Using Machine Learning to Develop a Predictive Understanding of the Impacts of Extreme Water Cycle Perturbations on River Water Quality

Charuleka Varadharajan[1], Vipin Kumar[2], Jared Willard[2], Jacob Zwart[3], Jeff Sadler[3], Helen Weierbach[1], Talita Perciano[4], Juliane Mueller[4], Valerie Hendrix[4], Danielle Christianson[4]

[1] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory
[2] Department of Computer Science and Engineering, University of Minnesota
[3] United States Geological Survey, Integrated Information Dissemination Division, Water Mission Area
[4] Computing Sciences Area, Lawrence Berkeley National Laboratory

Focal Area(s)

This whitepaper addresses to two focal areas – (3) Insight gleaned from complex data using Artificial Intelligence (AI), and other advanced techniques (primary), and (2) Predictive modeling through the use of AI techniques and AI-derived model components (secondary). This topic is directly relevant to four DOE Earth and Environmental Systems Science Division Grand Challenges: integrated water cycle, biogeochemistry, drivers and responses in the Earth system, and data-model integration[1].

Science Challenge

10-year Vision: Use a modular AI/ML framework to understand and predict how stream water quality (for a variety of constituents) will respond to perturbations at short (days) and long (years) time scales, and at reach to basin spatial scales. Here, we describe how a framework comprising data integration, big data analytics, information theory, and knowledge-guided AI can address fundamental science questions and predict the impacts of perturbations such as floods and droughts on water quality.

Rationale

Motivation **-** Hydrologic disturbances such as floods and droughts can have drastic, long-term consequences to water quality in rivers and streams [2,3]. For example, flood events lead to increased hydrologic connectivity and runoff, resulting in soil erosion and constituent mobilization [4,5]. Droughts drive eutrophication and evaporation, resulting in altered redox conditions and increased salt, nutrient, and contaminant concentrations [6]. Thus, water quality changes due to floods and droughts can have direct, sometimes immediate and expensive, repercussions on human and ecosystem health [7,8]. *Hence, there is a pressing need to understand and predict water quality response and resilience to extreme water cycle disturbances at local to watershed to regional scales.*

Scientific gaps – ***Currently, physics-based deterministic watershed models cannot predict river water quality, much less the impacts of disturbance, at large spatial scales.*** High-fidelity multiphysics reactive transport models are inherently limited by incomplete process understanding that propagates to deficiencies in model structure. Moreover, these models computationally expensive, needing highly resolved spatial grids over vast domains to account for heterogeneous watershed characteristics that might influence water quality. Hybrid statistical/process-based basin-scale water quality models assume long-term steady-state behavior, and are just starting to incorporate temporal dynamics [9].

Many studies of how water quality varies with space, time, and disturbance, are focused on reach to catchment scales and show that local characteristics and complex processes impact solute transport [10–12]. However, ***it is challenging to determine the persistence and impacts on water quality resulting from abrupt hydrologic disturbances and their landscape interactions at watershed to basin scales.*** Thus, studying the impacts of disturbance events also requires new ML strategies to identify and classify response patterns to disturbances in complex watershed systems, particularly at larger spatial scales.

Current barriers **–** The prediction of water quality response to perturbations is challenging due to:

1) *Process complexity* – Water quality involves a wide range of constituents from base physical variables (e.g. temperature, salinity) to several chemical and biological constituents, each influenced by

DOI: https://doi.org/10.2172/1769795
DOE BER, AI4ESP 2021 #AI4ESP1135 (https://www.ai4esp.org)

different drivers in different regions and at varying timescales. For example, water temperature is strongly driven by climate but can also be influenced by snowmelt, groundwater influx, and reservoir and power plant operations. The extent to which different drivers control water quality can change during a disturbance – e.g., groundwater influences on salinity increase during drought. Biogeochemical reactions that impact constituents such as nitrogen (e.g. denitrification) add further complexity.
2) *Scaling and generalizability* – Hydrological disturbances span spatial and temporal scales (e.g., from highly localized flood events that span a few days to drought events that span several years over large regions). The scale of the disturbance also impacts the drivers of water quality. Building ML/hybrid models that translate across spatiotemporal scales is a significant challenge. Moreover, due to the spatial heterogeneity of watershed characteristics, scaling from data-rich small-scale test beds to other regions is challenging. Though parameter regionalization and classification of process-based models by catchment have attempted this type of scaling with mixed results [13], a robust way of bridging scales for accurate predictions of key water quality variables has not been identified.
3) *Persistence and memory effects* – In some cases, the impact can persist well beyond the disturbance, particularly in systems with long residence times. Compound events such as alternating flood and drought scenarios also impact responses differently due to system memory. These scenarios require models that include the lagged effects of drivers over long time-scales (sometimes decades).
4) *Multivariate, high dimensional data* – The complex interactions between processes affecting water quality necessitate the use of highly diverse multi-scale, multimodal data to improve predictability regardless of whether we use deterministic, ML, or hybrid models. These data requirements raise barriers such as discovering and integrating relevant data, representing complex data in ML models at relevant scales, high data dimensionality, and when necessary, untangling impacts of correlated drivers.
5) *Data discovery/sparsity* – Water quality monitoring data are sparse compared to the hydrobiogeochemical complexity involved, and often co-located datasets on water quality and its drivers are either not available or not easily discoverable if available at all. Traditionally, ML techniques are 'data-hungry' and advances in ML modeling need to overcome the data sparsity challenges.

Potential impacts – This approach would enable predictions of changes to water quality for different disturbances, provide decision makers an understanding of the drivers for more optimized management, and determine where measurements are needed to reduce prediction uncertainties.

Narrative

Machine learning models have shown early promise in predicting stream temperatures, a master water quality variable, and are much more computationally efficient compared to multiphysics reactive transport models. For example, Jia et al. (*In Press*) used a graph-based LSTM neural network pre-trained with deterministic model output to predict basin-scale stream temperatures across the Delaware River Basin [14]. Rahmani et al. (2021) used an LSTM neural network to predict point-scale stream temperatures for 118 catchments in the continental United States [15]. ML-based models for stream water quality variables beyond temperature are in development, but exist for lakes [16]. Machine learning approaches have shown promise for understanding episodic hydrochemical dynamics and can help interpret complex concentration-discharge (C-Q) relationships if the relationships are non-stationary [17,18].

There is an opportunity to have a modular, open source framework that brings together a variety of ML and related approaches to address the barriers, as described below.

1) Data discovery, integration, assimilation - Capabilities that enable discovery and integration of relevant water cycle and disturbance data for modeling would accelerate ML(Cholia et al., *submitted*)[19]. Assimilation of new data into models (e.g. Brajard et al. 2020) can also improve prediction outcomes[20].

2) Representation of complex data in ML models - Creating advanced multiscale graph-based data representations could represent river network structure and carry a diverse set of node/edge attributes to embed multimodal, multiscale, and multitemporal information, and correlate structures across different scales and time steps. Probabilistic Graphical Models are a powerful method to potentially incorporate prior knowledge. New optimization methods for data representation are needed and could include, e.g., removing dependencies on array-like representations and convolution operations.

3) ***Hypotheses and knowledge generation*** - Accurate, efficient ML models can be built by first conducting data-driven analyses to infer physical information – such as identifying key drivers of changes in water quality to inform ML model selection, architecture, hyperparameters and input features. The data-driven analyses can include a combination of statistical analyses, causal inference, feature detection and classification [21,22]. These analyses would improve ML model performance, reduce computational run-times and avoid overfitting, as compared to traditional machine learning approaches.

4) ML model improvements for extremes – Typical ML loss functions based on mean squared error make the ML models best suited to predict mean behavior. New approaches are needed to predict extremes. One possibility is to encode prior knowledge of data representing extreme events as an additional feature in the ML training. Another option is to devise new types of loss functions, which could, for example, use weight multipliers to encode the importance of certain events. Using an adaptive model refinement approach (see Mueller et al., *submitted*[23]), scale-aware ML will switch between different scales, using coarse models when no rapid changes are expected and finely-resolved models when a disturbance event is expected.

5) Physics-informed machine learning - Future climate projections anticipate more intense extreme events, which make it challenging to use ML models because they can only predict the range of events in the historical record. Physics-driven ML that uses deterministic model output for pre-training could help fill this gap, and additionally enable extrapolation of predictions to other regions without data.

6) Meta transfer learning - Recently, meta transfer learning has enabled the strategic transfer of models to improve prediction of data-sparse locations. This approach uses a meta-learning model to predict transfer performance of ML models using geophysical attributes and past performance metrics [24]. Meta transfer learning flexibly allows the use of any type of model built on well-observed locations, including the aforementioned ML models for extreme events, and is thus suitable for increased regional scaling.

7) ***Ensemble approaches*** - Most current ML approaches are brute-force and train several models with different architectures to identify the model with the least error. Ensemble approaches aim to exploit the diversity of predictive skill from different models and are showing promise at improving predictions and uncertainty quantification [25]. These approaches can be formulated in many ways such as running simulations on a class of models, pre-training ML models on different process models, or transferring models from similar locations that are better observed within a meta transfer learning framework.

8) Optimizing data collection - Arguably the best way to improve model predictions is to acquire relevant data when and where needed for reducing model uncertainties. Observing System Simulation Experiments (OSSEs) aim to identify how many and what types of observations are needed to reach a desired model performance metric [26]. OSSEs could be improved to expand beyond using process- based models to also include ML models, which may value a different set of observations. Other ML techniques in conjunction with OSSEs such as active learning methods can inform data collection [14].

This research will utilize unique DOE capabilities and datasets such as NERSC, data in the ESS-DIVE repository and AmeriFlux, and other software tools built for DOE projects. The framework will enable reproducible analysis and be open source to allow contributed plug and play modules. Data products will follow FAIR principles (e.g. standards for data integration).

Using Machine Learning to Develop a Predictive Understanding of the Impacts of Extreme Water Cycle Perturbations on River Water Quality

Suggested Partners/Experts

Jordan Read or Alison Appling (U.S. Geological Survey)

References

1. U.S. DOE. *Climate and Environmental Sciences Division Strategic Plan 2018–2023*. (2018). doi:DOE/SC–0192
2. Nilsson, C. & Renofalt, B. M. Linking Flow Regime and Water Quality in Rivers: a Challenge to Adaptive Catchment Management. *Ecol. Soc.* **13**, (2008).
3. Mosley, L. M. Drought impacts on the water quality of freshwater systems; review and integration. *Earth-Science Rev.* **140**, 203–214 (2015).
4. Lyubimova, T., Lepikhin, A., Parshakova, Y. & Tiunov, A. The risk of river pollution due to washout from contaminated floodplain water bodies during periods of high magnitude floods. *J. Hydrol.* **534**, 579–589 (2016).
5. Roussiez, V., Probst, A. & Probst, J.-L. Significance of floods in metal dynamics and export in a small agricultural catchment. *J. Hydrol.* **499**, 71–81 (2013).
6. Murdoch, P. S., Baron, J. S. & Miller, T. L. Potential Effects of Climate Change on Surface-Water Quality in North America. *JAWRA J. Am. Water Resour. Assoc.* **36**, 347–366 (2000).
7. Chang, H. & Bonnette, M. R. Climate change and water-related ecosystem services: impacts of drought in California, USA. *Ecosyst. Heal. Sustain.* **2**, e01254 (2016).
8. Gleick, P. H. Introduction: studies from the water sector of the national assessment. *JAWRA J. Am. Water Resour. Assoc.* **35**, 1297–1300 (1999).
9. Chanat, J. G. & Yang, G. Exploring Drivers of Regional Water-Quality Change Using Differential Spatially Referenced Regression—A Pilot Study in the Chesapeake Bay Watershed. *Water Resour. Res.* **54**, 8120–8145 (2018).
10. Lintern, A. *et al.* What Are the Key Catchment Characteristics Affecting Spatial Differences in Riverine Water Quality? *Water Resour. Res.* **54**, 7252–7272 (2018).
11. Winnick, M. J. *et al.* Snowmelt controls on concentration-discharge relationships and the balance of oxidative and acid-base weathering fluxes in an alpine catchment, East River, Colorado. *Water Resour. Res.* **53**, 2507–2523 (2017).
12. Kim, H., Dietrich, W. E., Thurnhoffer, B. M., Bishop, J. K. B. & Fung, I. Y. Controls on solute concentration-discharge relationships revealed by simultaneous hydrochemistry observations of hillslope runoff and stream flow: The importance of critical zone structure. *Water Resour. Res.* **53**, 1424–1443 (2017).
13. Archfield, S. A. *et al.* Accelerating advances in continental domain hydrologic modeling. *Water Resour. Res.* **51**, 10078–10091 (2015).
14. Jia, X. *et al.* Graph-based Reinforcement Learning for Active Learning in Real Time: An Application in Modeling River Networks. *In Press.* Proceedings of the 2019 SIAM International Conference on Data Mining.
15. Rahmani, F. *et al.* Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.* (2020). doi:10.1088/1748-9326/abd501
16. Hanson, P. C. *et al.* Predicting lake surface water phosphorus dynamics using process-guided machine learning. *Ecol. Modell.* **430**, 109136 (2020).
17. Zhang, Q., Harman, C. J. & Kirchner, J. W. Evaluation of statistical methods for quantifying

fractal scaling in water-quality time series with irregular sampling. *Hydrol. Earth Syst. Sci.* **22**, 1175–1192 (2018).

18. Hamshaw, S. D., Dewoolkar, M. M., Schroth, A. W., Wemple, B. C. & Rizzo, D. M. A New Machine-Learning Approach for Classifying Hysteresis in Suspended-Sediment Discharge Relationships Using High-Frequency Monitoring Data. *Water Resour. Res.* **54**, 4040–4058 (2018).
19. Cholia, S., Varadharajan, C. & Pastorello, G. Z. *Integrating Models with Real-time Field Data for Extreme Events: From Field Sensors to Models and Back with AI in the Loop*. (Submitted, AI4ESP 2021 whitepaper).
20. Brajard, J., Carrassi, A., Bocquet, M. & Bertino, L. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *J. Comput. Sci.* **44**, 101171 (2020).
21. Hubbard, S. S., Varadharajan, C., Wu, Y., Wainwright, H. & Dwivedi, D. Emerging technologies and radical collaboration to advance predictive understanding of watershed hydrobiogeochemistry. *Hydrol. Process.* **n/a**, (2020).
22. Tsai, W.-P., Fang, K., Ji, X., Lawson, K. & Shen, C. Revealing Causal Controls of Storage-Streamflow Relationships With a Data-Centric Bayesian Framework Combining Machine Learning and Process-Based Modeling. *Front. Water* **2**, 40 (2020).
23. Mueller, J., Varadharajan, C., Woodburn, E. & Koven, C. D. *Machine Learning for Adaptive Model Refinement to Bridge Scales*. (Submitted, AI4ESP 2021 whitepaper).
24. Willard, J. D. *et al.* Predicting Water Temperature Dynamics of Unmonitored Lakes with Meta Transfer Learning. (2020). https://arxiv.org/abs/2011.05369
25. Jiang, S., Ren, L., Yang, X., Ma, M. & Liu, Y. Multi-model ensemble hydrologic prediction and uncertainties analysis. *Proc. Int. Assoc. Hydrol. Sci.* **364**, 249–254 (2014).
26. Masutani, M. *et al.* Observing System Simulation Experiments BT - Data Assimilation: Making Sense of Observations. in (eds. Lahoz, W., Khattatov, B. & Menard, R.) 647–679 (Springer Berlin Heidelberg, 2010). doi:10.1007/978-3-540-74703-1_24