NASA Science Mission Directorate
Research Opportunities in Space and Earth Sciences
NNH14ZDA001N-CMAC
A.40 CMAC 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 (CMAC) Program 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 information technologies and education activities required to enable and facilitate NASA's Earth system modeling efforts. The research and development themes specific to this solicitation included computing center based advanced data management and data analysis system, Earth system science researcher's workbench, strategy to restructure an earthquake model to distributed heterogeneous multicore computing architecture, and an annual interactive workshop and internship in specialized computational and software engineering techniques for Earth system modeling and data analysis. The ESD has selected 7 out of a total of 23 proposals received in response to this solicitation. The total funding for these investigations, over a period of three years, is approximately $4.3 million.

Ralph Dunlap/University of Colorado, Boulder
Cupid: An Integrated Development Environment for Earth System Models

IDEs are advanced software engineering tools that improve developer productivity and help to manage large codebases and complex software engineering processes involving a diverse set of tools. The work outlined in this proposal adds capabilities to and increases the robustness of the Cupid Integrated Development Environment (IDE), supported under ROSES/CMAC 2011, with a focus on adding new tooling to support development, execution, and runtime analysis of the Goddard Earth Observing System Model, Version 5 (GEOS-5). The work will result in a set of intelligent, integrated development tools that increase the overall usability of GEOS-5, simplify model execution and debugging, provide visualization of coupling control and data flows, and improve interoperability of GEOS-5 with other models based on the Earth System Modeling Framework. The work takes a critical step toward workflow consistency and code interoperability and opens the door to modeling infrastructure training across NASA centers.

Cupid extends the capabilities of the popular Eclipse IDE, adding customized tooling for geoscience models based on the Earth System Modeling Framework (ESMF) and the National Unified Operational Prediction Capability (NUOPC) software layer. The NUOPC software layer is designed to increase interoperability of ESM components and it is being implemented in the Community Earth System Model (CESM), Navy regional

and global coupled models, the NOAA Environmental Modeling System (NEMS), and Model E at NASA GISS. Cupid's existing tooling includes a static source code analysis engine and code generation facility that simplifies creation and modification of ESMF and NUOPC model components. Cupid's target user base includes model developers, especially those making architectural or infrastructure-level changes, such as introducing new model components or modifying the coupling configuration.

We propose to extend Cupid's existing tools, which are static in nature, with improved capabilities for executing and debugging ESMF applications. This includes the addition of control and data flow visualizations and improved compliance checking capabilities. The resulting tools will simplify the process of verifying correct model coupling implementations and will provide a more informative user interface for working with ESMF-based models. Additionally, we will optimize and improve Cupid's existing code analysis and code generation engine, and improve its technical documentation and overall robustness. GEOS-5 will serve as the primary test application for the proposed new features. At the same time, to promote continuity, we will continue an active partnership with the Model E team, addressing practical issues that hinder use of the IDE for end-to-end workflows.

In parallel with Cupid development, we will make improvements to GEOS-5 infrastructure to ensure its convergence with the latest developments of ESMF and the NUOPC interoperability layer. We propose to unify overlapping portions of the Modeling Analysis and Prediction Layer (MAPL) used in GEOS-5 with NUOPC. The resulting unified infrastructure will provide significant technical and scientific advantages to GEOS-5, such as making the latest ESMF features available to the model (e.g., unstructured grid remapping), improving GEOS-5's coupling interoperability with Model E, NCEP, GFDL, and Navy models, and reducing duplication of effort associated with maintaining multiple infrastructures.

Finally, to address the lack of availability of training materials about Earth system modeling infrastructure, we will team with Georgia Tech Professional Education (GTPE) to design and produce a set of on-line training modules about infrastructure for Earth system models and use of the Cupid IDE for model development. These materials will be made available for trainees and the general scientific community to use at a self-directed pace, and will complement traditional, face-to-face workshops.

Vipin Kumar/University of Minnesota
Scalable Analysis of Earth System Data Using Parallelized Graph-Based Approaches

Climate change is a defining environmental challenge facing our planet as rising temperatures and the transformation of global ecosystems are placing unprecedented strains on natural resources, man-made infrastructure, and our society. Responding to this challenge is difficult because there is a great deal of uncertainty in the projections of future climate. Key to addressing these issues is a deep, fundamental understanding of the global climate system and its complex physical processes. One of the primary tools used to study the properties of the climate system are general circulation models (GCMs).

Rapid gains in computational power and storage capacity have resulted in a data deluge for climate science in terms of observed and GCM simulated climate data. This proposal aims to address the need for better understanding of interactions within the climate system and improved model evaluation by developing a set of tools for analyzing graphs that are constructed as a representation of observed or simulated climate data and integrating those tools into the NASA Earth Exchange (NEX) for use and further refinement by the Earth science community. These graphs, also referred to as climate networks, have been proven to identify previously known and newly discovered teleconnections. At the same time, more general developments, such as those in complex network analysis and anomaly detection in graphs, show promise for applications to the analysis of changes in climate and the Earth system at a global scale.

However, most of these graph-based methods are computation-heavy and applying them to high resolution data sets lead to severe performance issues. Also, many of these existing algorithms are designed for single-core architectures and use mathematical primitives that require significant innovation before they can be deployed in a multi-core distributed computing environment. In this work, we propose to build a suite of graph algorithms that allows us to achieve the following goals:

- Enhance existing graph construction and network analysis algorithms that have been developed under the NSF Expeditions project and the NASA CMAC program to work on heterogeneous, high-resolution climate graphs with the goal of understanding whether various climate processes are incorporated and realistically represented in the current generation of GCMs, validating whether the responses of the natural system can be realistically assessed from the model output, and identifying differences between the projections made by different climate models or between different scenarios of a single climate model.
- Develop open-source software for these multi-core network analysis algorithms that compatible with the Apache Open Climate Workbench for easy integration.
- Integrate the new open-source complex networks analytics modules into the existing NEX data access and management infrastructure, to provide exploratory capabilities to climate scientists for further investigations into a wide range of complex Earth system processes and interactions at multiple spatial and temporal scales.

Data exploration with our approach will significantly advance the state of the art in Earth science research, because data sets can be compressed to their network structure to facilitate the validation and assessment of model data inter-comparison at minimal cost. Also, the resulting network provides a new information product and allows new types of relationship analyses. We expect that our research will play a crucial role for decision makers in assessing the vulnerability of socio-economic and natural systems to current and future changes in global and regional climate. Therefore, the proposed investigation is directly responsive to NASA's Plan for a Climate-Centric Architecture for Earth Observations and Applications from Space, specifically addressing climate continuity missions, modeling and computing, and climate monitoring and research, and product generation.

Kevin Schaefer/University of Colorado, Boulder
A Permafrost Benchmark System to Evaluate Permafrost Models

The Permafrost Benchmark System (PBS) will evaluate simulated permafrost dynamics against observed permafrost conditions. The project goals are 1) to develop a set of generic benchmarking tools capable of calculating performance statistics in multiple benchmarking efforts, and 2) develop benchmark datasets of permafrost dynamics based on available observations and 3) apply the PBS by evaluating multiple models. The PBS will provide an easily accessible, online tool to quickly evaluate model performance and guide model development without having to invest large resources into data preparation and organization. The chosen benchmark datasets include measurements of active layer thickness, permafrost temperature, and snow conditions. The ideal performance target is to match the observations within uncertainty, so the PBS benchmark datasets and evaluation metrics will account for observation uncertainty. The PBS will be a modular collection of tools built into the Community Surface Dynamics Modeling System (CSDMS) to leverage existing cyberinfrastructure and allow distribution of the benchmarking tools to the modeling community. The PBS fills a basic need of modeling teams to evaluate how well their models simulate permafrost dynamics, without a heavy investment in time and resources to organize the observations. This project will fill a gap in current benchmarking efforts, which currently do not include capabilities or observations to evaluate simulated permafrost dynamics.

John Schnase/Goddard Space Flight Center
Extending Climate Analytics-As-A-Service to the Earth System Grid Federation

We propose to build three extensions to prior-funded work on climate analytics-as-a-service that will benefit the Earth System Grid Federation (ESGF) as it addresses the Big Data challenges of future climate research: (1) We will build a cloud-based, high performance Virtual Real-Time Analytics Testbed supporting a select set of climate variables from six major reanalysis data sets. This near real-time capability will enable Cloudera Impala-based Structured Query Language (SQL) querying and SciHadoop-based MapReduce analytics over native NetCDF files and will provide a platform for community experimentation with emerging analytic technologies. (2) We will build a full-featured Reanalysis Ensemble Service comprising monthly means data from six reanalysis data sets. The service will provide a basic set of commonly used operations over the reanalysis collections. The operations will be made accessible through NASA's climate data analytics Web service and our client-side Climate Data Services (CDS) API. (3) We will build an Open Geospatial Consortium (OGC) WPS-compliant Web service interface to our climate data analytics service that will enable greater interoperability with next-generation ESGF capabilities. The CDS API will be extended to accommodate the new WPS Web service endpoints as well as ESGF's Web service endpoints. These activities address some of the most important technical challenges for server-side analytics and support the research community's requirements for improved interoperability and improved access to reanalysis data.

Herman Shugart/University of Virginia, Charlottesville

A Program for Education and Training in Computation for Environmental Sciences Students.

We propose to continue and to enhance our summer program for education and training in computation for students of environmental sciences. Our program, the Intensive Summer School in Computing for Environmental Sciences (ISSCENS) has already trained over 50 students in basic programming and the elements of software best practices over two sessions. After their training most students go on to summer internships where they apply their skills to projects in their fields. Training and experience in computing for students is essential for the future of scientific modeling; renewal of our program would enable us to expand upon our success with improved teaching methodologies and to place more students in NASA internships

---

Mark Simons/California Institute of Technology
High Performance Computing for Inclusion of Space Geodesy in Natural Hazard, Ground Water and Glacier Monitoring As Well as Disaster Assessment And Response

Space geodetic imaging from radar and optical satellites provide unique capabilities for monitoring deformation associated with earthquake faults, restless volcanic calderas, changing aquifers as well as glaciers, ice caps and ice sheets. We are now entering a golden age of geodetic quality space imaging enabled by investments from NASA and international space agencies to construct a constellation of radar and optical satellites including the ongoing US Landsat 8 mission and eventually the US/India NISAR radar satellite. These data can be used for monitoring surface movement, rapid post-disaster devastation assessment, and constraining physical models of underlying a variety of natural phenomena. This proposal will build on existing efforts to exploit high performance computing, in particular the use of large numbers of GPUs, to enable rapid production of massive quantities of geodetic imaging products, while at the same time furthering our efforts to allow for fully Bayesian exploration of models constrained by these data to improve our understanding of the fundamental underlying geophysical processes.

Thus, this effort focuses simultaneously on the creation of data products and on the rigorous inclusion of these data in physical models. From the data product perspective, we will build on a suite of existing tools developed for analysis of radar and optical geodetic imaging data that already exist, but currently are parallelized assuming a shared memory model on a single compute node using conventional CPU cores. Our goal is to explore the use of single and multiple GPUs in various steps of the processing chain (e.g., sub-pixel offset estimation, InSAR phase unwrapping, and InSAR time series analysis). From the modeling perspective, we have a Bayesian sampler that already takes advantage of 10's of GPUs simultaneously across multiple nodes to explore high dimensional (~1000) spaces for problems with relatively simple forward models. We propose to investigate approaches that remain scalable to machines with 1000's of GPUs to enable the next generation of physical models, as well as problems that may have smaller numbers of parameters but with more complex forward models (e.g. highly non-linear models).

Jia Zhang/Carnegie Mellon University
Climate Data Analytics Workflow Management

When we tried to integrate our previous CMAC projects to build a recommendation engine over climate data analytics web services, we came across an interesting cultural disagreement. Existing workflow tools, created by computer scientists, require Earth scientists to meticulously design their multi-step experiments before analyzing data. However, this is oftentimes contradictory to an Earth scientist's daily routine of conducting research and exploration. We hope to resolve this dispute. Imagine this: An Earth scientist starts her day applying JPL published climate data processing algorithms over ARGO deep ocean temperature and AMSRE sea surface temperature datasets. Throughout the day, she tunes the algorithm parameters to study various aspects of the data. Suddenly, she notices some interesting results. She then turns to a computer scientist and asks, "can you reproduce my results?" By tracking and reverse engineering her activities, the computer scientist creates a workflow. The Earth scientist can now rerun the workflow to validate her findings, modify the workflow to discover further variations, or publish the workflow to share the knowledge. In this way, we aim to revolutionize computer-supported Earth science.

We thus propose this project to realize the vision, aiming to pave a novel path to integrate our previous CMAC projects and create a sustainable building block toward Earth science big data analytics and knowledge sharing. Our plan is deliberated in four synergistic objectives. (1) We propose to study how Earth scientists conduct data analytics research in their daily work, develop a provenance model to record their activities, and to develop a technology to automatically generate workflows for scientists from the provenance. (2) We propose to develop Climate Model Diagnostic Analyzer (CMDA) 2.0 to enable multi-aspect physics-based and phenomenon-oriented model performance evaluations and diagnoses through the comprehensive and synergistic use of multiple observational data, reanalysis data, and model outputs. (3) We propose to build a data-centric provenance repository, and establish a PDSW (People, Data, Service, Workflow) knowledge network to support workflow recommendation. (4) We plan to develop mechanisms to ensure the scalability and performance of the expected CMDA 2.0 and recommendation system, leveraging the Apache OODT system technology.

Our proposed research well aligns with the long-term goals of the National Climate Modeling and Analytics Initiative. The community-approved, metrics-based performance evaluation web-service will allow a user to select a metric from the list of several community-approved metrics and to evaluate model performance using the metric as well as the reference dataset. This service will facilitate the use of reference datasets that are generated in support of the model-data intercomparison projects such as Obs4MIPs and Ana4MIPs. The on-demand user-customized climate dataset generation web-service will allow a user to customize and generate user-specific climatology datasets from Obs4MIPs project, Ana4MIPs project, and model outputs (e.g., CMIP5 now and CMIP6 in the future 2019-2020) available in Earth System Grid Federation. The interactive data visualization web-service will allow a user to interactively visualize data sets. Most of the plotting features will be dynamically adjusted on demand. The dynamic customization of

the output plots will facilitate comparisons with other plots and preparation for publications and presentations. The data-centric repository infrastructure will allow us to catch richer provenance to further facilitate knowledge sharing and scientific collaboration in the Earth science community. Our proposed research will provide a robust cyberinfrastructure to enable and facilitate transformative Earth science research. The expected software will be contributed to the Apache Open Climate Workbench under Apache License 2.0.