Toward watershed simulations combining machine learning and high-resolution process-based models: Initial results from the ExaSheds project

Scott L. Painter,[1] Ethan T. Coon,[1] Dan Lu,[2] Shih-Chieh Kao,[1] Goutam Konapala,[1] Julien Loiseau[3], Irina P. Demeshko[3], J. David Moulton[4], and Carl I. Steefel[5]

[1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830
[2]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN
[3]Computer and Computational Sciences Division, Los Alamos National Laboratory, Los Alamos, NM
[4]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM
[5]Earth and Environmental Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA

Contact: (paintersl@ornl.gov)

Project Lead Principle Investigator (PI): Carl I. Steefel (LBNL)

BER Program: Other (Data Management)

Project: ExaSheds

Project Abstract:

Sustainable management of watershed systems depends on quantitative modeling capability for the hydrologic and biogeochemical processes that control watershed system dynamics and water availability and quality. The ExaSheds project is pursuing a new watershed predictive capability that combines data-driven machine learning (ML) approaches and process-based (PB) hydrobiogeochemical simulation while taking advantage of leadership-class supercomputers. One component of that long-term vision is a PB-ML hybrid capability that uses the output of high-resolution process-based simulations as one of the inputs to ML models.

We are testing a hybrid approach using multi-decadal stream discharge data from the conterminous US-scale CAMELS dataset and short records of stream discharge from the East River, Colorado, watershed. Long Short-Term Memory (LSTM) networks were used for the ML approach and well-established semidistributed hydrology models were used for the process-based models. The hybrid approach outperforms both the LSTM and the PB approaches. For the East River catchments, the hybrid model was successfully trained using only two years of training data. These results suggest that process-based simulations can improve the robustness of ML models when inputs are out of sample and allow them to be trained with shorter records.

We used semi-distributed models as the process-based models in this initial test, but the long-term vision is to replace those with high-resolution simulations running on leadership class computers. To that end, we are also refactoring the integrated watershed simulator ATS to run on heterogeneous computing architectures that combine CPUs and GPU coprocessors. ATS was refactored using the Kokkos library, which abstracts data and execution models from the process representations, thus allowing the same code to be mapped onto a variety of computer architectures including leadership class supercomputers. Three-dimensional solutions of Richards equation on the Summit supercomputer demonstrate the feasibility of the approach.