NASA Science Mission Directorate
Research Opportunities in Space and Earth Sciences – 2011
NNH11ZDA001N
A. 40 Computational Modeling Algorithms and Cyberinfrastructure
Abstracts of Selected Proposals

NASA's Science Mission Directorate, NASA Headquarters, Washington, DC, has selected proposals for the Computational Modeling Algorithms and Cyberinfrastructure program element in support of the Earth Science Division (ESD). This program element invests in technology developments to reduce the risk and cost of evolving NASA information systems to support future Earth system modeling efforts, including the integration of observational data into the model development, operations, and validation. This solicitation focuses on the computational technologies required to enable and facilitate the Modeling and Analysis Program and its supporting information systems.

This program element complements the modeling program with the following objectives: (1) To increase efficiency of the computational models can be gained from the advances in computational technology, including hardware, software, networks, and tools; (2) To build and enhance a common infrastructure which will dramatically reduce the time and energy spent by NASA scientists and engineers in the preparation of data for data-model intercomparison, model V&V, and model utilization; (3) To enable distributed model development and to provide a means to encourage open and rapid contribution, modern software engineering methodologies must be practiced to develop modular, agile, easily maintainable, and extensible code; and (4) To develop an experimental summer education program to supplement the higher education programs at research universities.

The ESD has selected 10 out of a total of 54 proposals received in response to this solicitation. The total funding for these investigations, over a period of two years for research and development projects and three years for an education project, is approximately $6.3 million.

Cecelia DeLuca/University of Colorado
CUPID: An IDE for Model Development and Modeler Training

We propose to work along two convergent paths: development of the Cupid IDE, which is a friendly environment for creating Earth System Modeling Framework (ESMF) components and coupled applications; and integration of the NASA GISS Model E with ESMF infrastructure. The project goals will be to work toward a GISS model with better parallel performance and modularity, and to enable that model to be used in a training and teaching environment in which Cupid is a part.
Cupid would support the creation of new compliant components, coupled system configuration, and model run and test. We will explore the integration of Cupid with other IDE efforts to assist with aspects of parallelism in code. The development team would initiate work with synthetic systems and incrementally incorporate components of

Model E, with the objective of enabling the full model to be manipulated within Cupid. During development of Cupid, the team will consult with ESMF developers, NASA GSFC, and partner organizations such as NOAA NCEP to incorporate and reconcile existing ESMF usability layers and conventions (e.g. Modeling Analysis and Prediction Layer-MAPL and National Unified Operational Prediction Capability-NUOPC Layer) in the Cupid design. The additional constraints that these impose for interoperability are likely to be valuable for providing structure to Cupid. The team will also interact with NASA Modeling Guru developers to investigate if or how such an IDE will interact with that system. The proposed work would be excellent preparation for modeling courses based on Model E, or for extension of these ideas to GEOS-5.

Seungwon Lee/Jet Propulsion Laboratory
Parallel Web-Service Climate Model Diagnostic Analyzer

The latest Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report stressed the need for the comprehensive and innovative evaluation of climate models with newly available global observations. The traditional approach to climate model evaluation, which is the comparison of a single parameter at a time, identifies symptomatic model biases and errors but fails to diagnose the model problems. The model diagnosis process requires physics-based multi-variable comparisons, which typically involve large-volume and heterogeneous datasets, and computationally demanding and data-intensive operations. We will develop a computationally efficient information system to enable the physics-based multi-variable model performance evaluations and diagnoses through the comprehensive and synergistic use of multiple observational data, reanalysis data, and model outputs. This proposal is a response to the NRA topic of computational model and analysis algorithms.

Satellite observations have been widely used in model-data intercomparisons and model evaluation studies. These studies normally involve the comparison of a single parameter at a time using a time and space average. For example, modeling cloud-related processes in global climate models requires cloud parameterizations that provide quantitative rules for expressing the location, frequency of occurrence, and intensity of the clouds in terms of multiple large-scale model-resolved parameters such as temperature, pressure, humidity, and wind. One can evaluate the performance of the cloud parameterization by comparing the cloud water content with satellite data and can identify symptomatic model biases or errors. However, in order to understand the cause of the biases and errors, one has to simultaneously investigate several parameters that are integrated in the cloud parameterization.

Such studies, aimed at a multi-parameter model diagnosis, require locating, understanding, and manipulating multi-source observation datasets, model outputs, and (re)analysis outputs that are physically distributed, massive in volume, heterogeneous in format, and provide little information on data quality and production legacy. Additionally, these studies involve various data preparation and processing steps that can easily become computationally very demanding since many datasets have to be combined

and processed simultaneously. It is notorious that scientists spend more than 60% of their research time on just preparing the dataset before it can be analyzed for their research.

To address these challenges, we will build Parallel Web-Service Climate Model Diagnostic Analyzer (PAWS-CMDA) that will enable a streamlined and structured preparation of multiple large-volume and heterogeneous datasets, and provide a computationally efficient approach to processing the datasets for model diagnosis. We will leverage the existing information technologies and scientific tools that we developed in our current NASA ROSES COUND, MAP, and AIST projects. We will utilize the open-source Web-service technology and Parallel Python to achieve the distributed parallel service-oriented system. We will make PAWS-CMDA complementary to other climate model analysis tools currently available to the research community (e.g. PCMDI's CDAT and NCAR's CCMVal) by focusing on the missing capabilities such as co-location, conditional sampling, and probability distribution function and cluster analysis of multiple-instrument datasets. Users will be able to choose between two ways to interface with PAWS-CMDA: (1) a web browser interface for quick and easy exploratory runs and (2) programming language interfaces (e.g. Matlab, Python, IDL) for heavy duty usage and batch runs.

Christian Mattmann/Jet Propulsion Laboratory
Next Generation Cyberinfrastructure to Support Comparison of Satellite Observations with Climate Models
We will deliver the next generation infrastructure that will closely interconnect NASA remote sensing observations with the DOE-funded Earth System Grid Federation (ESGF).

NASA's role in the Intergovernmental Panel on Climate Change (IPCC) is growing and has solidified in the form of the obs4MIPs project, whose goal is twofold: (1) the identification of NASA's key remote sensing observations that are readily comparable with the model output datasets part of the World Climate Research Program s (WCRP) Coupled Model Intercomparison Project (CMIP) and the upcoming IPCC 5th Assessment Report (AR5); and (2) the static publication of a handful of those datasets to the DOE-funded Earth System Grid Federation, the home for all IPCC generated AR5 data.

The existing obs4MIPs NASA data publication process brings with it limitations, that mainly involve the laborious process to prepare the NASA datasets for publication; and the fact that statically generating those datasets can cause those published to quickly become out of sync. Furthermore, once the NASA datasets have been published to the ESGF, climate researchers desire the capability to easily plug those datasets into their tools, and to perform model to observational dataset comparisons.

We will deliver a cyberinfrastructure that overcomes these challenges. The cyberinfrastructure will provide automatic conversion of NASA HDF-EOS/HDF datasets into CF/NetCDF datasets compatible with the ESGF; the ability to perform model checking on those converted datasets using the Climate Model Output Rewriter (CMOR-

2) checker; and the ability to automatically publish the converted datasets into the Earth System Grid Federation using its available tooling and infrastructure. In addition, we will deploy the cyberinfrastructure at three NASA DAACs, including PO.DAAC, ASDC and GES DISC. An existing and emerging model comparison tool, the Regional Climate Model Evaluation System (RCMES), will be plugged into the developed cyberinfrastructure, for use in CORDEX and other relevant IPCC comparisons. Finally, all software standards and processes built as part of our cyberinfrastructure will be delivered to the NASA HPC Modeling Guru system, and we will leverage that knowledge to create a community of interest for comparing NASA remote sensing data with model output.

Our collaboration includes the principals behind the ESGF, including one of its lead architects as well as the PI; the experts from NASA JPL who led the technical development and standards development of early efforts to connect NASA and the ESGF; and participants from three NASA DAACs, some of whom are already publishing data to the ESGF. We also include on our team experts in the development of open source software, which we will leverage and construct as part of our proposed effort, and which we will use as one tool promoting the sustainability of our effort.

Nikunj Oza/Ames Research Center
Integrating Parallel and Distributed Data Mining Algorithms into the NASA Earth Exchange (NEX)

There is an urgent need in global climate change science for efficient model and/or data analysis algorithms that can be deployed in distributed and parallel environments because of the proliferation of large and heterogeneous data sets. Members of our team from NASA Ames Research Center and the University of Minnesota have been developing new distributed data mining algorithms and developing distributed versions of algorithms originally developed to run on a single machine. We propose to integrate these algorithms together with the Terrestrial Observation and Prediction System (TOPS), an ecological nowcasting and forecasting system, on the NASA Earth Exchange (NEX). The resulting system will allow scientists to mine numerous Earth science data sources, model outputs, and nowcasts and forecasts that result from TOPS s seamless integration of these data sources and models, using distributed data mining algorithms, by leveraging NASA Ames s supercomputing assets. We will also develop a framework and interface under which data mining algorithm developers can make their algorithms available for use by scientists in our system, model developers can set up their models to run within our system and make their results available, and data source providers can make their data available, all with as little effort as possible. The resulting system will greatly reduce the now substantial time that scientists spend on data preparation, and will also reduce the effort of data mining algorithm developers, model developers, and data source providers to make their developments available to the Earth science community.

Rahul Ramachandran/University of Alabama Huntsville

Collaborative Workbench to Accelerate Science Algorithm Development

Motivation / Problem Statement. There are significant untapped resources for information and knowledge creation within the Earth science community in the form of data, algorithms, services, analysis workflows or scripts, and the related knowledge about these resources. Despite the huge growth in social networking and collaboration platforms, these resources often reside on an investigator's workstation or laboratory and are rarely shared. A major reason for this is that there are very few scientific collaboration platforms, and those that exist typically require the use of a new set of analysis tools and paradigms to leverage the shared infrastructure. As a result, adoption for science research is inhibited by the high cost to an individual scientist of switching from his or her own familiar environment and set of tools to a new environment and tool set.

Proposed Solution. The proposed Earth science Collaborative Workbench (CWB) will eliminate this barrier by augmenting a scientist's current research environment and tool set to allow him to easily share diverse data and algorithms. The CWB will leverage evolving technologies such as commodity computing and social networking to design an architecture for scalable collaboration that will support the emerging vision of an Earth Science Collaboratory.

Approach. The project team will implement the CWB on the robust, open source Eclipse framework, to be compatible with widely used scientific analysis tools such as IDL. The myScience Catalog built into CWB will capture and track metadata and provenance about data and algorithms for the researchers in a non-intrusive manner, with minimal overhead. Seamless interfaces to multiple Cloud services will support sharing algorithms, data, and analysis results, as well as access to storage and compute resources. A Community Catalog will track the use of shared science artifacts and manage collaborations among researchers. In order to test the effectiveness of this new collaboration environment, the team will assemble components into three prototype systems targeting accelerated science algorithm development, with increasing Cloud integration in each prototype. The Science investigators, with their colleagues among the widely distributed algorithm development teams, will evaluate analysis and collaboration capabilities of each prototype for NASA's upcoming Global Precipitation Monitoring (GPM) mission.

Significance. The proposed work innovates in three areas:
1) The project will implement a bottom-up rather than a top-down approach for building an Earth Science Collaboratory. Our bottom-up approach will incrementally advance the collaboratory concept and tool set by adding features into researchers' existing tool sets, to promote collaboration without requiring them to learn new tools.
2) The proposed work leverages emerging Cloud Computing technology to provide both storage and computing infrastructure. The CWB will integrate the use of Cloud Computing into mainstream scientific research.
3) Specific components to be developed for CWB currently do not exist elsewhere, and these components will build the foundation for a community collaboratory. The

myScience Catalog will provide a much-needed active metadata and provenance management capability to individual scientists. Cloud extensions will allow tools in the CWB to leverage any of several different Cloud services.

Relevance. The CWB and associated collaboration tools will result in prototypes of new capabilities for collaboration using state-of-the-art information technologies. This project addresses the Cyberinfrastructure and Technology to Enable Seamless Research Environments areas of CMAC solicited research as described for the CMAC program, by specifically addressing tools and technologies for data management and capabilities for innovative collaborative environments.

John Schnase/Goddard Space Flight Center
MERRA Analytic Services

We propose to build MERRA Analytic Services (MERRA/AS), a cyberinfrastructure resource for developing and evaluating a new generation of climate data analysis capabilities. MERRA/AS will support OBS4MIP activities by reducing the time spent in the preparation of Modern Era Retrospective-Analysis for Research and Applications (MERRA) data used in data-model intercomparison. It will also provide a testbed for experimental development of high-performance analytics. MERRA/AS will be a cloud-based service built around the Virtual Climate Data Server (vCDS) technology that is currently used by the NASA Center for Climate Simulation (NCCS) to deliver Intergovernmental Panel on Climate Change (IPCC) data to the Earth System Grid (ESG). Crucial to its effectiveness, MERRA/AS s servers will use a workflow-generated realizable object capability to perform analyses over the MERRA data using the MapReduce approach to parallel storage-based computation. The results produced by these operations will be stored by the vCDS, which will also be able to host code sets for those who wish to explore the use of MapReduce for more advanced analytics. While the work described here will focus on the MERRA collection, these technologies can be used to publish other reanalysis, observational, and ancillary OBS4MIP data to ESG and, importantly, offer an architectural approach to climate data services that can be generalized to applications and customers beyond the traditional climate research community.

Khawaja Shams/Jet Propulsion Laboratory
Cloud Enabled Scientific Collaborative Research Environment (CESCRE)

The Cloud Enabled Scientific Collaborative Research Environment (CESCRE) will significantly simplify and speed up the discovery and processing of science data by science collaborators. It eliminates the current serial, expensive, and cumbersome process and replaces it with a collaboration environment where each scientist can directly access the science data and/or results they are interested in. In addition to the sharing and reusing of data, it also provides an environment to share algorithms and ideas by taking

advantage of industry advances in virtualization and cloud computing. It fundamentally advances the way scientists share data, algorithms, and results.

NASA scientists lack a collaboration environment that streamlines data processing, configuration, and sharing raw or processed results. Productivity is hindered by software configuration, local infrastructure, data download, and inability to share data or algorithms. It is prohibitively expensive to correlate data across multiple instruments, as it requires domain expertise on the instrument, data types, and installation and execution of specific processing libraries. There are unnecessary delays in obtaining raw data or processed results because analysis requires one to:

a) Provision sufficient storage/compute capacity locally for raw and processed data
b) Configure algorithms and software along with the required libraries on their machines
c) Download raw data before processing
d) Rely on processing speed of the local infrastructure

Sharing algorithms via source code requires cumbersome software compilation as well as its dependencies. Sharing raw data and results is difficult, resulting in redundant processing across multiple scientists analyzing the same dataset.

We propose CESCRE to provide a seamless collaboration environment that streamlines data acquisition, processing, sharing of results as well as algorithms. Our approach fundamentally advances the collaboration and processing capabilities through advanced use of cloud computing.

Co-location of Storage and Compute Capacity: Cloud computing enables us to co-locate storage with virtually limitless computational capacity that can be provisioned on-demand. Co-location obviates unnecessary provisioning of local storage or downloading of data into the local infrastructure as computational capacity can be elastically provisioned in the cloud, where the stored data are available on the local network.

Preconfigured Machine Images: Sharing software through virtual machine images has advantages over the traditional binary or source distribution. We believe this will be the de-facto means of distribution in the future. CESCRE will provide means to share machine images amongst scientists and to catalog them. It will populate the catalog with images containing popular algorithms to streamline provisioning a machine and performing analysis.

Parallelization and Cloud Orchestration: We will integrate CESCRE with a distributed scheduling environment, Polyphony. We will utilize the elastic capacity in the cloud by orchestrating tasks across multiple machines concurrently. Polyphony has a track record for enabling MSL, MER, CARVE, and engineering applications to leverage cloud computing in a highly optimized environment.

Collaboration: Co-location of data enables sharing of raw data as well as higher-level data products. The pre-configured machine images will enable software developers and research scientists to share their implementations and algorithms with the community. Parallelization of algorithms in a shared environment will minimize duplicate processing and lead to collaborative processing of large datasets. It will enable efficient sharing of computational resources, data, and algorithms across the NASA community.

End-to-End Validation of CESCRE with InSAR: We will integrate CESCRE multiple SAR packages developed to process Level-0 raw data through to higher-level data products and modeling outputs.

Herman Shugart/University of Virginia
A Program for Computational Education and Internship Training for Environmental Science Students

The goal of this proposal is to develop a summer school in computer programming for motivated environmental science students. The summer school would teach basic software engineering with an object-oriented scripting language, followed by use of a compiled language. The students would then be instructed in high-performance and parallel computing for both distributed and shared-memory systems. At the end of this training program the students would be qualified to begin internships with NASA laboratories in computational fields.

Monish Tandale/Optimal Synthesis Inc.
Accelerating Earthquake Simulations on General-Purpose Graphics Processors

Numerical simulations play a crucial role in NASA s research in weather forecasting, global climate models and predictive models of complex interconnected solid-Earth processes. The increased resolution and complexity of the simulation models and their associated assimilation systems is driving the requirements for NASA s High-End Computing (HEC) resources. High Performance Computing (HPC) can no longer rely on the upward-spiraling processor clock speeds, and exploiting parallelism across increasing number of processor cores is the only alternative to extract additional computational performance. The central objective of the proposed research and development effort is to make the computational models and the associated analysis algorithms more efficient in a distributed parallel computing environment, and to demonstrate their operation on emerging general-purpose graphics processing units (GPGPU).
Under the research discipline of Solid Earth and Natural Hazards, NASA has supported the development of Virtual California (VC), which is a topologically realistic numerical simulation of earthquakes occurring on the fault systems of California. One of the demands placed upon simulations is the accurate reproduction of the observed earthquake statistics (Gutenberg-Richter and Omori statistics) over 3-4 decades. This requires the use of a finer-resolution fault model which greatly increases the computational requirements. Conventional, low-cost computer architectures cannot meet these requirements.
However, recent revolutionary advances in general purpose graphic processing units have the potential to address problems such as these at moderate cost increment. Under multiple research initiatives with NASA, OSI has demonstrated the application of these machines for tackling complex computational tasks. Using this background experience, Optimal Synthesis Inc. in collaboration with UC Davis, proposes to leverage the emerging computational power of GPUs for accelerating earthquake simulation using Virtual California.
The proposed R&D effort will analyze the functional decomposition of VC code and identify opportunities to exploit parallelism on the GPU architecture, in order to increase the run-time performance. Single-GPU, multi-GPU and GPU-cluster implementations of Virtual California will be developed to speed up run-times and to enable use of higher

resolution fault models. Finally the benefits in earthquake prediction made possible due to GPU implementation of VC will be demonstrated by performing VC simulations with a high-resolution fault model of Northern California.

Jia Zhang/Northern Illinois University

A Community-Driven Workflow Recommendations and Reuse Infrastructure

As current satellite measurements rapidly magnify the accumulation of more than 40 years of scientific knowledge, new discoveries increasingly require collaborative integration and adaptation of various data-driven software components (tools). In recent years, scientists have learned how to codify tools into reusable software modules that can be chained into multi-step executable workflows. However, although computing technologies continue to improve, adoption via the sharing and reuse of modules and workflows remains a big challenge.

This project aims to tackle this challenge from a novel angle, to study how to leverage peer scientists best practice to help facilitate the discovery and reuse of Earth science modules developed by others. The fundamental hypothesis is that published modules can be analyzed in a similar manner to actors in social networks, in that modules interact with modules much as friends interact with friends. Thus, social networking theory and analysis techniques can be employed to offer advice on module reuse. The research will tackle two fundamental research questions: What implicit knowledge, derived from social network analysis, may be extracted to help scientists better understand existing workflows and modules? and how can such implicit knowledge be used to aid module and workflow reuse?

To this end, the project will create an infrastructure for recommending and reusing workflows and workflow components for Earth science communities and integrate this infrastructure with the NASA Earth Exchange (NEX), a platform supporting scientific collaboration and knowledge sharing in the Earth sciences. The integration with NEX will consists of (1) building a foundational social network-empowered knowledge base model; (2) developing algorithms to derive interaction patterns, both direct and indirect, from this knowledge base; (3) developing a search and semi-automatic composition service that is responsive to the evolving working context and needs of the individual scientist; and (4) developing a prototyping system as a plug-in to the NEX workflow design and management system (VisTrails) that will aid scientists in reusing workflow modules and extending them to more complex science problems.

Intellectual Merit: The proposed work will leverage social network analysis to intelligently extract hidden information from a shared computing environment. By modeling Earth science workflow modules as social entities and their dependencies as social relationships, this research will open up new vistas for applying social science to facilitate software reuse and distributed workflow development. By building interlinked knowledge networks, the research promises to extend understanding of Earth science workflow interoperability and composition, and deepen our understanding of how collaborative expertise develops. New algorithms will be developed to create and analyze networks over the knowledge base to enhance leading module and workflow search approaches.

The Broader Impacts of this proposal are three-fold. First, the research goal of facilitating Earth science module and workflow reuse promises to build a sustainable infrastructural component for NEX, to offer a recommend-as-you-go service enabling Earth scientists to focus on science and allowing the community to make more effective use of computational resources. Second, the research may also enable significantly more virtual collaboration through module and workflow reuse. Third, the software and intellectual products of this research will be broadly disseminated through publications and open-source mechanisms.