# EarthCube: Scientific Workflows with Open Community Software

## Category and Summary

This letter expresses our intent to pursue supplemental funding to address issues identified in the "Critical Milestone" category of EarthCube. We propose to add three new team members from the geoscience community to the Open Gateway Computing Environments (OGCE) project (NSF OCI Award #1032742). Together, we will collaborate on workflow and distributed execution research problems that have both general and geoscience-specific requirements. This work primarily falls within EarthCube's "E" capability category (see http://earthcube.ning.com/page/capabilities), with additional relevance to capability categories A, B, and F. We also expect this work to explore important community governance models through the Apache Software Foundation that will have broader impact on the EarthCube community (Categories L and N). Our work will address all five desirable characteristics of EarthCube proposals (http://www.nsf.gov/geo/earthcube/eagerguidance.jsp).

## Scientific Motivation

**Scientific Workflows for Geoscience:** EarthCube has identified a number of desirable cyberinfrastructure capabilities for geoscience advancement, including many aspects of scientific workflows (discovery, provenance tracking, execution in distributed environments, data management, and controlled sharing and reuse). There are many solutions to these problems, including tools that concentrate on composition/expression (Taverna, Kepler), provenance (Karma), and high throughput of very large workflow graphs (Condor DAGman, Pegasus).

The Linked Environments for Atmospheric Discovery (LEAD) project (which included both Mattocks and Ramachandran on this team) pioneered comprehensive workflow approaches for integrating complex geoscience applications on distributed supercomputing resources. The OGCE project has extended LEAD's workflow infrastructure to make it a general-purpose workflow system for Grid- and Cloud Computing that supports both long-running and on-demand computing. The OGCE workflow tools, distributed through the Apache Airavata incubator, have already proven their utility in computational chemistry, astronomy, astrophysics, bioinformatics, and nuclear physics. We will focus on extending OGCE scientific workflow tools to address the specific requirements of geoscience applications. Our goals are to capture the use cases and requirements of geoscience researchers within this open source software framework. We expect the geoscience use cases to both reuse general capabilities and, more interestingly, introduce innovative new requirements, such as workflow complexity, real-time processing, event triggering, and human intervention into workflow executions. By exploring geoscience workflows within a general software framework, our project will have specific impacts within geoscience and broader impacts on other scientific domains.

**Open Community Software:** Tackling scientific workflow problems across diverse domains demands sustainable and well-governed software. We contend that we must go beyond simple open source, incorporating open community processes to do this. Our workflow software solutions will be developed through Apache Software Foundation processes as part of the Apache Airavata incubator. These processes go beyond the usual "code on the web" open source model to encourage diversity of stakeholders in a specific piece of software. These include a) well-defined processes for constructing a diverse Program Management Committee (PMC) with specific responsibilities; b) well-defined processes and incentives for adding new PMC members and code committers with write access (not just read access) to the source code; c) open, public decision making processes on project goals and design decisions; and d) good software engineering processes. Apache provides incentives to its member projects for adding new members with full voting and code contribution rights. We believe this distinguishes our effort from other workflow software efforts and from cyberinfrastructure software generally, offering the possibility of transforming the way cyberinfrastructure software is developed.

## Resulting Advancements

This proposal will extend the OGCE project to include three new project members. These build on existing collaborations, so we expect significant near term impact. From the current OGCE collaboration, Chathura Herath (senior researcher), Suresh Marru (OGCE Co-PI), and Marlon Pierce (OGCE PI) will participate. The *VLab project*, led by Prof. Renata Wentzcovitch, has developed sophisticated workflows that also must address high throughput computing requirements as part of their first principals materials science computations (da Silveira et al, TeraGrid 2011). These computations need to be brought to national-scale resources (XSEDE, OSG), which will be the focus of this collaboration. Pierce is a former Co-PI of the VLab project. The *Ocean Land Atmosphere Model (OLAM) gateway* project led by Dr. Craig Mattocks from the University of Miami's Rosenstiel School of Marine and Atmospheric Science will prototype the workflow system to integrate the global circulation/climate prediction model with local high-resolution land surface characteristics databases. The model's unique, flexible mesh refinement capability orchestrated with high resolution data will generate regional climate change projections with fine-grained resolutions and will provide rainfall projections that can aid in planning, water management and decision-making to ensure that threatened communities are sustainable and fragile ecosystems are more resilient. We initially propose to build tools to support classroom usage. The *Data Mining Solutions Center (DMSC) Portal effort* led by Dr. Rahul Ramachandran provides researchers both access to and the ability to use over one hundred different mining algorithms exposed as services. The researchers can use the portal to search for distributed data, move the distributed data to the computational resource, create and execute data mining workflows, share workflows or full experiments with each other, and finally publish science stories presenting their results backed by data and workflows. As part of this work, we will replace the existing DMSC customized workflow components with the OGCE tools for better sustainability, scaling and robustness, and for enabling Cloud Computing support.

In summary, the proposed work efforts will research specific problems that are aligned with EarthCube's generally desired workflow capabilities, including hierarchical and parametric sweep workflows for continuous data streams, user interaction with workflow execution, workflow sharing, and robust fault handling and exception models in distributed systems. This development will take place within an open community model through the Apache Software Foundation, which has a great potential value in leveraging established workflow languages, standards and tools from the business domain as it allows for sharing of workflow definition documents, using commercial and open source tools, and leveraging existing training, support, documentation and community activity. The workflows deployed for EarthCube will demonstrate the ability to integrate with legacy geoscience applications, provide flexibility and adaptability to experiments, track history and provenance, reuse and share workflows, hierarchical and parametric sweep workflow composition and execution, supporting long running applications and large number of concurrent executions.

## Team Members

Craig Mattocks (University of Miami), Renata Wentzcovitch (University of Minnesota), Rahul Ramachandran (University of Alabama, Huntsville), Suresh Marru (Indiana University), Chathura Herath (Indiana University), Marlon Pierce (Indiana University).