Title: Mapping hydrologic and biogeochemical information flows to improve predictive models and understand climate influence

Author: Zexuan Xu, Earth and Environmental Science Area, Lawrence Berkeley National Laboratory

Focal Area: Insights gleaned from complex data (both observed and simulated) using AI, big data analytics, and other advanced methods, including explainable AI and physics or knowledge-guide AI

Science Challenge:

Current hydrological and biogeochemical model benchmarking is insufficient to evaluate the performance discrepancies, which is important in model development. There are research gaps and urgent needs to reveal the components of the model are performing poorly, or attribute performance shortcomings to the uncertainties of data, parameters, and/or mechanistic uncertainties. On the other hand, when models are performing well, rarely is it investigated whether they are "right for the right reasons", meaning that the model outputs are accurate and that accuracy is the result of accurate representations of constituent processes, rather than the result of cross canceling errors. Additionally, the spatiotemporal variability in the dominant mechanism controlling geochemical fluxes, and two-way feedbacks with changing climate at various timescales are missing with advanced analytics tools.

Rationale:

The existing challenges in filling the research gaps include: 1) Scarcity of observational datasets, particularly geochemistry data, in and beyond the intensive-instrumented region/watershed, for such information flow mapping studies; 2) The hurdles of mechanistic model, e.g., processes complexitivity, computationally intensive, data requirement and scalability, hinder the statistical robustness of information flow over many model realizations; 3) The developments, evaluations and applications of data-driven machine learning (ML)/AI in hydrological and biogeochemical models are limited, thus result in insufficient representative of the model diversity hence predictivities and capabilities for the science questions.

The success of the proposed research could: 1) Gain insights of hydrological and biogeochemical interaction in addition to the physical-based models. This is particularly helpful to identify the environmental factors that sustain "hot spot and hot moment" of biogeochemical at the watershed scale, and addresses one of the key research questions of the EESSD's biogeochemistry grand challenges; 2) Enhance hydrological and biogeochemical predictability of both mechanistic and ML models with causation and knowledge gained from the information flow analysis. Particularly, the information flow analysis results could provide additional and/or refine features for the ML models, and assist model selection with appropriate metrics.

Narrative:

The advantages of information theory have been recognized over traditional statistical approaches spread and used in the earth system science. Information, or entropy, quantified as $H = p \log p \, dx$ (Shannon, 1948), where p is the probability that random variable X_i is equal to x, is a measure of the total uncertainty contained within variable X. Information theory quantifies the

DOI: https://doi.org/10.2172/1769747

amount of uncertainty reduction in a dependent variable that originated from knowledge of a driver variable. The variation of entropy is computed as transfer entropy (TE), an indicator of causality that requires no prior knowledge or explicit functional relationship between the predictors and response over time (Schreiber, 2000). In this manner, information theoretic analyses have been used to define process networks of functional interactions and feedback among earth systems variables (Goodwell et al., 2018; Ma et al., 2018; Ruddell and Kumar, 2009a, 2009b). Process networks can define dominant controls on fluxes of carbon and water and the timescales over which they occur (Larsen and Harvey, 2017; Tennant et al., 2019), and delineate the dominant spatial pathways through which fluxes occur within a watershed (Rinderer et al., 2018). Similarly, by quantifying the mutual (i.e., shared) information between model predictions and an observational time series and normalizing that shared information by the total information content present within the observations, it is possible to quantify precisely how much of the information content apparent in an output variable of interest is captured by a model. Usually, TE is computed directly from data, with a significance level associated with the calculated strength of the directional interaction between the two variables. However, the application of information theory is rare in model benchmarking for understanding the hydrological and biogeochemical physics, and assessing model performance.

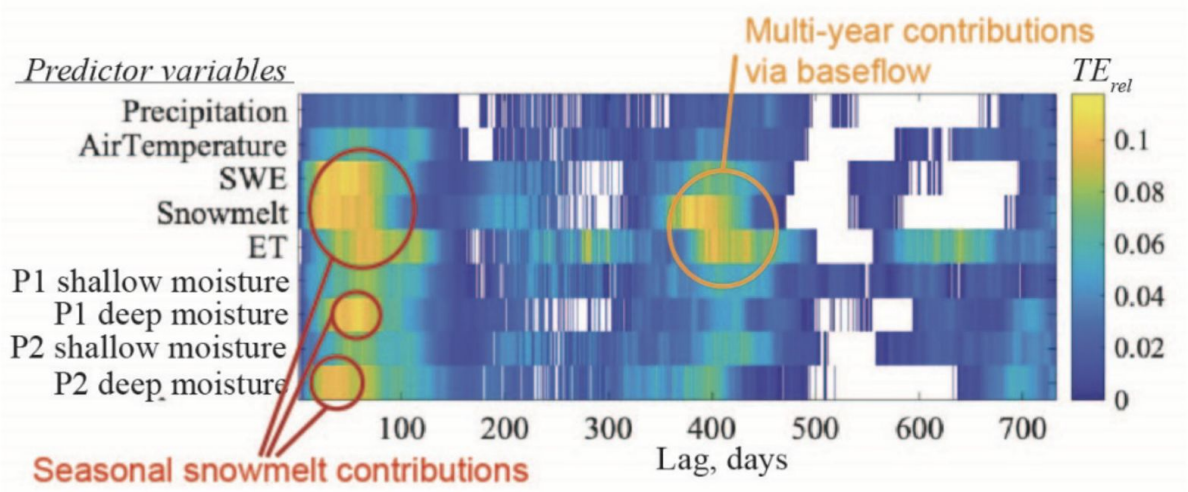

Fig. 1. Information flows at the Dry Creek Experimental Watershed, Idaho, from predictor variables (y-axis) to stream discharge at a range of time lags (x-axis). Colors quantify transfer entropy relative to the total uncertainty in discharge. In this catchment, discharge is dominated by seasonal snowmelt contributions. Each year's snowmelt also contributes to the next year's streamflow via percolation to deep groundwater (modified from Tennant et al., 2019).

Compared to the information flow mapping techniques, classical methods (e.g., sensitivity analysis) do not fully capture the appropriate timescale over which hydrologic transport processes occur. Artificial neural network (ANN) models are generally considered state-of-the-art in their ability to replicate the causation between predictor variables to a response variable, therefore ideal estimators of the benchmark performance should be achievable by a

mechanistic model containing accurate physics and parameter values. For example, even simple one-layer ANN could map all of the hydroclimatic, hydrological, and biogeochemical predictors to the biogeochemical response variable of interest, with potential integration of other deep learning techniques (Nearing et al., 2018). Geochemical fluxes should be partitioned into discharge and concentration components, and aggregated to evaluate the mutual information for each set of outputs to determine each mechanistic model component's contribution to error. Integrated with the mechanistic modeling developments and applications, e.g., the high-resolution ATS biogeochemical model, can provide additional modeling data for benchmarking, with the projections of climate change forcing simulated by larger-scale, refined-resolution earth system models, e.g., E3SM.

The variability of hydroclimate, e.g., extreme precipitation/drought, and/or wet years vs. dry years, could influence dominant mechanisms controlling the geochemical behaviors. However, observational data are usually insufficient for such analysis and lack capabilities for such analysis. Benchmarking comparison of mapping information flow could add insights on the understanding and evaluation of extreme events on biogeochemical processes across temporal and scale scales, as they are sensitive to the perturbation of surface-subsurface water system and land cover modification induced by climate change. It is expected that climatological phenomena affect the whole watershed synchronously, whereas the benchmarking provides generalizable information contained within both global and local, site or scale specific processes.

The long-term vision of this proposed research is to create an information flow analysis toolbox and framework for a larger group of model benchmarking and analysis. The success of this research could provide a better understanding of the dominant processes controlling biogeochemical cycling under different hydroclimatic regimes and different spatial scales. This research could create a benchmarking community and build up a bridge for sharing knowledge, and connecting model development, evaluation and applications under climate change across scales. The benchmarking and diagnosis analyses will lead to suggestions for improvement of the mechanistic models, with an effective incorporation. Multiple open-source codes and software have been created and will be leveraged for applying information-theoretic analysis and performing model benchmarking. New development of software products and tools are still essential for generating the processing workflow for preparing data time-series for information-theoretic analysis, applying information-theoretic analysis, generating data-driven predictions of nitrogen fluxes. All newly generated codes should make use of open-source libraries and open-source licenses. All raw and intermediate data products should be made available to the public in accordance with the reproducible research standards.

This paper is inspired by the USGS Powell Center for Synthesis working group on watershed storage and control, the workshop on Critical Timescales of Hydrologic Transport held on May 22-24, 2019 at Berkeley Institute of Data Science, UC Berkeley, the Early Career Critical Zone Network-of-networks Workshop during the AGU Fall Meeting, and the DOE SBR university-lead proposal (unfunded) "The nitrogen cycling informationscape: mapping watershed information flows to improve predictive models and understand climate influence" in 2019.

Suggested Partners/Experts:

Laurel Larsen, Associate Professor, University of California, Berkeley

Benjamin Ruddell, Associate Professor, Northern Arizona University

Praveen Kumar, Professor, University of Illinois at Urbana-Champaign

References:

Goodwell, A.E., Kumar, P., Fellows, A.W., Flerchinger, G.N., 2018. Dynamic process connectivity explains ecohydrologic responses to rainfall pulses and drought. Proc. Natl. Acad. Sci. 115, E8604–E8613.

Larsen, L.G., Harvey, J.W., 2017. Disrupted carbon cycling in restored and unrestored urban streams: Critical timescales and controls. Limnol. Oceanogr. 62

Ma, H., Larsen, L.G., Wagner, R.W., 2018. Ecogeomorphic Feedbacks that Grow Deltas. J. Geophys. Res. Earth Surf. 123, 3228–3250. https://doi.org/10.1029/2018JF004706

Nearing, G.S., Gupta, H.V., 2018. Ensembles vs. information theory: supporting science under uncertainty. Front. Earth Sci. 12, 653–660. https://doi.org/10.1007/s11707-018-0709-9

Rinderer, M., Ali, G., Larsen, L., 2018. Assessing structural, functional and effective hydrologic connectivity with brain neuroscience methods: State-of-the-art and research directions. Earth Sci. Rev.

Ruddell, B.L., Kumar, P., 2009a. Ecohydrologic process networks: 1. Identification. Water Resour. Res. 45, doi: 10.1029/2008WR007279.

Ruddell, B.L., Kumar, P., 2009b. Ecohydrologic process networks: 2. Analysis and characterization. Water Resour. Res. 45, W03420.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech J 27, 379–423.

Tennant, C., Larsen, L.G., Bellugi, D., Moges, E., Ma, H., Zhang, L., The utility of information flow in formulating discharge forecast models: a case study from an arid snow dominated catchment. Water Resour. Res., WR024908