Collaborative Research: CyberWater—An open and sustainable framework for diverse data and model integration with provenance and access to HPC

Award #s: 1835785, 1835817, 1835592, 1835338, 1835602, 1835656

NSF CSSI PI meeting, Seattle, WA, February 13-14th, 2020

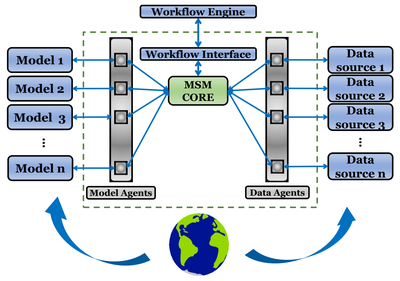
Xu Liang[1], Yao Liang[2], Daniel Luna[1], Ranran Chen[2], Yuan Cao[2], Yuankun Fu[2], Sudhakar Pamidighantam[3], Fengguang Song[2], Jerad Bales[4], Anthony Castronova[4], Ibrahim Demir[5], Richard Hooper[6], Witold Krajewski[5], Lan Lin[7], Ricardo Mantilla[5], Yang Zhang[8]

[1]University of Pittsburgh, [2]Indiana University-Purdue University Indianapolis, [3]Indiana University, [4] CUAHSI, [5]University of Iowa, [6] Tufts University, [7] Ball State University, [8]North Carolina State University

1. What is CyberWater

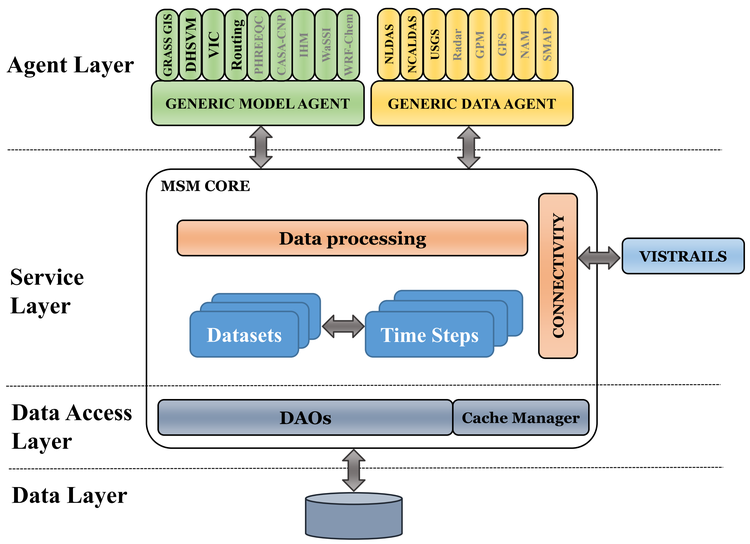
CyberWater is an open and sustainable modeling framework software system which enables not only an easy and incremental integration of diverse data sources and models for knowledge discovery and interdisciplinary team-work, but also reproducible computing and the seamless and on-demand access to various HPC resources.

CyberWater utilizes the visual interface of VisTrails (Bavoil et al., 2005), a scientific workflow management system that offers a nice and convenient graphical workflow interface, in addition to its unique capabilities, such as provenance and data visualization. CyberWater allows executing complex workflows with comprehensive interactions between remotely accessible heterogeneous data sources and diverse models and model coupling capabilities.



2. Motivation

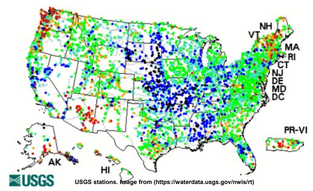
The challenges of accessing heterogeneous data sources and integrating diverse models are longstanding, yet there are few existing systems which can address such cyberinfrastructure challenges. Researchers need to make an effective use of various available data across disciplines to improve theories, algorithms, accuracy and reliability of the models' predictions and simulations, understanding of the complex behaviors of the various processes (e.g., physical, hydrological, atmospheric, biological, chemical) and the interactions among them. However, a large amount of such valuable data often goes unused. On the other hand, the complexity of the existing models makes them difficulty to be shared and used by others. Although a number of modeling systems exist for scientific communities, most of them do not support direct access to heterogeneous data sources over the Internet. In addition, significant challenges exist in these systems in connecting data and models in an effective and easy-to-use fashion when multiple steps are involved, such as data acquisition, pre-processing, fusion, model layout, simulations, forecasts, etc. Researchers find a high cost in learning and implementing these tools for creating, configuring, and coupling their models. Consequently, the various valuable data collected are not used and the strengths and limitations of the various developed models are not well evaluated, understood, and widely used. To this end, we develop CyberWater to expedite the process of fundamental scientific explorations/discoveries and significantly reduce time and effort for data and model integration applications.



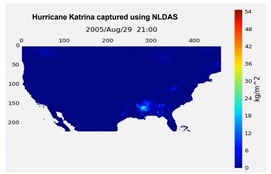
3. Current Status

CyberWater is currently capable of accessing four different Data products, through four Data Agents:

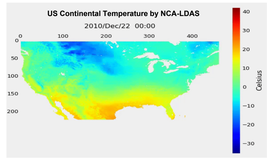
1. **USGS Data Agent:** Provide access to more than 18,000 different variables measured by the United States Geological Survey's (USGS) station infrastructure.



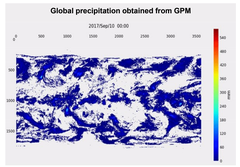
2. **NLDAS Data Agent:** Be capable of retrieving 10 different hydro-meteorological variables from the North American Land Data Assimilation System (NLDAS).



3. **NCA-LDAS Data Agent:** Similar to NLDAS, it can retrieve 19 different hydro-meteorological variables from the National Climate Assessment - Land Data Assimilation System (NCA-LDAS).



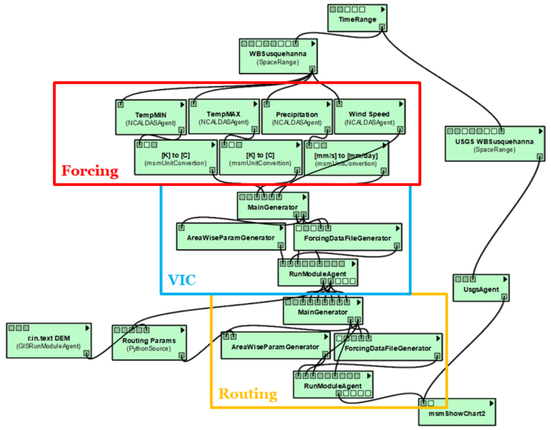
4. **GPM Data Agent:** Retrieve precipitation data from the Global Precipitation Measurement (GPM), a NASA satellite mission that measures and monitors precipitation around the world.



CyberWater currently also offers three different Model Agents:

1. **VIC Model:** The Variable Infiltration Capacity model is a macro-scale hydrologic model, capable of simulating water and energy fluxes in a distributed fashion (Liang et al., 1994).
2. **Routing Model:** The Routing model takes time-series of surface and subsurface runoff which can be from either a data source or model outputs (e.g., the VIC model outputs) and computes overland and channel routings for a given watershed (Hernández and Liang, 2018).
3. **DHSVM:** The Distributed Hydrology Soil Vegetation Model (DHSVM) (Wigmosta et al., 1994) is a high-resolution model that simulates the effects of topography, soil, vegetation, and weather on hydrologic processes within watersheds.

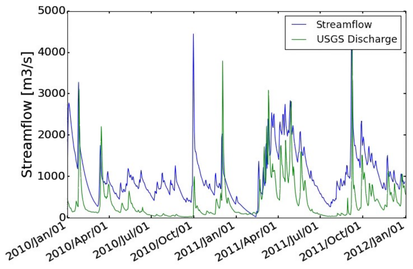
CyberWater includes a set of tools called the Generic Model Agent Tools. They are created to allow the user to build his/her own Model Agents and integrate his/her very own model(s) into CyberWater with little to no coding required at all. At present, these tools allow us to develop the Model Agents for VIC, DHSVM and the Routing Model without coding and obtain the same results as those using the specific Model Agents we wrote.



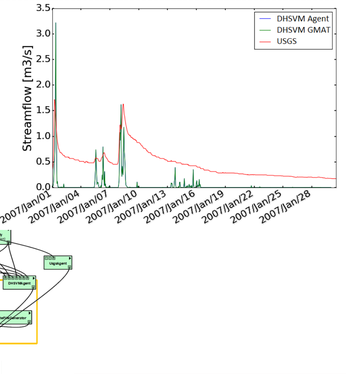
Example of using the Generic Tools, generating the VIC Agent and the Routing Agent, to couple VIC and the routing models in a workflow

4. Example Workflows

Here we illustrate how to use CyberWater to build a full simulation, using the West-Branch Susquehanna river basin in PA, as an example. This is a large basin, with an area of more than 17,700 km². This simulation will require the use of the water-balance-mode of the VIC (v4.0) Model, coupled with the Routing Model. The uncalibrated **streamflows simulated from the coupled models are compared** with the USGS observed streamflows at the USGS station 01553500.



The following example depicts the usage of th Generic Model Agent Tools to generate the Mod Agent for the DHSVM model to perform a simulatio of the Indiantown Run streamflow, a small basin 14 km² in PA. The results are compared with thos using the Model Agent specifically written for th DHSVM and the streamflow measurements fro USGS.



5. Future Work

CyberWater is currently going through meticulous testing to ensure the desirable software quality. We are also working on integrating new Data Agents such as:

1. The **North American Mesoscale Forecast System (NAM)**, high-resolution model forecast information of the US generated from an NOAA model.
2. The **Global Forecast System (GFS)**, global low-resolution forecasts of more than 700 atmospheric and land-soil variables generated from an NOAA model.
3. The Precipitation measurements with **Doppler Radar** made by the NOAA's National Weather Service (NWS).

Additionally, CyberWater will also incorporate new Model Agents, allowing the users to use different models in geosciences, such as:

1. The **PH-Redox Equilibrium model in C (PHREEQC):** an environmental model that simulates chemical reactions in an aqueous medium.
2. The **Carnegie-Ames Stanford Approach including Carbon-Nitrogen-Phosphorous cycles (CASA-CNP):** a biogeochemical model.
3. The **Water Supply Stress Index (WaSSI)**, a **hydrologically-based ecosystem model** to simulate Water Stress.
4. The **Weather Research and Forecasting model coupled with Chemistry** (WRF-Chem).

6. References

- **Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T., Vo, H.T., 2005. VisTrails: Enabling interactive multiple-view visualizations. Proc. IEEE Vis. Conf. 18. https://doi.org/10.1109/VISUAL.2005.1532788.**
- **Hernandez, F. and Liang, X., 2018. Hybridizing Bayesian and variational data assimilation for high-resolution hydrologic forecasting. Hydrology and Earth System Sciences. V.22. 11, 5759--5779. https://www.hydrol-earth-syst-sci.net/22/5759/2018/. DOI: 10.5194/hess-22-5759-2018.**
- **Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. Atmos. 99, 14415–14428. https://doi.org/10.1029/94JD00483.**
- **Wigmosta, M.S., Vail, L.W., Lettenmaier, D.P., 1994. A distributed hydrology-vegetation model for complex terrain. Water Resour. Res. 30. https://doi.org/10.1029/94WR00436**