On Demand Machine Learning for Multi-Fidelity Biogeochemistry in River Basins Impacted by Climate Extremes

Carl Steefel[1], Dipankar Dwivedi[1], Guillem Sole-Mari[1], Zexuan Xu[1], Ilhan, Ozgen[1], Allan Leal[2], Utkarsh Mital[1]

1. Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory
2. ETH Zürich, Switzerland

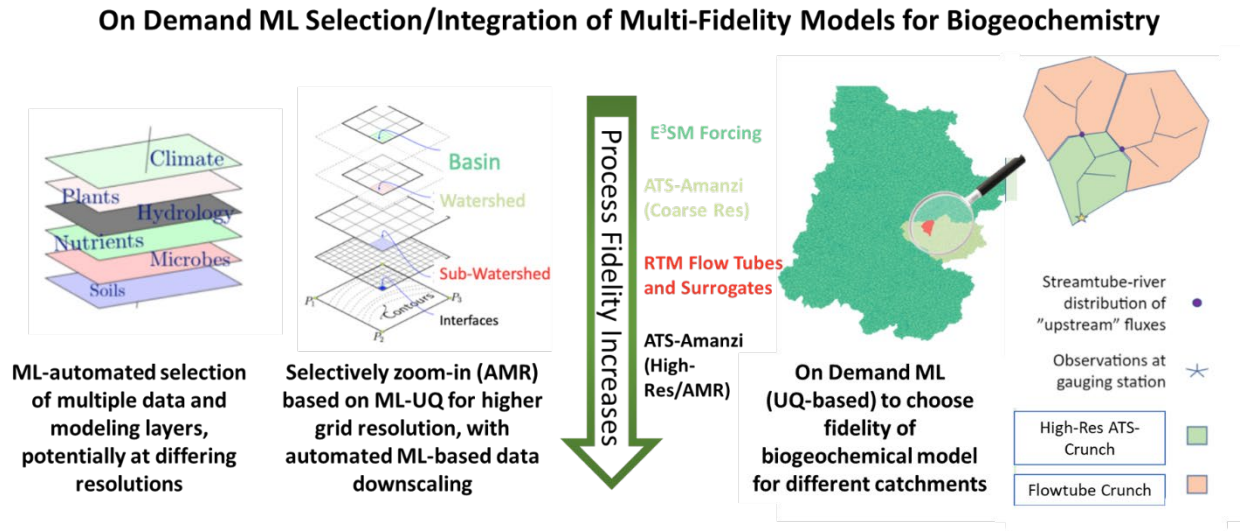
Focal Area 2: Predictive modeling of watershed to river basin scale biogeochemistry through the use of AI techniques and AI-derived model components and the use of AI to design a prediction system comprised of a hierarchy of multi-fidelity models, including AI driven model/component/parameterization selection.

Science Challenge: A full treatment of water flow and biogeochemical reactive transport at watershed to river basin scales is not currently possible, and yet there is a critical need to provide improved estimates of fluxes of nutrients and contaminants for both scientific and water management purposes at these scales. The use of hybrid multi-fidelity predictive models that take advantage of ML techniques offers an attractive option to overcome the obstacles associated with computational expense, especially insofar as it is possible to maintain process fidelity for heterogeneously distributed biogeochemical processes interacting with the hydrological cycle—a situation that is expected to be particularly pronounced during climate extremes.

Rationale: Watersheds funnel rain and snowmelt to river basins where they are used by municipalities, agriculture, and energy producers, but they may also be sources of contamination that affect water quality in the river system. These effects may be more pronounced during episodes of climate extreme. For example, flooding can release large amounts of nutrients or contaminants from dispersed sources, while drought can reduce the dilution expected for a typical river basin. Under these extreme climate conditions, the biogeochemical functions of watersheds and river basins are poorly understood, and far from predictable with scientifically-defensible consideration of mechanistic coupled processes.

Narrative: The need for dramatically improved prediction of river basin scale biogeochemical function is dramatic, but the computational challenges are daunting, But machine learning (ML) can play an important role in at least five ways: 1) ML can facilitate the inclusion of diverse big data in physics-based models for water and biogeochemistry through downscaling and upscaling approaches (Mital et al. 2020; 2021 submitted), 2) ML can achieve improved predictability by enabling calibration and validation of models for given river basins and watersheds (Cromwell et al 2021), 3) ML can enable the development of surrogate and reduced order/dimension models that capture watershed and river basin function with reduced computational expense, 4) On The Fly ML can be used for automation of uncertainty quantification (UQ) to choose dynamically the level of fidelity and computational expense that is adequate for a given river basin-scale simulation, and 5) On Demand ML can be used to gradually reduce the number of full predictive calculations that are needed to describe the watershed to river basin-scale biogeochemical function, essentially replacing full physics-based simulation with continuously improving surrogate models (Leal et al 2020).

DOI: https://doi.org/10.2172/1769757
DOE BER, AI4ESP 2021 #AI4ESP1126 (https://www.ai4esp.org)



The different levels within the multi-fidelity simulation framework include:

ML-automated selection of data and modeling layers: The predictive water and biogeochemical high resolution and surrogate models require multiple data layers (climate forcing, vegetation, microbes, soils, …) that may exist at different scales. These data layers can be selected based on ML approaches and automatically downscaled or upscaled to the desired grid resolution.

ML-facilitated high-resolution biogeochemical reactive transport simulations: The next generation exascale high-performance parallel computing and quantum computing systems acquired by the U.S. Department of Energy (DOE) will enable running Earth System Models (ESMs) at hyper-resolution. In addition, porting of DOE software to heterogeneous (GPU-CPU) architectures, especially *Amanzi-ATS+Crunch* as part of the ExaSheds program, will enable catchment-scale mechanistic hydrology and biogeochemical reactive transport simulations at increasingly higher resolutions. The predictive capabilities of these hydro-biogeochemical models can be enhanced through the integration of downscaled meteorological forcing from larger-scale ESMs, such as E^{3}SM. Despite the increase in computing resources, the computational cost of high-resolution hydro-biogeochemical models may still be prohibitive. A foreseeable strategy to cope with this high computational cost is adaptive mesh refinement (AMR), where the mesh is dynamically refined in regions of interest and coarsened elsewhere (Özgen-Xian et al 2020). Here, the multi-physics nature of these simulations requires a mesh refinement strategy that considers multiple data from both hydrology and biogeochemistry without overrefining the mesh. The results of these high-resolution hydro-biogeochemical models may fill the gaps of observational datasets and provide additional synthetic data for machine learning applications, for example, surrogate models (based on Gaussian process regression, dynamic mode decomposition, random forest, or neural networks) and the on-demand multi-fidelity framework.

1D flowtube approaches for fully transient flow and reactive transport: A form of efficient and lower-fidelity modeling approach is to reduce the problem dimensionality by representing the flow system as an ensemble of one-dimensional flow-lines or tubes. This type of approach has been used in the past to model reactive transport through aquifers under saturated, steady-state groundwater flow (Ackerer et al 2020). Biogeochemistry is then simulated on the equivalent ensemble of one-dimensional tubes, and results can then be remapped to the original 2D or 3D medium. Reducing the problem to a single dimension may significantly reduce computational

times, thanks to greatly increased parallel scalability, and the loss of accuracy may be low as long as the impact of transverse dispersion on the overall results is negligible. However, the seasonal dynamics of watersheds involve critical differences with respect to steady-state groundwater flow that require further development of the method. Due to strong seasonal variations in boundary conditions and forcings, water flow direction will change over time, and the medium will go through local and global water loss and accumulation, resulting in water saturation variations. This means that the ensemble of tubes that represents the watershed or any portion of it should be dynamic, periodically updating its geometry, and able to correctly account for local and global water loss and accumulation, thus fulfilling conservation of mass. This first of its kind methodology will be developed and included in the suite of available models for On Demand ML.

Surrogate models from training on synthetic high-resolution data: Much of our understanding of ecosystem function, including the integrated water and biogeochemical cycles, stems from high-fidelity physics-based models. Although these high-fidelity models can provide a detailed simulation of how the future climate extremes will impact biogeochemistry in river basins, they are computationally expensive and extremely time-consuming, making them unsuitable for the multitudinous runs needed to evaluate the complex interactions of processes ranging from the bedrock to the canopy. The use of surrogates or emulators within the On Demand machine learning multi-fidelity framework can significantly limit the prohibitive computational costs of high-fidelity simulations while capturing the dynamics of underlying processes. The overarching benefit of surrogates is their ability to reduce complexity by learning the state variables' dynamics directly from the observational data or full output of a custom-built model (i.e., *synthetic high-resolution data*). As a whole, these data-informed or custom-built model-informed surrogates can easily be developed on the fly using machine-learning techniques, such as Gaussian process regression, dynamic mode decomposition, random forest, and neural networks (e.g., Lu and Tartakovsky, 2021), and seamlessly integrated with the larger multi-fidelity framework discussed in this White Paper.

On Demand ML-based model fidelity selection: Considering biogeochemical processes in reactive transport simulations is computationally expensive. By using an on demand machine learning (ODML) algorithm (Leal et al 2020), however, the computing costs for biogeochemistry and transport calculations can be reduced by orders of magnitude. The ODML model will start with zero knowledge at the beginning of the simulation. It will then gradually learn key biogeochemical calculations during the reactive transport simulation. These key calculations are then used as often as possible to predict similar calculations. The predictions are much faster because they do not require iterative algorithms; just a fast matrix-vector multiplication. In addition to ODML, other ML approaches can provide a uncertainty quantification (UQ) based approach relying on selective comparison with observational data and high resolution physics-based simulations to automatically choose the fidelity of a biogeochemical approach (e.g., high resolution RTM versus 1D flowtube versus surrogate) to balance the demands of computational efficiency and process fidelity.

Expected Results: The vision offered by this White Paper is to improve the scientific understanding and predictability of biogeochemical fluxes during climate extremes, an objective that is currently not possible with the existing set of approaches and tools. Machine learning will play a critical role in making this possible by providing variable resolution data, automated grid refinement, On the Fly model fidelity selection, and On Demand creation of surrogate models that overcome the computational challenges of pure physics-based models.

Suggested Partners/Experts: This vision can be achieved by partnering with ORNL (Scott Painter, the USGS (Allison Appling, David Lesmes), and the University of Texas Austin (Alexander Sun) to develop a comprehensive multi-fidelity approach for both hydrology and biogeochemistry. To take advantage of exascale computing resources, we can partner with ORNL (Ethan Coon, Scott Painter) and LANL (David Moulton) on GPU computing (Amanzi-ATS on heterogeneous architectures via ExaSheds project), as well as with the Exascale Computing Project, where groundbreaking work on linear and nonlinear solvers for GPU systems is underway (Alexander et al. 2020). The effort will make full use of DOE facilities like ARM, SAIL, and NERSC.

References

Ackerer J, Steefel C, Liu F, Bart R, Safeeq M, O'Geen A, Hunsaker C and Bales R (2020). Determining how Critical Zone structure constrains hydrogeochemical behavior of watersheds: Learning from an elevation gradient in California's Sierra Nevada. *Frontiers in Water* **2:23**. DOI: 10.3389/frwa.2020.0002.

Alexander, F., A. Almgren, … D.B. Kothe, …C.I. Steefel, … (2020). Exascale applications: skin in the game, *Philosophical Transactions of the Royal Society A* **378** (2166), 20190056.

Cromwell Erol, Shuai Pin, Jiang Peishi, Coon Ethan T., Painter Scott L., Moulton J. David, Lin Youzuo, Chen Xingyuan (2021). Estimating watershed subsurface permeability from stream discharge data using Deep Neural Networks. *Frontiers in Water*. DOI: 10.3389/feart.2021.613011.

Leal, A.M., Kyas, S., Kulik, D.A. and Saar, M.O. (2020). Accelerating reactive transport modeling: on-demand machine learning algorithm for chemical equilibrium calculations. *Transport in Porous Media* ***133*****(2)**: 161-204.

Lu H, Tartakovsky DM (2021) Dynamic mode decomposition for construction of reduced-order models of hyperbolic problems with shocks. *J. Machine Learning for Modeling and Computation* **2**: 1-29.

Mital, U., Dwivedi, D., Brown, J.B., Faybishenko, B., Painter, S.L. and Steefel, CI (2020). Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. *Frontiers in Water **2**: 20.* DOI: *10.3389/frwa*.

Mital U, Dwivedi D, Özgen-Xian I, Brown JB, Steefel CI (submitted) Modeling spatial distribution of snow water equivalent by coupling precipitation and temperature with Lidar maps. *Journal of Geophysical Research Letters*.

Özgen-Xian, I., G. Kesserwani, D. Caviedes-Voullième, S. Molins, Z. Xu, D. Dwivedi1, J. D. Moulton, and C. I. Steefel (2020) Wavelet-based local mesh refinement for rainfall-runoff simulations. *Journal of Hydroinformatics*. DOI: 10.2166/hydro.2020.198