Title: **Overview of the ExaSheds Project**

Carl Steefel[1]*, Scott Painter[2], Dave Moulton[3], Xingyuan Chen[4], Dipankar Dwivedi[1],Ethan Coon[2], Utkarsh Mital[1], J.E. Brown[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA
[2]Oak Ridge National Laboratory, Oak Ridge, TN
[3]Los Alamos National Laboratory, Los Alamos, NM
[4]Pacific Northwest National Laboratory, Richland, WA

Contact: (CISteefel@lbl.gov)

Project Lead Principle Investigator (PI): Steefel

BER Program: Other

Project: ExaSheds

Project Abstract:

Advances in machine learning capability, data quality and quantity, and computational capabilities create a unique opportunity to dramatically improve our watershed system modeling capability by making better use of diverse and spatially extensive data available from remote sensing, future networks of sensors, detailed site-level investigations, and highly resolved simulations. We are developing and testing a new watershed predictive capability that uses data-intensive machine learning in combination with BER's unique hydro-biogeochemical simulation capability, while leveraging leadership-class computational resources. The initial focus of the project is on the Upper Colorado River, which includes the East River Watershed under study in the Berkeley Lab SFA.

The ExaSheds seed project includes research tasks that are prototyping and evaluating strategies to address key aspects of data-driven ML applied to watershed modeling and tasks that are adapting BER's watershed computational tools to heterogeneous computer architectures. In this poster, we summarize the overall vision for the ExaSheds project and highlight some initial results. Additional results can be found in an accompanying poster led by S. Painter.

Task 1 has focused on demonstrating and evaluating the use of ML to inform watershed model inputs. For this purpose, we conducted this study at the Upper Colorado Water Resource Region (UCWRR) and used precipitation data, which is a critical forcing input for watershed models. High-resolution precipitation (<100 m) inputs are needed for accurately predicting watershed-scale hydrological and biogeochemical dynamics, but precipitation data are currently available at much coarser resolutions from PRISM (800 m; daily) and NLDAS (~12.5 km; hourly). To convert coarse-resolution precipitation data to the model resolution, referred to as "downscaling", point scale measurements are needed. To impute missing data, we used the Random Forest to learn correlations with stations that have complete records. Subsequently, we used Random Forests and Convolutional Neural Networks to generate high-resolution (~100 m) precipitation data by incorporating the effects of various factors such as elevation, vegetation, and land-use. We also identified the relative importance of these factors and examined weather station data both near to and far from the area of interest to investigate the multi-scale spatial patterns of precipitation. Results indicate that the proposed strategy generates high-resolution precipitation data by learning from lower-resolution precipitation coupled with high-resolution features with reasonable accuracy.