Deep Learning for Hydro-Biogeochemistry Processes

Authors

Li Li (lili@engr.psu.edu), Wei Zhi, Chaopeng Shen, Dept. Civil & Environ. Engr., Penn State Univ.
Eric Pierce, Dan Lu, Oak Ridge National Laboratory

Focal Area(s)

The white paper focuses on areas 2 and 3. The proposed deep learning modeling approach can provide forecasting capabilities and mechanism-based understanding of terrestrial and aquatic hydro-biogeochemical processes at watershed and river basin scales. Knowledge gleaned here can also provide insights for Earth System Models.

Science Challenge

Deep learning approaches have gained momentum in hydrological forecasting [Shen, 2018] but not in understanding terrestrial and aquatic hydro-biogeochemical processes and in forecasting water quality. Hydro-biogeochemical processes regulate the concentration levels and timing of stream solutes including nutrients, dissolved inorganic and organic carbon, among others. A significant barrier to the development of deep learning approaches in this arena is the scarcity of soil and rock biogeochemical data and stream chemistry. These data often have large temporal gaps (multiple years to decades), low sampling frequency (weekly to seasonally), and sparse spatial coverage. Data scarcity presents a major barrier for mechanism-based understanding of solute transformation and transport from land to aquatic systems; it also imposes a significant roadblock for water quality forecasting. In contrast to this data scarcity challenge in subsurface and chemistry data, earth surface characteristics and hydrometeorological data have become largely available at high spatial and temporal resolution. Here we ask the question: **what and to what extent can we learn about the dynamics of hydro-biogeochemical processes from intensive hydrometeorological and earth surface data and scarce water chemistry measurements?** One of our recent studies has shown the promise of training a deep learning model using largely available hydrometeorology data to forecast dissolved oxygen, an important water quality measure (Fig. 1) [Zhi et al., 2021].

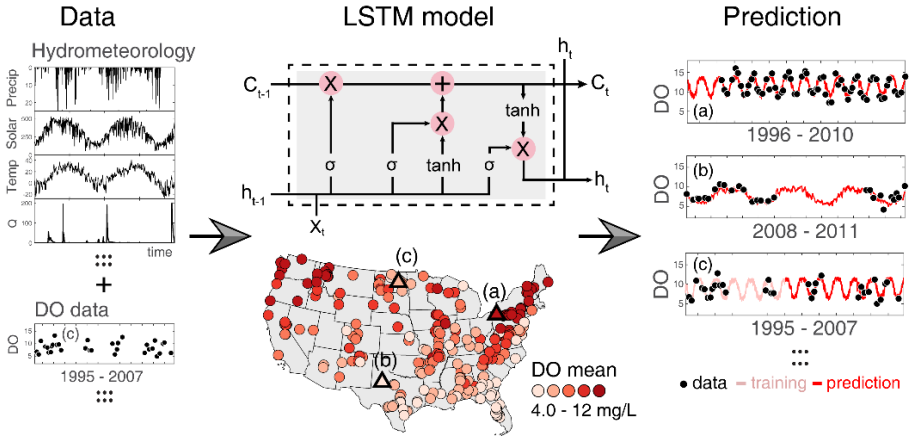

Fig. 1. Using hydrometeorological data to forecast water

We propose to develop deep learning models to understand and forecast the spatial-temporal dynamics of stream solutes at continental scales, using conus as a test-bed. The water chemistry measures can include, for example, carbon (dissolved organic and inorganic carbon), total nitrogen (TN), nitrate, and total phosphorous (TP), all of which are of societal significance. For example, TN and TP have caused persisting eutrophication

DOI: https://doi.org/10.2172/1769693
DOE BER, AI4ESP 2021 #AI4ESP1077 (https://www.ai4esp.org)

worldwide as a result of intensive agriculture and urbanization. In the U.S. alone, over 60% of U.S. estuaries and coastal water bodies have been degraded by excessive nutrient inputs.

Rationale

Watershed hydro-biogeochemical forecasting have traditionally used process-based models that are computationally expensive, scale dependent, and parameter non-unique. Existing usage of machine learning has focused on non-deep approaches using logistic regression, boosted classification trees, and artificial neural networks (for example, [Tesoriero et al., 2015]). One of the significant challenges in using deep learning approaches in the biogeochemistry and water quality fields is the data scarcity. Gauged basins with hydrometeorology and discharge data are often "chemically ungauged" (e.g., lack water chemistry data). With the exception of a few water quality quantities that can be measured by sensors, most solutes are measured biweekly or bimonthly, often missing large events such storms, and are often insufficient to train robust models. The Long-Short Term Memory (LSTM) network has previously been used to predict Dissolved Oxygen (DO), an important water quality measure, in individual basins with intensively measured, high frequency water quality data. Most places however do not have such luxury of rich water quality data. Our recent work indicate that intensive, largely available hydrometeorology data and watershed attributes can be used together with sparse DO measurements to overcome this challenge and forecast water quality [Zhi et al., 2021].

LSTM models can be developed for an array of stream solutes at the Conus scale where data are available. The development of LSTM models will 1) mark a new application of deep learning models in a new field, 2) offer the capability to understand watershed-scale hydro-biogeochemical processes and forecast water chemistry where data are typically sparse and scattered, and 3) alleviate the problem of data scarcity in the biogeochemistry and water quality community. The "filled" data can be used to understand temporal trend of water quality response to changing climate and human perturbation, a question that has long puzzled the water quality community because of the lack of consistent, historical data. The model output can also inform decision making and policy management, and offer insights on when and where to collect more data. In addition, insights gained at the watershed and river basins scales can provide guidelines for incorporating watershed-scale processes into Earth System Models.

Narrative

Datasets. We will use various data sources. For example, the GAGES II (Geospatial Attributes of Gages for Evaluating Streamflow, Version II) dataset has stream chemistry and daily discharge data for thousands of sites. An example solute, bicarbonate, is in Fig. 2. The GAGES-II data also include basin characteristics (e.g., climate features, geology, soils, topography, land use). Climate forcing such as precipitation, maximum and minimum air temperature, vapor pressure, and solar radiation will be from a gridded meteorological dataset (DAYMET, https://daymet.ornl.gov/) using the Google Earth Engine.

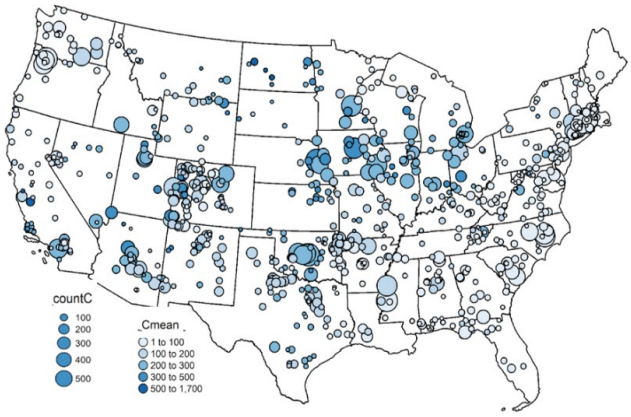

Fig. 2. Concentration map and data count for bicarbonate, an important water quality

The LSTM deep learning model. The LSTM network is a special type of recurrent neural networks that has overcome some weakness in traditional RNN and had been widely used to long-term dependence in time-series tasks. An LSTM layer consists of a set of recurrently connected blocks (i.e., memory cells) to store and pass sequential information. Each LSTM memory cell has three information gates (i.e., input gate, forget gate, and output gate) and two states (i.e., cell state and hidden state) to control what to flow in, what to forget, and what to memorize across time steps, allowing the network to learn long-term dependencies. This makes LSTM especially suitable for hydrological simulation where the lag times between precipitation and discharge can be up to years (e.g., water storage). LSTM architecture has been recently found to be useful for natural systems that store and release substance at multiple rates and exhibit hysteretic behaviors. A Bayesian LSTM can also be used to not only quantify output uncertainty but also to avoid overfitting caused by small training data [*Lu et al.*, 2019]. The LSTM network will be implemented in the open-source machine learning framework PyTorch. To accelerate model training, we will set up the code to leverage the highly GPU-optimized library of CUDA Deep Neural Network library (cuDNN). The cell and hidden states will be randomly initialized from a uniform distribution and we will perform multiple runs to consider the influence of randomness on training results. The models can be trained using daily time-series of hydrometeorological variables (i.e., precipitation, solar radiation, maximum and minimum air temperature, vapor pressure, day length, discharge), basin watershed attributes, and sparse water quality data.

The trained models and modeled data output will significantly alleviate the data scarcity challenge in the hydro-biogeochemistry community. They will also stimulate new questions and answers in understanding controls of surface and groundwater chemistry and in illuminating climate-carbon-water feedbacks.

References

Lu, D., Liu, S. and Ricciuto, D. (2019), An Efficient Bayesian Method for Advancing the Application of Deep Learning in Earth Science, paper presented at 2019 International Conference on Data Mining Workshops (ICDMW), 8-11 Nov. 2019.

Shen, C. (2018), A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, *Water Resour. Res.*, 54(11), 8558-8593.

Tesoriero, A.J., Terziotti, S. and Abrams, D.B. (2015), Predicting Redox Conditions in Groundwater at a Regional Scale, *Environmental Science & Technology*, 49(16), 9657-9664.

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C. and Li, L. (2021), from Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale?, *Environmental Science & Technology*, in press.