

Poster #21-1**From Data to Models and Analytics: An Integrated Scientific Computing Platform to Accelerate Environmental Research**

Aashish Chaudhary^{1*}, Charuleka Varadharajan², John Tourtellott¹, Shreyas Cholia², Matt Handerson², Doruk Ozturk¹

¹ Kitware Inc., Clifton Park, NY

² Lawrence Berkeley National Lab, National Laboratory, Berkeley, CA;

Contact: aashish.chaudhary@kitware.com

BER Program: SBIR

Project Website: <https://github.com/OpenDataAnalytics/gaia>

Advances in sensor-based field observations, multi-scale models, and computational technologies in Environmental Systems Science (ESS) provide new opportunities for applied research and development in academia and industry. These same advances are producing spatially and temporally heterogeneous data of growing size with different measurement frequencies and increasingly varying types and formats. As a result, current and foreseen model-data integration efforts require handling wide varieties and large volumes of data for pre-processing, simulation run, and post-processing which consequently requires the use of complex technologies such as data and resource management, high performance (HPC) and cloud computing. However, current tools and efforts for model-data integration in the ESS domain lack turnkey solutions for model-data integration workflows that involve large and complex data management, scalable pre and post-processing, and a complete suite of scientific visualization features. What makes it more challenging is that the current efforts are either fragmented or specialized for one-off systems. To reduce this fragmentation and to provide scalable model-data integration solutions, there is a need for an integrated software platform to advance research and development using simulation models. Furthermore, there is an urgent need to rethink traditional workflows which involves moving data between resources as needed for visualization and hypothesis formulation as the data is getting larger and computing costs (both HPC and cloud) are becoming cheaper.

To address the challenges described above Kitware and its collaborators are developing an integrated open source platform, Gaia, providing turnkey solutions for model-data integration using modern web technologies, advances in HPC and cloud computing and high-performance visualization. A key feature of our platform is that it minimizes and automates data movement between data storage and computing systems. Another critical feature of our design is that it enables users to perform high-fidelity post-processing by performing high-performance visualization server-side and in-situ on the compute cluster close to the data. Gaia will use Kitware developed VTK, ParaView, Girder, and community-led Jupyter effort in a loosely coupled modular design. In Phase I, we prototyped Gaia, enabling unified data access and server-side preprocessing operations. In Phase II, we will integrate simulation codes and post-processing capabilities while continue to enhance Phase I features to achieve a production-ready platform before the end of Phase II. We will also deploy Gaia at DOE HPC and cloud systems.