On AI Prediction of Hydrological Processes Based on Integration of Retrospective and Forecasting ML Techniques

Boris Faybishenko (EESA), Lavanya Ramakrishnan (CRD), Tom Powell (EESA), Bhavna Arora (EESA), John Wu and Deb Agarwal (all CRD), Lawrence Berkeley National Laboratory.

Focal areas: (1) Data assimilation enabled by machine learning, unsupervised learning (including deep learning), and (2) Predictive modeling through the use of AI techniques and AI-derived model components, comprising a hierarchy of models.

What is the 10-year vision?

The application of retrospective-predictive modeling will provide the information of what constitutes good governance for water, and what is needed for advancing water priorities and to transform water governance for the upcoming decades (Water Affordability…, 2020). The retrospective-predictive modeling approach is well suited for accurate forecasting of extreme meteorological and hydrological events, droughts and floods. The retrospective-predictive approach meets the FAIR principles of findability, accessibility, interoperability, and reusability. The approach will help increase the accuracy of simulations of extreme hydrological events at multiple spatial scales -- from a local to a sub-watershed and to the watershed scale, and multiple temporal scales -- from short-term to long-term scales. Solving the input and training data challenges through the help of machine learning algorithms for navigating software, predicting the many defects of datasets, advanced data correlation, enabling automation and providing graphical analysis of data coverage. This approach is suitable to resolve the following typical challenges of the AI/ML predictions: *compatibility* of data and models, *portability*, i.e., porting a model from one environment to another, *computing power* (advanced models may require large capacities of computer power), *scalability* (an ML environment should be able to scale up over time to meet performance and accuracy requirements), and *model size* (the model hosting environment should have sufficient storage and processing capabilities).

The 10-year vision includes the development of a scalable open-source software for advanced analysis techniques such as data mining, machine learning, pattern scaling, and visualization will enable scientific discovery for advancing water priorities. Advanced workflows will enable the reproducibility of research results and simulations and the ability to easily apply new techniques to existing datasets.

A modern and effective cyberinfrastructure for archiving, managing, analyzing, and visualizing experimental, observational, and model-generated data is critical for supporting scientific investigation of Earth system processes. Solving the AI/ML challenges of modeling and risk assessment of "extreme" water cycles through the help of self-learning algorithms for navigating software, enabling automation and providing graphical analysis of data coverage.

Goal

An important goal for the community is to present a concept of the application of robust multi-level AI prediction techniques, based on a combination of a *retrospective (training)* pipeline, generating multiple models to using different ML algorithms, and a *predictive* pipeline providing predictions of new data based on the model(s) generated in the retrospective pipeline. (*Note*: The goal corresponds to Task 3 of the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography--see https://www.ai2es.org/research/foundational-research-in-trustworthy-ai-ml).

DOI: https://doi.org/10.2172/1769756
DOE BER, AI4ESP 2021 #AI4ESP1042 (https://www.ai4esp.org)

Science Challenges

Traditional ML methodology being used in earth systems predictions has focused on the model development process, involving selecting the most appropriate algorithms for a given problem. Nearing et al (2020) have recently suggested that a central challenge in hydrology is how to integrate hydrological theory and machine learning techniques to add value to prediction systems. Lessons learned from the history of hydrological modeling motivate several clear next steps toward integrating machine learning into hydrological modeling workflows. It has become apparent that rigorous data and model preparation processes are necessary to ensure the stability and efficient performance of the models to ensure accurate data driven decisions. The challenging problem is that there is no QA/QC and scaling perspectives of datasets.

The other challenging problem that is constantly debated is whether training data are needed. For example, experiments by Kratzert et al. (2019) showed that DL models gave, on average, better daily streamflow predictions in basins where the model had never seen training data, compared to a process-based model. On the other hand, in their report of the IAHS community-wide effort to outline key 'Unsolved Problems in Hydrology' (UPH), Bl¨oschl et al. (2019) indicated that *"[m]ost hydrologists would probably agree that [extrapolating to changing conditions] will require a more process-based rather than a calibration-based approach as calibrated conceptual models do not usually extrapolate well."*

Hydrological models generally contain parameters that cannot be measured directly, but can only be meaningfully inferred by calibration against a historical record of input–output data (e.g., Vrugt et al., 2006; Araya1 and Ghezzehei, 2019). In addition, data used for the model development and initial conditions for modeling are often biased due to multiple errors, missing data, bad data, outliers, etc., so that the developed model would have been trained on such failure scenarios, causing inaccurate model predictions, and require continuous validation of ML models.

Rationale

The data is needed for effective Earth system research to better understand climatic and environmental variabilities and change over multiple scales continue to increase, both in volume and complexity. The rationale for using the AI prediction of ecohydrological processes is based on an integration of retrospective and forecasting ML techniques, which correspond to the research goals identified in the EESSD Strategic Plan (EESSD Strategic Plan, Page 19): "Develop and use innovative computational tools, testbeds, benchmarks, diagnostics, and metrics, as well as data at process and global scales, to evaluate and validate models and to characterize and understand sources of uncertainty in both observations and model simulations." Accurate predictions of extreme events--floods and droughts--are necessary for planning and risk assessment of using water and energy resources (e.g., Lee et al., 2005).

Narrative

The concept of the application of robust multi-level AI prediction techniques is based on a combination of a *retrospective (training)* pipeline, generating multiple models, which are evaluated using different algorithms to select the best model(s), and a *predictive* pipeline providing predictions of new data based on the model(s) generated in the retrospective pipeline. Retrospective modeling, which is the data- and solution-oriented data preparation pipeline, focuses on analyzing and interpreting observed patterns and structures in physics-based data to enable learning, reasoning, and decision making. The retrospective modeling pipeline also includes an application of novel ML algorithms for QA/QC for the development

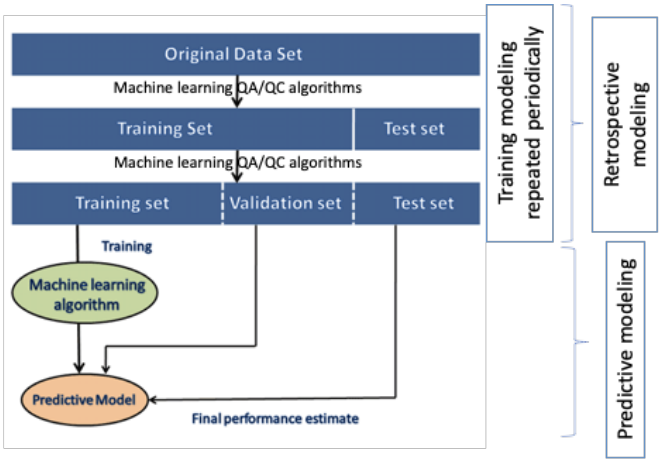

Figure 1. Flowchart of the retrospective and predictive machine learning modeling (modified from: https://www.javatpoint.com/how-to-get-datasets-for-machine-learning)

of training data. The length of the training data is increasing as new information becomes available, and new models are also developed. A schematic of the integration of the retrospective and predictive modeling is shown in Figure 1.

Deep learning as part of the retrospective pipeline uses neural networks to learn important features directly from data. In particular, neural networks are able to combine multiple nonlinear processing layers, typical for climatic and hydrological processes, using simple elements operating in parallel, achieving state-of-the-art accuracy in object classification and model development. Models are periodically trained as new real observations become available, using a large set of monitoring data, and neural network architectures are updated, usually including some convolutional layers. Although training the models is computationally intensive, it can be accelerated by using a high-performance GPU. The following ML methods are suitable for time series and spatial distribution analysis: Multiple Linear Regression, Stepwise regression and ridge and lasso regression, Logistic regression, Weighted regression, Nonlinear regression, Generalized Additive Models, Random Forest, Gradient Boosting, and Bayesian network analysis. For example, Dou and Yang (2018) compared an application of four different machine learning approaches in different terrestrial ecosystems.

For example, a Bayesian-network model can be used to assess the dynamic probabilistic dependencies between the hydrological and meteorological variables. An example of the Bayesian Network of Soil Water-Climate Balance Components is shown in Figure 2. Such networks are representations of a joint probability distribution of a set of random variables with a possible mutual causal relationship, and are represented using directed acyclic graphs. The procedure includes the development of a series of probabilistic functions and rank candidate distributions (e.g., the Akaike Bayesian information criteria), based on the prior information about the model parameters and functions. The Bayesian anomaly detection approach using Extreme Value Theory (EVT) is an effective probabilistic framework (Clifton et al., 2008; Guggilam et al., 2019) to detect anomalies and to explicitly model the normal and anomalous data as part of both retrospective and predictive modeling pipelines.

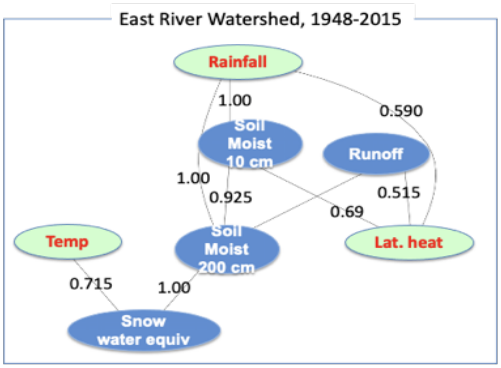

Figure 2. An example of the Bayesian Network of Soil Water-Climate Balance Components for the East River Watershed. (Numbers – arcs strengths>0.5)

Key collaborators

Clemson University, SC (Prof. Fred Molz).

References

Araya1, S.N. and T.A. Ghezzehei (2019). Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations, *Water Resources Research*, 55, 5715–5737.

Blöschl, G. (2017). Debates—Hypothesis testing in hydrology: Introduction. *Water Resources Research*, 53, 1767– 1769.

Clifton, D.A., L. Tarassenko, N. McGrogan, D. King, S. King, and P. Anuzis (2008). Bayesian extreme value statistics for novelty detection in gas-turbine engines. In: *Aerospace Conference*, IEEE, 1–11.

Dou, X. and Y. Yang (2018). Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems, *Computers and Electronics in Agriculture*, 148, 95–106

EESSD Strategic Plan (2018). Earth and Environmental Systems Sciences Division, Strategic Plan 2018–2023, DOE/SC–0192, May 2018.

Guggilam,S., S. M. Arshad Zaidi, V.Chandola, and A.Patra, Bayesian Anomaly Detection Using Extreme Value Theory (2019), arXiv:1905.12150v1.

Kratzert,F., D.Klotz, M.Herrnegger A.K. Sampson, S.Hochreiter, G.S.Nearing, Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, *Water Resources Research*, 55, 12, 2019

Lee, T.C.K., et al. (2005). Bayesian Climate Change Detection and Attribution Assessment, *Journal of Climate* 18(13):2429-2440

Nearing, G.S., Ruddell, B. L., Clark, M. P., Nijssen, B., Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, 19(11), 1835– 1852.

Vrugt, J.H., V.Gupta, S.C.Dekker, S.Sorooshian, T.Wagener, W.Bouten (2006). Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model, *Journal of Hydrology*, 325, 1–4, 288-307.

Zenil, H. (2020). Towards demystifying Shannon entropy, lossless compression and approaches to machine learning. *Proceedings*, MDPI, 47(1), 24; https://doi.org/10.3390/proceedings2020047024

Water Affordability and Equity: Re-Imaging Water Services – A Report from the 2020 Virtual Aspen-Nicholas Water Forum, 2020. https://www.aspeninstitute.org/wp-content/uploads/2020/12/Water-Forum-Consolidated-Report-2020.pdf