FEATURE

The Rise of Machine Learning in Hydrology

Xiang Li and John L. Nieber

SINCE 2015, ALPHAGO AND ITS SUCCESSOR PROGRAMS have defeated human Go professionals using artificial intelligence. AlphaGo was developed to test how well a neural network using deep learning can compete at the game Go and other board games, such as chess, without being taught the rules. The tremendous growth in AI, machine learning (ML), and big data has opened a new era, sometimes called the Fourth Industrial Revolution. From the targeting of more effective business advertisements to the greater accuracy of live captions on the media, ML has fundamentally changed the way we live and work. Now data scientists, computer scientists, and hydrologists are coming together to use ML to advance our understanding of hydrological systems.

Although the unreasonably effective predictive performance of ML models may make them appear mysterious to some, they are not unintelligible to practitioners. In the simplest terms, any applicable ML model can be broken down into three components: general model architecture, purpose-oriented loss function, and an optimization algorithm. (Technically, a linear regression is also an ML model.) These components can be customized and redesigned to address a specific problem. With appropriate modification, an ML algorithm can be transformed to solve a well-defined problem with data-rich scenarios even in specialized science and engineering domains. This generalizability is a blueprint for ML applications.

In the natural sciences, ML is already having an enormous impact. The increasing availability of large volumes of earth science data provides unprecedented opportunities for advancing the applicability of ML in hydrology. Data from remote-sensing satellites, stream gauges, field sensors, large-scale earth system simulations, and other sources, when integrated with ML techniques, hold great potential for modeling hydrological systems. Using a model similar to what Google Translator uses to translate a sentence from one language into another, ML can predict watershed discharge even more accurately than process-based hydrologic models in certain scenarios. Furthermore, such a model can also be used for hydrological regionalization, without the need to use catchment characteristics for gauged predictions.

> *It makes sense to draw on the forecasting ability of machine learning to solve hydrology problems while incorporating principles of hydrology.*

Using Hydrology Knowledge to Guide Machine Learning

ML models are quite distinct from traditional process-based hydrological models. While process-based models encode hydrological processes (such as infiltration and surface runoff) in mathematical representations formed within hydrological cycles, ML, which usually ignores the wealth of accumulated hydrologic knowledge, aims to leverage the capability of data to explain the hidden patterns in data. Process-based models do not fully leverage the information hidden in data since current hydrologic understanding might not comprehensively explain all interesting data patterns. To bridge this gap between ML and hydrology knowledge, a research avenue called knowledge-guided machine learning (KGML) has emerged, capturing the interest of both academia and industry, where hydrology is an important application field. In a nutshell, the information behind the big data in earth science can be transformed into hydrologic knowledge.

In 2020 the University of Minnesota held a three-day virtual workshop that brought together researchers from around the world to discuss the KGML framework. In the hydrology session, one of the presentations addressed how to use KGML to predict basin discharge by incorporating hydrologic knowledge into an ML model. This approach demonstrated some success at emulating the streamflow mechanism of the well-known hydrologic model SWAT.

In one small watershed in southeast Minnesota,

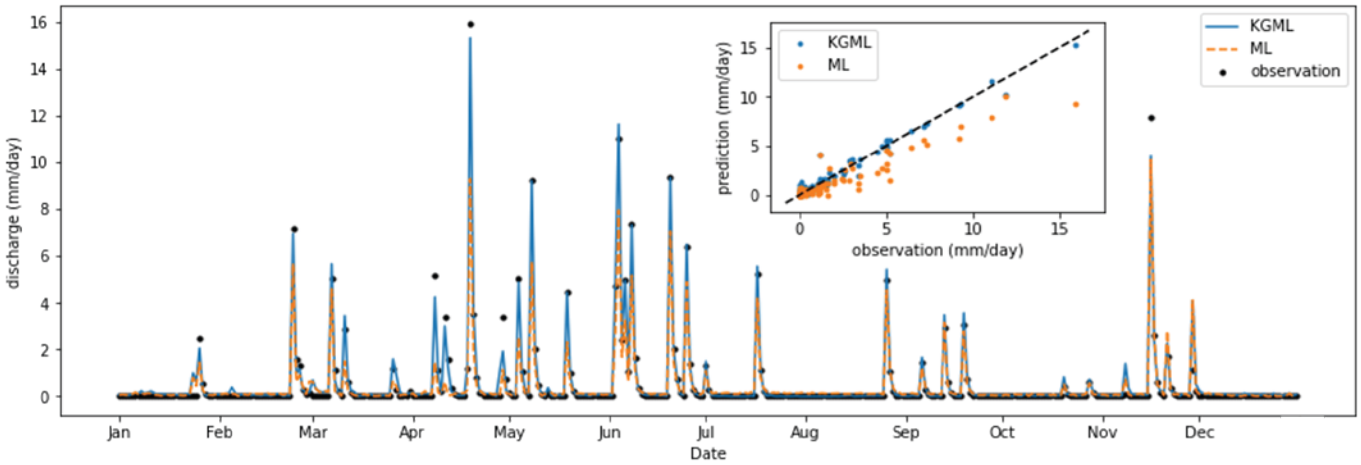

Figure 1. KGML emulation of the SWAT model in the South Branch of the Root River at Garden Meadow in southeastern Minnesota. KGML (solid blue line) performs better at streamflow prediction than does the pure ML model (dashed orange line). Observation = SWAT synthetic data. The figure shows performance comparison in one simulation year.

KGML satisfactorily emulated the SWAT-generated discharge when the ML model adopted concepts of hydrologic system memories, such as soil moisture and snow accumulation. As shown in Figure 1, KGML improves streamflow prediction compared with the case when no physics is included in the ML model. The inset scatter plot clearly shows that KGML predictions match the SWAT synthetic data more consistently. Through the whole testing period, the NSE (Nash-Sutcliffe Efficiency) score improves from 0.57 to 0.76 when implementing KGML.

In a similar workshop held in 2021, the development of KGML in hydrology was again discussed. One interesting hydrology application of KGML is to infer catchment characteristics from hydrologic responses. With the guidance of hydrology knowledge, ML infers noisy, uncertain, or sometimes hard-to-measure catchment characteristics from environmental drivers and streamflow data. Using these data, it was shown that KGML could reasonably estimate soil porosity, or even catchment elevation, slope, and other characteristics, with a 16% improvement in accuracy. This inference of catchment characteristics exhibits the capability of ML to improve hydrologic understanding when integrated with hydrologic knowledge.

The International Association of Hydrological Sciences (IAHS) proposed dedicating the decade 2013–2022 to "Prediction under Change." On the road to understanding the hydrologic principles that can help predict hydrology responses in changing environments, traditional process-based models have produced hydrologic discoveries for decades. Such models, however, do not perform well in all basins, implying a need to develop innovative tools to discover hydrologic knowledge. The development of big data in hydrology has led to an increasing trend to take advantage of the predictive power of ML in hydrology. Although ML models outperform process-based models in some instances, pure ML models will definitely not replace hydrological models. ML models rely heavily on data richness, and they are not as clearly interpretable as process-based models. It makes sense to draw on the forecasting ability of ML to solve hydrology problems while incorporating principles of hydrology (such as conservation of mass and conservation of energy).

A New Era in Modeling Hydrological Systems

Although still at an early stage, ML and KGML both exhibit remarkable potential in hydrology. Advances in scientific discovery and our understanding of complex hydrologic systems await help from these epoch-making, data-driven methods.

To move forward, KGML will require significant collaborative research among data scientists, computer scientists, and hydrologists. Research that couples process-based hydrological models with data-driven ML models will help shed light on complicated watershed systems. As interdisciplinary research efforts advance models of complex hydrological systems with assistance from ML, more definitive answers to questions raised by the theme "Prediction under Change" will likely emerge in the coming decade. ■

Xiang Li (lixx5000@umn.edu) is a Ph.D. candidate at the University of Minnesota. His research focuses on deep learning applications in hydrology and groundwater modeling. **John L. Nieber** (nieber@umn.edu) is a professor in the Department of Bioproducts and Biosystems Engineering at the University of Minnesota. He is a past president of the American Institute of Hydrology, a fellow at the University of Minnesota Institute on the Environment, and immediate past-chair of the Soil Physics and Hydrology division in the Soil Science Society of America.