Open Hydrospheric Modeling Framework (OHMF) Roadmap

Prepared by the OHMF Concept Group

1. Purpose

To improve understanding of the complex behaviors of the various processes (e.g., physical, hydrological) and their interactions involved in the Earth System, as well as the accuracy and reliability of model predictions of weather, floods, droughts, and climate variability, researchers need to be able to make good use of the available data across disciplines to improve their theories, algorithms, models, and validations. However, a large amount of such valuable data often goes unused, due to the significant overhead of time and effort needed to discover, access, understand, and prepare the data. Similarly, there are many models available (e.g., hydrological/land surface models, routing models), but the complexity of these models necessitates a long lead time, even for a domain scientist, to learn how to use the models. Consequently, the strengths and limitations of the various models are not well evaluated or widely understood. For the user community, complexities related to individual models, different data requirements of models, and the myriad data formats, coordinate systems, and resolutions cause even more difficulties than those encountered by the research community. This combination of the variety and complexity of models and the usability of existing diverse data presents one of the most critical challenges in Earth Science. Hence, the development of an open modeling framework, which can integrate data and models easily and incrementally for knowledge discovery and management, is fundamentally important and urgent, not only to the research community and operational professionals, but also to policy makers and other users.

The ultimate goal of the OHMF is to build an open modeling framework, which should significantly reduce the time and effort on the part of users in the preparatory work for data and model comparisons, model testing and validations, and fundamental knowledge discoveries. In such a framework, components/modules interact via user-configured open interfaces, so that various hydrological models and data sources can be easily added and composed to interoperate, through scientific workflows.

2. Communication

One of the key aspects to the successful development of the OHMF is the enabling of communities of end users and researchers/builders to easily participate and contribute to the efforts of building the OHMF. To this end, we will carry out the following:

- Conduct webinars over the course of the project.
- Conduct community workshop to foster in-depth discussions and collect feedback.
- Work with other EarthCube Concept Groups (CGs), particularly the Earth Systems Modeling CG, to identify linkages.
- Invite members of other EarthCube groups to present related work at our community workshop.
- Participate in workshops and webinars of other EarthCube groups.

- Coordinate end-user evaluation and testing, e.g., by the NWS Ohio River Forecast Center(OHRFC) and the community at large. OHRFC is one of 13 River Forecast Centers (RFCs) of NWS and will serve as this project's operational EarthCube testbed. In the next phase of the OHMF, we will extend the OHMF testbed from OHRFC to other RFCs, as well as to other end-users (both operational and research).
- Leverage CUAHSI's community-engagement mechanisms to share findings of the OHMF Concept Group and gather community inputs and feedback.
- Link the Integrated Water Resources Science and Services (IWRSS) community with EarthCube community. Collect feedback from IWRSS on the project prototype. The NWS OHRFC, as the project's operational EarthCube testbed, will be integrated into the backbone of IWRSS.

During the EAGER phase of the OHMF development, team members will communicate through emails and regular teleconferences. Technical communication between the development team members, testers, and end-users will be effected over a track system hosted at the University of Pittsburgh. Over this track system, requirements, changes, bugs, defects, and questions will be reported, attended to, and statistically analyzed.

3. Challenges

- The design of OHMF should be as general as possible, so that technology updates will not outdate or disable the OHMF architecture. For example, although a specific workflow management tool will be adopted for the OHMF development, the overall OHMF design concepts and architecture do not rely on any specific workflow tool. They can be ported to another workflow system, whenever made advantageous by newer technology.
- The architecture needs to be robust and resilient enough to be straightforward for installation on different platforms. It should not require technologies that are unique to a specific platform.
- The design of OHMF also needs to be able to adapt to and incorporate evolving technologies in data access and management.
- Data discovery, access (coordination/communication with, and leveraging the work of, Data Access Services and Data Discovery Services Community Groups).
- Scalability (e.g., of the OHRFC) as more data sources and models are incorporated.
- Constraints imposed by network security and firewalls (e.g., NOAA/NWS Advanced Interactive Processing System (AWIPS) firewall) on the free-flow of data, particularly in the case of some operational end-users.
- Integration of OHMF solutions into existing frameworks and systems.
- End-users must be able to seamlessly integrate OHMF solutions into their existing workflows and research/operational paradigms.
- OHMF solutions must be extensible, allowing end-users to expand the capability of the OHMF solutions that was not initially anticipated.

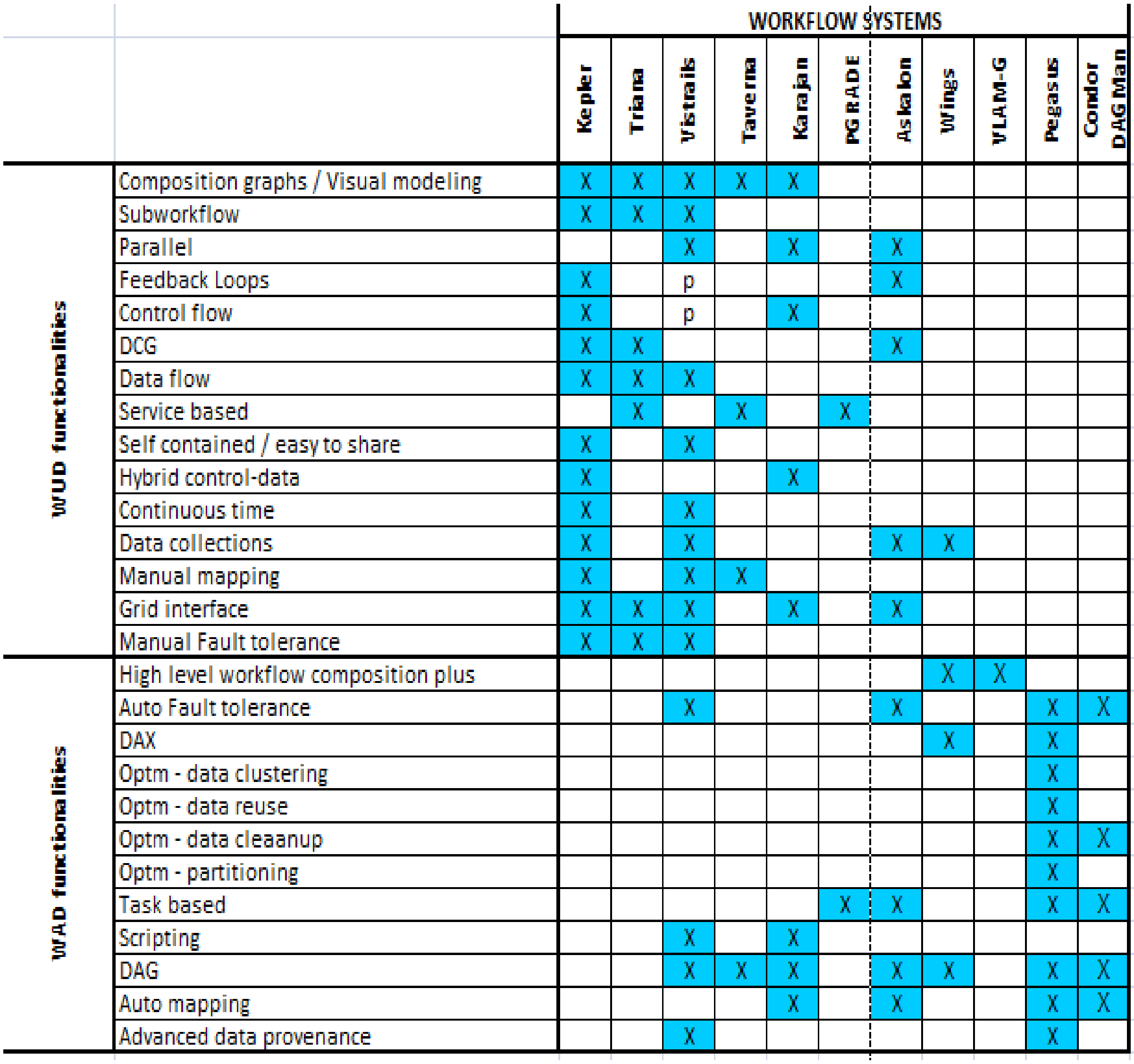
4. Requirements

- Openness/extendibility: Different models or modules can be easily added to and integrated in the OHMF through our proposed open modeling approach.
- Data independence: Scientific algorithms/models are separated from their associated disparate data sources (e.g., NASA satellite data, NOAA data, USGS data, wireless sensor network data), with different formats/organizations/coordinate systems, etc., through a well-defined layered-architecture.
- Flexibility: (a) Data and models operating on different spatial and temporal scales are easily exchangeable and/or fusible, based on our multi-scale data fusion approach and (b) model results can be easily analyzed and visualized online, and saved in different common formats.
- Usability: Well-defined web-based Graphical User Interface (GUI).
- Reliability and performance: Support cloud computing for OHMF system's reliability and performance.
- Models easily connected to OHMF: For OHMF prototype, we will focus on using hydrological/land surface models (e.g., NOAH, VIC).
- Inclusion of high level components: Open modeling framework, scientific workflow, component/module interfaces, layered architecture, multi-scale data fusion algorithms.
- Open to community participation: Architecture and design allow the community to freely contribute to and use the OHMF.
- Integration: OHMF solutions are easily integrated with other EarthCube components, through a standard strategy.
- Workflow instance prototype: Implement in OHMF prototype a workflow prototype that represents various scientific processes where data from NASA, NOAA, and USGS are used and transformed, models are run, model results are analyzed, and best estimates are obtained.
- Implement OHMF in a real-time operational environment, using NOAA/NWS Ohio River Forecast Center (RFC) as prototype testbed.
- Capability to handle requests for retrospective forecasts or model simulations for periods and locations with missing or unavailable data.

5. Status

Regarding workflows to be considered in the OHMF, we have reviewed the literature on applications of workflow systems over the past years and categorized the different systems and their functionalities (Fig. 1). All the applications use workflow systems for two major types of purposes: (1) to understand physical/environmental processes and (2) to optimize data/CPU-intensive modeling of natural processes.

For the first type, scientists use workflows to try different hypotheses and theories. They modify the solution procedures and typically run small-to-medium size datasets in order to quickly obtain intermediate results and to improve their hypotheses. For these users, defining the flow is more critical than actual implementation (e.g., tools that capture the desired analyses and/or simulations).



		WORKFLOW SYSTEMS										
		Kepler	Triana	Vistrails	Taverna	Karajan	PGRADE	Askalon	Wings	VLAM-G	Pegasus	Condor DAG Man
WUD functionalities	Composition graphs / Visual modeling	X	X	X	X	X						
	Subworkflow	X	X	X								
	Parallel			X		X		X				
	Feedback Loops	X		p				X				
	Control flow	X		p		X						
	DCG	X	X					X				
	Data flow	X	X	X								
	Service based		X		X		X					
	Self contained / easy to share	X		X								
	Hybrid control-data	X				X						
	Continuous time	X		X								
	Data collections	X		X				X	X			
	Manual mapping	X		X	X							
	Grid interface	X	X	X		X		X				
	Manual Fault tolerance	X	X	X								
WAD functionalities	High level workflow composition plus								X	X		
	Auto Fault tolerance			X				X			X	X
	DAX								X		X	
	Optm - data clustering										X	
	Optm - data reuse										X	
	Optm - data cleaanup										X	X
	Optm - partitioning										X	
	Task based						X	X			X	X
	Scripting			X		X						
	DAG			X	X	X		X	X		X	X
	Auto mapping					X		X			X	X
	Advanced data provenance			X							X	

Figure 1.WUD and WAD functionalities and their associated workflow systems.

For the second type, with flow definitions already conceptually tested, users run the selected workflow on large datasets to intensively test their hypotheses. Such applications require efficient uses of storage and CPU. For these users, actual implementation is more critical than defining the flow. In some applications, the workflow definition is considered to be just another constraint, along with all the other constraints (e.g., those on resources, time, security, and storage), to be used in finding the most efficient execution.

In Figure 1, which summarizes major functions of the various existing workflow systems, we use WUD to represent "workflow-under-definition,"(mainly used for Type 1 applications) and WAD to represent "workflow-already-defined," (mainly used for Type 2 applications).

For WUD, it is important to provide users with tools that help materialize and conceptualize ideas; while for WAD, the priority is to provide algorithms to distribute the CPU and storage demands on resources (e.g., to schedule algorithms). In some cases, existing workflow engines (WES) satisfy only WUD or WAD but not both. Literature has shown a lack of workflow systems that can provide both easy tools for WUD and efficient execution tools for WAD. It has been suggested that a home-made workflow engine (MOTEUR) satisfies some of the features typically found useful for both WUD and WAD applications.

6. Solutions

- Develop an open data and open modeling system for OHMF by (1) developing data source agents to automatically access heterogeneous data from various data sources, (2) opening up interfaces between model components, (3) using layered architecture, and (4) combining with a workflow system.
- Develop a testbed for evaluating and testing the concepts and ideas of the OHMF through a prototype system.

7. Process

The development process of OHMF will be based on the procedure of the Unified Process (UP). That is, our development will be realized through cycles in a way that the most risky and technical challenging issues will be looked into and addressed first. Each cycle will result in a generation of a product release. Each release will provide some useful functions, not just architectural components. In this way, we can ensure that our product is usable from the very beginning. This also ensures that the functionalities developed can be tested and become more mature from one release to the next. Also, the technical non-functional risks encountered over the course of development will be solved as soon as possible.

We also plan to use the CMM (Capability Maturity Model) model to compare our product to a standardized method in terms of repeatability, process definition, management and finally, optimization that will improve the system gradually by its interaction with the community.

8. Timeline

Tasks for the 1st year:

- Develop OHMF prototype based on our open data open model framework, including data source agents, open model/component interfaces, and testing and evaluation of their effectiveness and ease of use for heterogeneous data sources and models. The layered architecture framework and the workflow system will also be developed and initially tested.
- Develop OHMF testbed with OHRFC.
- Conduct community outreach activities and interact with other EC groups.
- Develop prototype use case (OHRFC).

Tasks for the 2nd year:

- Extend developed OHMF prototype to other NWS RFCs.
- Continue testing and evaluating developed OHMF prototype.
- Begin integration of OHMF with other EC groups (Data Access Services and Data Discovery Services Community Groups; Interop and Earth System Modeling Concept Group, Workflow Community Group).

Tasks for the 3rd year:

- Test OHMF performance for data-intensive systems (e.g., parallelism, DISC, Cassandra) and incorporate cloud computing technology.
- Expand OHMF to sophisticated non-RFC end-users, such as USGS, USACE, US Bureau of Reclamation, researchers from universities/institutions and national labs.
- Identify other EC groups' systems that may be ready for integration with OHMF and formulate an integration plan.

Tasks for the 4th year:

- Expand OHMF to "less-sophisticated" end-users, such as NWS WFOs, possibly emergency managers, as well as general users.
- Continue to integrate OHMF into EC.

Tasks for the 5th year:

- Fully integrate OHMF into EC.
- Provide clear metrics of the utility of the OHMF-EarthCube solution.

9. Management

The core team members of the OHMF project will collaboratively manage the project to ensure it achieves its overall goals, while individual team members will be responsible for the tasks within her/his knowledge domains. The outreach activities to the different user and research communities will be mainly overseen by the project's partners. Interactions with other EC groups will be conducted through collaborative efforts between the core team members and the project partners. The project PI will be responsible for the coordination among team members, adherence to schedules, progress reports and other documentation, and project reviews and demonstrations.

10. Risks

- Integration between EC groups: As in commercial software projects, integration of EC concept groups' work and/or systems poses certain risks. The various groups need to work together to effect an automatic and open communication protocol among different EC components.
- Technology changes: The evolution of technology is more a certainty than a risk. Therefore, our system needs to be strategically designed to always be able to adapt to changes (e.g., keep loose couplings and choose tools that are also committed to loose couplings).
- AWIPS2: Our first end-user (NWS) is undergoing major updates and modifications to its technological forecast system and platform. Therefore, we need to keep such system updates on track in order to quickly adapt to any upcoming problems and changes.
- User working environment integration: OHMF needs to be seamlessly integrated into current environments/workflows of end-users to enhance their capability.
- Adoption resistance to OHMF by user communities (related to trust in security, privacy, data quality, etc.).
- Possible incompatibilities between open infrastructure of OHMF and requirements of specific users (e.g., OHRFC).