Knowledge-Guided Machine Learning (KGML) Platform to Predict Integrated Water Cycle and Associated extremes

Dipankar Dwivedi[1], Grey Nearing[2], Hoshin Gupta[3], Alden Keefe Sampson[4], Laura Condon[3], Benjamin L. Ruddell[5], Daniel Klotz[6], Frederik Kratzert[6], Uwe Ehret[7], Laura Read[4], Praveen Kumar[8], Ty Ferre[3], and Carl Steefel[1]

[1]Lawrence Berkeley National Laboratory, [2]University of California Davis, [3]The University of Arizona, [4]Upstream Tech, [5]Northern Arizona University, [6]Johannes Kepler University Linz, Austria, [7]Karlsruhe Institute of Technology (KIT), Germany, [8]University of Illinois at Urbana-Champaign

Focal Area(s): Predictive modeling through the use of AI techniques and insight gleaned from complex data (both observed and simulated).

Science Challenge: Although advanced predictive capabilities of the water cycle are critical to address environmental needs and develop sustainable solutions for energy demands, there is no robust framework, to say the least, that seamlessly integrates local to intermediate to global scales and a gamut of biogeophysical information to enhance understanding of the integrated water cycle and its associated extremes.

Rationale: Hydrological functioning of watersheds and river basins responds to changes in precipitation patterns, climate, extreme weather, geomorphology, vegetation cover, nutrient and contaminant loading, as well as disturbances (e.g. land-use change) and extreme events (e.g. droughts, floods, and wildfire). Hydrology exhibits tremendous variability across scales; local processes influence medium to large scale responses and vice versa. To ensure water security—critical for safeguarding human health, ensuring energy and food production, and enhancing economic growth—a robust predictive understanding of integrated water cycles and associated extremes is needed. To understand the water cycle, the science community has made significant efforts, such as developing unifying concepts and linking small scale (e.g., hillslope) process understanding with large scale (e.g., watershed) hydrological responses while respecting landscape heterogeneity, a dominant control [1–5]. Due to this, considerable progress has been achieved toward understanding the integrated water cycle, yet large uncertainties remain regarding to what extent local-scale heterogeneities and anomalies, including natural and anthropogenic system components, drive larger-scale hydrological processes and phenomena, and how persistent and predictable are these interactions [6, 7]. An integrated Holoarchic understanding is needed to address underlying challenges posed by the complexity of the water cycle as follows: (a) predictability of extremes, (b) short, medium and long-term predictions (triggered by perturbations at all scales), (c) frequency and intensity of hydrological events (affected by local fluxes and coupled medium to large scale variability), and (d) predictions of hydro-biogeochemistry.

Narrative: We propose to address the scientific gap regarding how to combine biogeophysical understanding with AI/ML for hydrological systems from local to intermediate to global scales. To be clear, we are not proposing to solve a specific problem in physical hydrology, and we are not proposing to build a new modeling system (although both will be outcomes of this venture).

DOI: https://doi.org/10.2172/1769733
DOE BER, AI4ESP 2021 #AI4ESP1037 (https://www.ai4esp.org)

We propose to build a knowledge-guided machine learning (KGML) platform, which is a research question in its own right and must be tackled in ways unique to the hydrological sciences. However, the framework will be generalizable to many other fields of geosciences.

Hydrology has unique challenges, related to understanding and predicting integrated, scale-relevant, dynamic behaviors, that are different from other disciplines. Although there are many strategies for combining physics with machine learning, there is no one-size-fits-all solution to the KGML problem that works across disciplines. We see this as the single most important question for the future of the hydrological sciences, and addressing it in an ad-hoc way (e.g., by making KGML a subcomponent of specific hypothesis-driven research projects) is inadequate. Addressing this problem will require truly cross-disciplinary teams (of computer scientists and hydrologists).

The practical outcome we intend is a community research and modeling program built from the ground-up with an "AI-First" perspective. Start with what we can learn from data and then learn where biogeophysical theory adds value in different areas, timescales, and hydrological regimes. Our vision is for the hydrological sciences community to build a community modeling program (like ACME, CLM, JULES, NOAA, CHTESSLE, LIS, etc.) from the ground up, with machine learning. We can nurture this program in testbeds such as the Delaware River Basin, Upper Colorado Water Resources Region, and East River Watershed, the site of the Berkeley Lab's Scientific Focus Area. These testbeds offer unique opportunities by exploiting long-term rich data (leveraged through DOE and USGS investments; e.g., *SFA, ExaSheds, SAIL, NGWOS*) and furthering much needed *open-science* through collaborations. Additionally, to fully leverage AI's strength in big data applications, testbeds will be expanded to incorporate large scale national and global datasets to evaluate the generalizability of modeling hypotheses. This will provide infrastructure for a new type of hypothesis testing, causal analysis, and simulation that is (i) *explainable* and (ii) *yields improved forecasting and prediction* across timescales and hydrological regimes. Not to mention, these initiatives will honor the FAIR principles in all aspects to maximize utility and impacts. Although there are now numerous examples of all of these aspects of KGML in hydrology, the community currently lacks this "AI-First" perspective. We argue that treating AI as simply a tool in standard research and modeling workflows is not optimal for capturing AI/ML's full power.

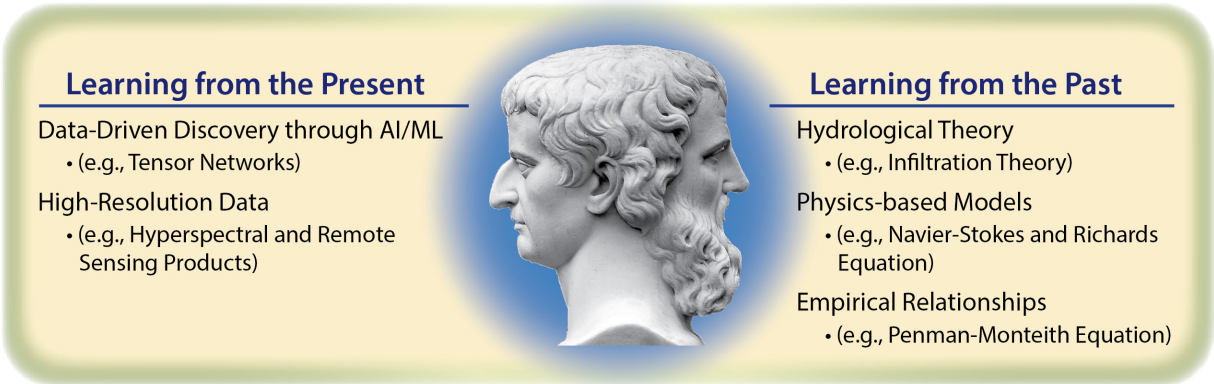

Figure 1: ***The two-faced Roman Deity, Janus, could see into the past with one face and the future with the other. Analogous to Janus, the KGML framework implies change and transition through learning from the past and moving into the future. A rich body of literature from the past can guide advanced AI/ML approaches to reliably predict integrated water cycle and associated extremes.***

Opportunities, Approach, and Activities: There are two sources of information - data and theory - and no universal strategy for combining these two types of information exists (see, Figure 1). KGML will give us ways to structure complex amalgamations of scientific theory and data-driven discovery in the same modeling package (e.g., tensor networks). There are at least eight classes of KGML strategies (to be discussed in our white paper) to integrate scientific theory and AI/ML

strategies, yet the community does not yet have a strong grasp of how these strategies function for the types of problems we work on. The challenge we face is technical, but also cultural. Current funding programs in hydrological science primarily fund individual hypothesis-driven questions that may or may not include AI/ML-driven tools, and no avenue exists for large-scale research and development of AI/ML as a central platform for hydrological science. AI/ML is our best path toward extracting the most information from data and models, and the community currently lacks a program to develop this capability except in piecemeal as components of individual projects. We need a program like ACME that is developed from an "AI-First" perspective. Examples of specific challenges that such a platform would address are:

1. How to resolve data sparsity issues in ML, as they require many observations on the ground (e.g., getting good precipitation data is challenging)?
2. How to make full use of all the information available? On the one hand, ML faces data scarcity, and on the other hand, we are not even close to fully utilizing all the observational data that is already available because our data-ingesting abilities are limited. Further, we choose to focus on a handful of variables that we assume to be important rather than considering the full range of data variables.
3. How to train AI/ML approaches without exposing them to implicit biases? Note that a traditional approach is doing hydrology backward, i.e., inversion from response variables such as precipitation. However, ML is likely to learn the error (implicit bias) of the data products when mapping them backward.
4. How to integrate different sources of information (e.g., getting a better estimate of precipitation from a suite of observations)?
5. How to reduce uncertainty when using indirect information through inference? Direct measurements are limited in hydrology; only a few state variables (e.g., river stage) can be measured, and the rest are inferred. These inferences introduce large uncertainties.
6. How can we explicitly use the rich data collected from the Earth's skin to say something about what is going on beneath?
7. How to utilize all the data and the theory and the linkages that are missing across them?
8. How do we transfer knowledge across time and space?

Why is ML/AI Key? The laws of physics are universal and invariant in space-time. New AI/ML advancements within the KGML framework should take advantage of big data, theory, and physical laws (Figure 1). A suite of AI/ML models in KGML will essentially preserve what they have learned from the past while updating themselves to predict what is happening.

Significance for Extreme Water Cycles: The KGML framework will combine observations and simulation capabilities to (a) enhance predictive capabilities for integrated water cycles and extremes events that are likely to occur more frequently, (b) maximally quantify and reduce the uncertainty by using all available data, and (c) more reliably and universally project future climate and extreme environmental states.

What is the 10-Year Vision? In ten years, the community should have a solid understanding of two things. First, how to rapidly deploy KGML tools for measuring the information value of new hypotheses. Second, we should have some understanding of which parts of our current body of hydrological theory add information (value) to predictions of different aspects of the water cycle at different spatiotemporal scales, in the context of models that can extract as much information as possible directly from (necessarily incomplete) Earth-observing datasets. Answers to these questions will come from integrating hydrological theory piece-by-piece into comprehensive multi-scale AI/ML models.

References

[1] K. Schulz, R. Seppelt, E. Zehe, H.-J. Vogel, and S. Attinger, "Importance of spatial structures in advancing hydrological sciences," *Water Resources Research*, vol. 42, no. 3, 2006.

[2] M. Sivapalan, S. Thompson, C. Harman, N. Basu, and P. Kumar, "Water cycle dynamics in a changing environment: Improving predictability through synthesis," *Water Resources Research*, vol. 47, no. 10, 2011.

[3] S. S. Hubbard, C. Varadharajan, Y. Wu, H. Wainwright, and D. Dwivedi, "Emerging technologies and radical collaboration to advance predictive understanding of watershed hydrobiogeochemistry," *Hydrological Processes*, vol. 34, no. 15, pp. 3175–3182, 2020.

[4] X. Chen, R. M. Lee, D. Dwivedi, K. Son, Y. Fang, X. Zhang, E. Graham, J. Stegen, J. B. Fisher, D. Moulton, *et al.*, "Integrating field observations and process-based modeling to predict watershed water quality under environmental perturbations," *Journal of Hydrology*, p. 125762, 2020.

[5] T. Wagener, M. Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H. V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, and J. S. Wilson, "The future of hydrology: An evolving science for a changing world," *Water Resources Research*, vol. 46, no. 5, 2010.

[6] G. Blöschl, M. F. Bierkens, A. Chambel, C. Cudennec, G. Destouni, A. Fiori, J. W. Kirchner, J. J. McDonnell, H. H. Savenije, M. Sivapalan, *et al.*, "Twenty-three unsolved problems in hydrology (uph)–a community perspective," *Hydrological sciences journal*, vol. 64, no. 10, pp. 1141–1158, 2019.

[7] G. Destouni, F. Jaramillo, and C. Prieto, "Hydroclimatic shifts driven by human water use for food and energy production," *Nature Climate Change*, vol. 3, no. 3, pp. 213–217, 2013.