, Wilson D, Snowden DP, Knee KR. Data management update for the integrated ocean observing system (ioos®). In *Oceans-St. John's, 2014*, 2014; 1–10. IEEE.
18. Thornton PE, Running SW, White MA. Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology* 1997; **190**(3):214–251.
19. Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB. *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. Data set*, 2014. (Available from: http://daac.ornl.gov) [accessed on January 2014], Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.
20. N. C. for Atmospheric Research Staff. *The climate data guide: ERA40*, 2014. (Available from: https://climatedataguide.ucar.edu/climate-data/era40) [accessed on January 2014].
21. Hayashi S, Gesing S, Quick R, Teige S, Ganote C, Wu LS, Prout E. Galaxy based BLAST submission to distributed national high throughput computing resources. *Presented at the International Symposium on Grids and Clouds (ISGC) 2014*, Taipei, Taiwan, 23–28 March 2014. Appears in Proceedings of the International Symposium on Grids and Clouds, PoS ISGC2014 (025).
22. LeBauer DS, Wang D, Richter KT, Davidson CC, Dietze MC. Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs* 2013; **83**(2):133–154.

DOI: 10.1002/cpe