Title: ExaSheds: Advancing Watershed System Science using Machine Learning and Data-Intensive Extreme-Scale Simulation

Carl Steefel[1]*, Scott Painter[2], Xingyuan Chen[3], Ethan Coon[2], Dipankar Dwivedi[1], David Moulton[4], Alison Appling[5], Alexander Sun[6]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]Pacific Northwest National Laboratory, Richland, WA; [4]Los Alamos National Laboratory, Los Alamos, NM; [5]U.S. Geological Survey; [6]University of Texas, Austin

Contact: (CISteefel@lbl.gov)

Project Lead Principle Investigators (PI): Steefel and Painter

BER Program: Data Management

Project: ExaSheds

Project Website: https://exasheds.org

Project Abstract: The ExaSheds project is exploring synergies between data-driven machine learning (ML) approaches and process-based hydro-biogeochemical high-performance computing (HPC) simulations with the objective of improving predictive capability for watershed and river basin function. The new set of hybrid ML+HPC and ML tools are being developed to enable scientific discovery of both present-day watershed function, and also to extend predictive capability to future climate scenarios where ML-only approaches may be challenged.

The project is pursuing several research themes and applying the resulting tools in integrated demonstrations studies of watershed to basin-scale function in two study areas: the Upper Colorado Water Resources Region (UCWRR) and the Delaware River Basin (DRB).

The project is using ML to generate difficult-to-observe model inputs from sparse, coarse, and indirectly related observations. We developed an approach to sequential imputation of missing spatio-temporal precipitation data (Mital et al, 2020) and modeling the distribution of snow-water equivalent by combining precipitation and temperature with LIDAR maps (Mital et al, submitted).

The use of ML+HPC hybrid models is being explored as an alternative to traditional inverse modeling. As a preliminary step, we have developed a deep neural network (DNN)-based inverse modeling method that estimates the subsurface permeability of a watershed from stream discharge data (Cromwell et al, 2021).

We are exploring the use of hybrid ML-physics models for hydrology. We developed a hybrid streamflow model that use output of a hydrological model as one of the inputs to a Long Short-Term Memory (LSTM) network. That hybrid approach outperforms standalone LSTM on diverse catchments (Konapala et al, 2020) and when trained on short datasets (Lu et al, 2021).

We are developing a new set of reduced dimension and reduced order models for biogeochemistry. Our initial effort has focused on developing an approach for representing hillslope fluxes using an ensemble of flowtubes that is applicable to transient, variably saturated conditions.

A simulation (hydrology and geochemistry) capability on heterogeneous computer (GPU+CPU) architectures is being developed, using the DOE software Amanzi-ATS and Crunch. This is

providing vastly improved simulation throughput capacity to support model-data integration and enabling large-scale simulations on leadership class computing facilities.

References

Konapala et al, 2020. *US. Environmental Research Letters*, 15(10), p.104022.
Mital et al, 2020. *Frontiers in Water 2: 20. doi: 10.3389/frwa*.
Cromwell et al, 2021. *Frontiers in Earth Science* 9:613011. doi: 10.3389/feart.2021.613011
Lu et al, 2021. *Journal of Hydrometeorology*, *22*(6), 1421-1438.