Agile data management and synthesis for heterogeneous, multiscale watershed datasets

Charuleka Varadharajan, Boris Faybishenko, Roelof Versteeg and Deb Agarwal

The goal of the data management and assimilation (DMA) component of the U.S. DOE's Genomes to Watershed and Watershed Function Science Focus Areas (SFA) is support of SFA data management needs and efficient generation of data synthesis products to enable SFA science and render datasets available for community use. The objective is to create a software platform to enable users to integrate diverse and disparate field, laboratory, and simulation datasets, including geological, geochemical, geophysical, microbiological, hydrological, and meteorological data across a range of spatial and temporal scales within the Rifle floodplain and the East River watershed, Colorado.

Development priorities of data products are dynamically driven by project science needs. Agile development and modified User-Centered Design (UCD) techniques have and will continue to be used to develop synthesis products, and common infrastructure to archive SFA-generated data, obtain external datasets, perform quality assurance and quality control (QA/QC), perform gap-filling, implement synthesis pipelines, provide user analytics interfaces, support parameter estimation, implement uncertainty quantification, and manage data.

In the Genomes to Watershed SFA, data collection and QA/QC activities have focused on data and products needed for on-going scientific analyses, and hydrological and geochemical modeling. Sensor-based datasets from the DOE-LM installed System Operation and Analysis at Remote Sites (SOARS) network that collects data from water-level pressure transducers, vadose zone and groundwater thermistors, and meteorological stations, as well as data from laboratory characterization of groundwater samples have been curated and archived in a database. An example of an accomplishment is development of automated QA/QC methods to identify and flag issues in the data. Based on the analysis of field water-level data, we provided recommendations on reinstallation and calibration of pressure transducers installed in monitoring wells. Additionally, in support of QC of the geochemical dataset, we developed an approach of flagging samples based on an evaluation of the ionic balance of water samples. Additionally, we have developed, tested and released a broker and portal that integrates the diverse datasets from the different distributed databases where project data are curated. The development of the portal user interface followed UCD methods, resulting in an intuitive product that presents the highest priority datasets and capabilities needed by users including data search, visualization and download.

Our agile approach has enabled us to build a DMA system that is keeping pace with the project science needs while utilizing limited resources.