EDITORIAL
published: 11 May 2021
doi: 10.3389/frwa.2021.681023

Editorial: Broadening the Use of Machine Learning in Hydrology

Chaopeng Shen[1]*, *Xingyuan Chen*[2] *and Eric Laloy*[3]

[1] *Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, United States,* [2] *Earth System Measurements & Data, Pacific Northwest National Laboratory, Richland, WA, United States,* [3] *Institute for Environment, Health and Safety, Belgian Nuclear Research Centre, Mol, Belgium*

Keywords: machine learning, hydrology, water cycle, integration, snow, groundwater, streamflow (runoff), precipitation

Editorial on the Research Topic

Broadening the Use of Machine Learning in Hydrology

The introduction of deep learning (DL) (LeCun et al., 2015) into hydrology around 2016–2018 (Tao et al., 2016; Laloy et al., 2017, 2018; Shen, 2018; Shen et al., 2018), especially the use of long short-term memory (LSTM) as a dynamical modeling tool for soil moisture and streamflow (Fang et al., 2017; Kratzert et al., 2019), has ignited a surge in machine learning applications across all domains of hydrology. At the core, machine learning is a set of tools that allow us to build and train models that extract and reproduce the spatial and temporal patterns in the datasets they encounter. In particular, the central philosophy of DL has been to minimize the intervention of the human experts in feature design and to facilitate maximal extraction of information from data (Goodfellow et al., 2016). Improved prediction quality in hydrologic machine learning (ML) models has been achieved not by infusing process-based assumptions into the models, but by conducting extensive training of the models with large quantities of a priori data. It has been argued by Nearing et al. (2020) that there could be significantly more information in large-scale hydrological data sets than hydrologists have been able to translate into theory or process-based models. The hydrology community is poised to fully explore the power in the vast amount of data using machine learning in various subdomains of hydrology.

In this Research Topic, we sought to broaden the use of machine learning (ML) in hydrology rather than emphasizing the depth of a specific topic. We sought applications of machine learning in both data-rich and data-scarce settings. We are highly encouraged to see the diversity and breadth covered by the resulting collection of published papers, which have almost covered the entire water cycle. A variety of machine learning techniques have been adapted to address various challenges existing in predicting the hydrologic cycle, ranging from a dynamical modeling tool to event localization, and from information extraction to a hypothesis generator. In the following section, we briefly go over some editor-identified highlights of the papers.

Precipitation, as *the beginning of the hydrologic cycle*, is a major source of uncertainty, and most satellite products are still too coarse for water management purposes, making precipitation downscaling a high-stakes activity. Sun and Tang employed an attention-based, deep convolutional neural network (AU-Net) to downscale coarse-resolution satellite-based precipitation data products to 1 km resolution (learning from gauge-based precipitation data products), with the help of auxiliary predictors including elevation, vegetation index, and air temperature. Novel to hydrology, authors employed an attention mechanism that extracts multiscale features by fusing gauged data. However, there are often missing values in gauged precipitation data due to various instrumentation and data quality issues. Mital et al. developed a new sequential imputation algorithm based on a Random Forest technique for interpolating the missing values in

OPEN ACCESS

Edited and reviewed by:
Harrie-Jan Hendricks Franssen, Julich-Forschungszentrum, Helmholtz-Verband Deutscher Forschungszentren (HZ), Germany

****Correspondence:***
Chaopeng Shen
cshen@engr.psu.edu

Specialty section:
This article was submitted to Water and Hydrocomplexity, a section of the journal Frontiers in Water

Received: *15 March 2021*
Accepted: *01 April 2021*
Published: *11 May 2021*

Citation:
Shen C, Chen X and Laloy E (2021) Editorial: Broadening the Use of Machine Learning in Hydrology. Front. Water 3:681023. doi: 10.3389/frwa.2021.681023

spatio-temporal daily precipitation records. They found that, for reliable imputation, having a few strongly correlated references is more effective than having a larger number of weakly correlated references.

Snow is an important precipitation component that is even more difficult to measure (*in-situ* or remotely) than rainfall. Meyal et al. wrote one of the first papers to simulate a snow water equivalent (SWE) using LSTM, leveraging climatic and SWE data from five Snow Telemetry (SNOTEL) stations. They reported Nash Sutcliffe efficiency coefficient (NSE) values ranging from 0.85 to 0.96. The authors build an automated prediction system with online data ingestion. This standard application demonstrated the plausibility of using LSTM for large-scale operational SWE modeling. With only five training sites, however, it remains to be seen if the model can be applied to larger scales.

Streamflow is an important and human-relevant component of the hydrologic cycle. Duan et al. employed a temporal convolutional neural network (TCNN), a one-dimensional dilated convolutional unit with sequential or causal connections, for long-term streamflow projection in California. By comparing the performance of TCNN against other machine learning approaches including the LSTM, Duan et al. not only showed that TCNN excelled at capturing high flows, but also qualitatively demonstrated that TCNN yielded physically plausible estimations of streamflow in responding to precipitation under future extreme climate scenarios beyond the historic records (e.g., under high temperature and quadrupled precipitation), showing that causal convolutions could enhance the stability of ML models when extrapolated outside of their trained conditions.

While still dealing with surface water, Oppel and Mewes present a slightly different application that used machine learning to localize events. They compared several machine learning approaches ranging from support vector machines to extreme learning machines to identify the beginning and end of multiple flood events along with their associated volumes from hydrographs. They also demonstrated that the ML methods afford additional benefits in facilitating the automation of the workflow, which can lead to increased scalability for practical operations.

With the groundwater of the hydrological cycle, Sahu et al. trained a Multilayer Perceptron (MLP) model to predict three-point observations of groundwater levels using temperature, precipitation, river discharge, and past groundwater data as inputs. The authors conducted a sensitivity analysis of features' importance and observed that providing all available inputs to their MLP models was not necessarily the optimal choice. They also found that MLPs trained solely on temperature and historical groundwater level measurements as features were unreliable at all locations, which alluded to the dynamical linkage between surface hydrology and groundwater. Future sensitivity analysis will likely be accompanied by uncertainty estimates to ensure the robustness of the analysis. We also note more effort should be focused on finding ways to generalize these types of models outside of locations with data included in the training set. Groundwater flow problems, due to their lack of observation, the three-dimensional nature of the problem, and strong heterogeneity, are difficult to formulate into uniform learnable problems.

Diving deep into the subsurface environment, Generative Adversarial Networks (GAN) are becoming an alternative to Multiple-point Statistics (MPS) techniques to generate stochastic subsurface fields from training images. An open issue for all the training image-based simulation techniques (including GAN and MPS) is to generate consistent 3D field realizations when only 2D training data sets are available. This is especially relevant to groundwater hydrology for which it is difficult, if not impossible, to collect exhaustive and accurate data about the 3D subsurface distribution of rock types (or physical properties). Coiffier et al. introduced a novel approach termed Dimension Augmenter GAN (DiAGAN) that enables GANs to generate 3D fields from 2D examples. The method is simple to implement as it introduces a random cut sampling step between the generator and the discriminator of a standard GAN. Numerical experiments show that for complex binary subsurface media, the proposed approach is efficient and provides results of similar quality as those obtained by a state-of-the-art MPS method.

Around the world, many aspects of urban water systems, e.g., water supply, discharge, and stormwater management, require upgrades to adapt to the challenges of global change and urban growth. We expect there will be a substantial surge in applications of ML in urban water systems to improve their efficiency and transform them into smart cities. Allen-Dumas et al. wrote a thorough review that synthesized ways in which ML techniques have been applied to different parts of the urban water system in order to address multiple water hazards. They discussed ML applications in monitoring, early warning, prediction of urban water hazards (floods, drought, water contamination, soil erosion, and sediment transport), multi-hazard risks (compound risks), selection of best management practices, etc. They argued that by weaving together multiple ML methods for different risks, we can eventually arrive at a comprehensive watershed-to-community planning workflow for smart-city management of urban water resources.

In agreement with the general trend in the field of hydrology, the abovementioned papers have covered most components of the hydrologic cycle. Outside of this Research Topic, machine learning has been applied to soil moisture (Fang et al., 2019), soil data extraction (Chaney et al., 2019), hydrology-influenced water quality variables including in-stream water temperature (Rahmani et al., 2020) and dissolved oxygen (Zhi et al., 2021), human water management through reservoirs (Yang et al., 2019; Ouyang et al., 2021), subsurface reactive transport (Laloy and Jacques, 2019; He et al., 2020), and vadose zone hydrology (Bandai and Ghezzehei, 2021), among others. ML is not only applicable in data-rich regions but can also be leveraged by data-scarce regions (Feng et al., 2021; Ma et al., 2021). DL-native methods for uncertainty quantification have also emerged (Zhu et al., 2019; Fang et al., 2020). What is still missing to date includes vegetation hydraulics, glaciers, preferential flow, hyporheic exchange, and regional groundwater recharge, though this list is incomplete. We believe these components will be covered by machine learning approaches in the future.

While the broadening of ML has been, to some extent, achieved, one can also notice some limitations and unrealized potential. First, most of the abovementioned use cases are siloed to one variable, e.g., streamflow or precipitation. Second, many of the presented examples are built on small datasets, which means that instead of having learned universally-applicable physical laws, they were locally-fitted models based on the measurement sites in question. The implications of these limitations are that the models are not transferable outside the training region, their potential prediction failures are not yet sufficiently tested, and the information from one observed variable cannot influence the other variables.

There are many angles from which one can overcome the limitations. From a purely data-driven perspective, multi-task learning could allow multiple variables to interact and inform each other. A multiphysics land surface model can be trained to simultaneously predict multiple physical variables in the context of multi-task learning, which is known to improve all tasks. This is because many tasks can use shared representations and are thus constrained by multiple targets at the same time (Caruana, 1997). Alternatively, one may seek to organically tie in physical processes with machine learning, allowing known physical laws such as the mass balance and the law of flow to serve as the connective tissue between different model components. While there is a substantial amount of effort in the direction of knowledge-guided machine learning (Read et al., 2019), there are certainly many different paths toward the goal of integrating physics with machine learning. Outside of this Research Topic, there are methods for parameter learning (Tsai et al., 2020a) and physics-informed neural networks (He et al., 2020; Tartakovsky et al., 2020).

One of such pathways, perhaps a niche one, was documented in (Tsai et al., 2020b). This paper used machine learning to generate articulable hypotheses about which physical factor between soil texture, soil thickness, and slope *caused* water storage and streamflow to be linked in a certain way in a basin, and tested them using a physically-based model. While machine learning is very powerful, due to data limitations and factor covariation, it often cannot distinguish between causal or associative relationships, and what it found are therefore merely *hypotheses*. To test these competing hypotheses, Tsai et al. configured a physically-based hydrology model, PAWS+CLM (Shen and Phanikumar, 2010; Shen et al., 2013; Niu et al., 2017; Ji et al., 2019) to represent these hypotheses, e.g., they increased soil thickness or changed soil texture in one of the synthetic simulations and checked if the storage-streamflow relationships changed in agreement with the hypothesized effect as a result. The outcome of the process-based model can in fact be merged with the machine learning hypotheses in a Bayesian and algorithmic way, which implies this avenue can in fact be autonomously executed. While this paradigm is not expected to become popular any time soon, it does suggest physical models provide unique information that can fill in the gaps (in this case, assessment for a causal relationship) for machine learning methods.

Multiple pathways exist for ML to help to make advances in hydrology: (1) incorporating physics in ML models; (2) improving the interpretability of ML models; (3) developing coupled, physics-informed neural networks; (4) quantifying and propagating uncertainty in model results; (5) developing publicly available benchmark training data sets that can be used to aid and test new ML methods; and (6) building a community computational platform to allow sharing of ML pipelines with easy access to pre-trained ML models (e.g., similar to Model Zoo, https://modelzoo.co/), standardized application-ready datasets, interoperable process-based models, and supercomputing and/or cloud computing resources. Generating public benchmark training data sets (similar to ImageNet, http://www.image-net.org/) that researchers can use to build better ML models is the key to advancing applications of ML in Earth science domains (Dramsch, 2020; Maskey et al., 2020). There is a unique opportunity here to enhance the use of the new generation of remote sensing products that capture components of the water cycle (precipitation, snow, soil moisture, evapotranspiration, groundwater, and runoff), as well as coupled carbon and nutrient cycle components, with increasing spatial and temporal resolutions. Training data may also be generated from process-based models. Leveraging open-source resources from federal agencies is necessary for the success of such extensive and expensive effort. For example, NASA's Earth Sciences Data Systems (ESDS) have generated high-quality training data sets that are open and easily accessible. NOAA, USGS, and other federal agencies have been maintaining extensive observation networks and are developing a large number of integrated Earth system models. Standardized data management practices would significantly increase data usability, and we call for significant investment to support community efforts that address these challenges.

AUTHOR CONTRIBUTIONS

CS, XC, and EL edited the Research Topic and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

CS was sponsored by the U.S. National Science Foundation Award OAC #1940190. XC was supported by the ExaSheds project funded by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Government.

ACKNOWLEDGMENTS

As editors, we thank all authors and reviewers who contributed to this Research Topic. We also thank Frontiers for facilitating this process.

REFERENCES

Bandai, T., and Ghezzehei, T. A. (2021). Physics-informed neural networks with monotonicity constraints for richardson-richards equation: estimation of constitutive relationships and soil water flux density from volumetric water content measurements. *Water Resour. Res*. 57:e2020WR027642. doi: 10.1029/2020wr027642

Caruana, R. (1997). Multitask learning. *Mach. Learn*. 28, 41–75. https://doi.org/10/d3gsgj

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., et al. (2019). POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous united states. *Water Resour. Res.* 55, 2916–2938. https://doi.org/10/ggj68b

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *arXiv:2006.13311*. 61, 1–55. doi: 10.1016/bs.agph.2020.08.002

Fang, K., Kifer, D., Lawson, K., and Shen, C. (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resour. Res*. 56:e2020WR028095. doi: 10.1029/2020wr028095

Fang, K., Pan, M., and Shen, C. (2019). The value of SMAP for long-term soil moisture estimation with the help of deep learning. *IEEE Trans. Geosci. Remote Sensing*. 57, 2221–2233. https://doi.org/10/gghp3v

Fang, K., Shen, C., Kifer, D., and Yang, X. (2017). Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophys. Res. Lett*. 44, 11030–11039. https://doi.org/10/gcr7mq

Feng, D., Lawson, K., and Shen, C. (2021). Prediction in ungauged regions with sparse flow duration curves and input-selection ensemble modeling. *arXiv*. http://arxiv.org/abs/2011.13380

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press. Available online at: https://www.deeplearningbook.org/

He, Q., Barajas-Solano, D., Tartakovsky, G., and Tartakovsky, A. M. (2020). Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. *Adv. Water Resour*. 141:103610. doi: 10.1016/j.advwatres.2020.103610

Ji, X., Lesack, L. F. W., Melack, J. M., Wang, S., Riley, W. J., and Shen, C. (2019). Seasonal and interannual patterns and controls of hydrological fluxes in an amazon floodplain lake with a surface-subsurface process model. *Water Resour. Res*. 55, 3056–3075. https://doi.org/10/gghp4s

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci*. 23, 5089–5110. https://doi.org/10/gghmz4

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resour. Res*. 54, 381–406. https://doi.org/10/gdbxmz

Laloy, E., Hérault, R., Lee, J., Jacques, D., and Linde, N. (2017). Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Adv. Water Resour*. 110, 387–405. https://doi.org/10/gcqftj

Laloy, E., and Jacques, D. (2019). Emulation of CPU-demanding reactive transport models: a comparison of gaussian processes, polynomial chaos expansion, and deep neural networks. *Comput. Geosci*. 23, 1193–1215. doi: 10.1007/s10596-019-09875-y

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*. 521, 436–444. https://doi.org/10/bmqp

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resour. Res*. e2020WR028600. doi: 10.1029/2020wr028600

Maskey, M., Alemohammad, H., Murphy, K. J., and Ramachandran, R. (2020). Advancing AI for earth science: a data systems perspective. *Eos* 101. doi: 10.1029/2020EO151245

Nearing, G., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J., et al. (2020). What role does hydrological science play in the age of machine learning? *arXiv* doi: 10.31223/osf.io/3sx6g

Niu, J., Shen, C., Chambers, J., Melack, J. M., and Riley, W. J. (2017). Interannual variation in hydrologic budgets in an Amazonian watershed with a coupled subsurface—Land surface process model. *J. Hydromet*. 18, 2597–2617. https://doi.org/10/gcrcf8

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., and Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: a demonstration of effectiveness and a delineation of challenges. *arXiv*. arxiv:2101.04423

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C. (2020). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett*. 16:024025. doi: 10.1088/1748-9326/abd501

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resour. Res*. 55, 9173–9190. doi: 10.1029/2019wr024922

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res*. 54, 8558–8593. https://doi.org/10/gd8cqb

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS Opinions: incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci*. 22, 5639–5656. doi: 10.5194/hess-22-5639-2018

Shen, C., Niu, J., and Phanikumar, M. S. (2013). Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface—Land surface processes model. *Water Resour. Res*. 49, 2552–2572. https://doi.org/10/f5gcrx

Shen, C., and Phanikumar, M. S. (2010). A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resour*. 33, 1524–1541. https://doi.org/10/c4r8k5

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., and Ihler, A. (2016). A deep neural network modeling framework to reduce bias in satellite precipitation products. *J. Hydromet*. 17, 931–945. https://doi.org/10/ggj7gh

Tartakovsky, A. M., Marrero, C. O., Perdikaris, P., Tartakovsky, G. D., and Barajas-Solano, D. (2020). Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. *Water Resour. Res*. 56:e2019WR026731. doi: 10.1029/2019wr026731

Tsai, W.-P., Pan, M., Lawson, K., Liu, J., Feng, D., and Shen, C. (2020a). From parameter calibration to parameter learning: revolutionizing large-scale geoscientific modeling with big data. *arXiv:2007.15751.* http://arxiv.org/abs/2007.15751

Tsai, W. P., Fang, K., Ji, X., Lawson, K., and Shen, C. (2020b). Revealing causal controls of storage-streamflow relationships with a data-centric bayesian framework combining machine learning and process-based modeling. *Front. Water* 2:583000. doi: 10.3389/frwa.2020.583000

Yang, S., Yang, D., Chen, J., and Zhao, B. (2019). Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model. *J. Hydrol.* 579:124229. https://doi.org/10/ggj668

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., et al. (2021). From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol*. 55, 2357–2368. doi: 10.1021/acs.est.0c06783

Zhu, Y., Zabaras, N., Koutsourelakis, P.-S., and Perdikaris, P. (2019). Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comp. Phys*. 394, 56–81. https://doi.org/10/ggddhn

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.