FAIR data infrastructure and tools for AI-assisted streamflow prediction

Authors

Line Pouchard, Carlos Soto (BNL CSI), Marcia Branstetter, Giri Prakash (ORNL, ARM)

Focal Area(s) Areas

We discuss how the integration of AI into Earth Science models can impact streamflow predictions at both the science and data levels. Doing so, we address cross-cutting needs related to the goal of making data FAIR (Findable, Accessible, Interoperable, and Re-usable [1]) for seamless use with Artificial Intelligence/Machine Learning (AI/ML) in Earth System Science at DOE. A novel idea is that AI/ML itself can help with the FAIR data goal and address issues in targeted areas e.g. missing data, data quality and reduction. In addition, the interpretability of results obtained with new AI methods is poised to impact broader scientific challenges in hydrology..

Science Challenge

Our science question involves a streamflow scenario that uses AI/ML to predict discharge as a function of precipitation. This type of simulation for a watershed is important for flood risk management, ultimately impacting flood forecasting and the operation of power plants. Predicting and validating multi-decadal trends for a streamflow in a watershed as it impacts extreme events uses a rainfall-runoff model. This model requires precipitation data and land cover information that can be obtained from the E3SM coupled land-surface model ELM and observational data. The future AMF2 mobile field campaign Surface Atmosphere Integrated Field Laboratory (SAIL) at Crested Butte, Colorado in 2021 will collect atmospheric measurements to be used with existing surface data. Data from SAIL including precipitation rate, solar radiation, relative humidity, wind speed can be used in our scenario. Data from the East River Watershed Function SFA managed by LBL is also needed to provide precipitation, snow fall, and river flow. An AI model can predict the streamflow to compare predictions with physics-based results. AI can be further used to improve input data, both observational and computational, thus increasing FAIR-ness. Our scenario exemplifies many issues encountered when using AI for predictions, issues that are applicable to many scientific challenges using AI.

Rationale

Challenges related to data quality and data volume: AI research involving data from disparate sources requires significant effort by the PI, including analyzing the measurements' data quality and removing data points based on various data quality reports. Enabling AI at the data source to carry out the data quality analysis on-the-fly will help reduce the volume of unfiltered data to the users and lower the post-processing time by the PIs. Currently, data quality information is communicated in a variety of ways by the data sources and repositories. As an example, ARM provides data quality flags and data quality reports as an ancillary data file. In comparison, data quality information for Soil moisture and water flow data distributed by the NASA DAAC and USGS Water Resource Program are captured either in their geospatial maps or metadata files. Understanding and using such a variety of data quality information is a complicated and time-consuming activity for the PIs.

DOI: https://doi.org/10.2172/1769710

Challenges related to rainfall-runoff models: When using traditional rainfall-runoff models, streamflow predictions are hampered by the need to calibrate each individual watershed. In recent years, this burden has been alleviated by data-driven AI/ML techniques which have been shown to produce accurate streamflow as a function of precipitation events (rainfall-runoff processes). Li, et al. (2020), for example, used an LSTM (Long Short-Term Memory network [2]) to predict river discharge in a Houston watershed, and found this model to be more efficient than a process-based model. This and other AI techniques have also shown promise in characterizing whole regions by calibrating parameters in several catchments, and enabling streamflow predictions at scale to be obtained from many watersheds simultaneously.

There remain, however, important data, algorithmic, and usability challenges. Data scarcity (e.g., for individual catchments) and data quality, in particular, must be taken into account in AI model design, and concerted AI interpretability efforts must be made to avoid black box model outputs. Without such interpretability, it becomes very challenging to meaningfully relate source data to model results, i.e., relating cell data to predicted hydrological patterns. More broadly, for any appropriate use of AI in such streamflow scenarios, these data must be findable, properly described, and of sufficient quality to be used in AI models. Furthermore, scientists will not necessarily trust black box model results and require explanations for AI-obtained predictions, and synthetic data obtained with AI. AI/ML predictions for scientific purposes will not be trusted without compliance with FAIR data principles and robust AI explanations.

Narrative

AI/ML can have a transformational impact on scientific discovery by making inferences at scale for entire regions and watersheds, by explaining itself, and by improving data quality. Figure 1 shows the applications of AI to observations and model predictions that we address below. First, new developments in Deep Learning techniques such as Transformer architectures and variational auto-encoders can improve streamflow predictions. Such techniques also have the advantage of contributing to the interpretability of the results that will help inspire more confidence in the predictions themselves. Second, AI can improve Findability by mitigating the effects of missing data, both for traditional models and data-driven models. And third, from the data management perspective, AI can improve quality for raw data transferred from instruments to archives such as ARM, thus help alleviate the burden of manual curation.

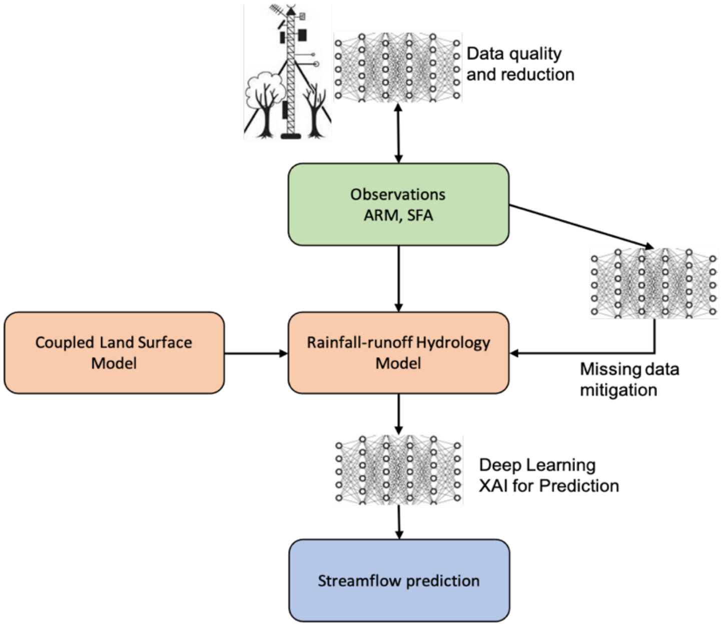

Figure 1: Applying AI to observations and predictions

Improving prediction accuracy and explainability with Transformers: Data-driven ML models have been successfully shown to produce hydrology predictions normally made with process-driven models. LSTM models in particular have excelled at capturing delayed reactions and data relationships, such as snowfall to snowmelt. However, because of their recurrent nature, LSTMs still lose representational capacity for such long-term data relationships at each iteration. Further, LSTMs suffer from the same black-box characterizations that burden the interpretability of many AI techniques.

LSTMs have been recently superseded by a new AI architecture called the Transformer [3]. This new method has improved performance relative to all prior ML models at processing sequential data. Transformers are non-recurrent and so can parallelize within input sequences, allowing for much faster training and inference, as well as supporting much larger and deeper networks which can capture more detailed and nuanced data properties. They also employ a novel self-attention mechanism which enables the same long-term dependency modeling which made LSTMs such a success. However, self-attention gives Transformers direct access to deep representations of all previous data points, and so they outperform LSTMs at capturing time-delayed relationships. Also, the self-attention mechanism presents an opportunity to interpret Transformer model results in a way not possible with alternatives such as LSTMs, by directly assigning the relative importance of each data point to each intermediate representation and to final predictions.

Using AI methods to mitigate the effect of missing data values: Missing observational data is a fundamental challenge to any hydrology modeling and predictive efforts. Given the diversity of instruments and campaigns that potentially record data related to precipitation - e.g. radar, aerosol for rain, snow fall, particles - finding values for missing data in existing archives for use in a model can be very time-consuming. Missing data parameters (e.g. in ARM, Ameriflux or Fluxnet data) to be input into a streamflow model can be predicted with a variety of AI methods, including supervised, self-supervised and unsupervised approaches. These ML methods learn how to fill missing values, capturing covariances between data features, rather than e.g. simple mean or median replacements. Multi-layer perceptrons (MLPs) and k-Nearest Neighbors (kNN) are two common approaches, which may be used in supervised or self-supervised manners. Unsupervised methods such as clustering may also be applied. In all cases, AI/ML approaches produce predicted replacement values for missing data points which incorporate knowledge derived from data relationships in the entire dataset.

De-noising radar data: AI for data quality and reduction

Another important data challenge for hydrology modeling is noisy and/or large source data. Radar data, for example, can benefit from denoising and reduction before archiving in databases. Data denoising and reduction are related problems in that both seek to separate and preserve only the valuable parts of a data source, while discarding noise and/or bulk. Generalizable AI/ML approaches such as Autoencoders [4,5] and Generative Adversarial Networks (GANs [6]) can be adapted for both of these tasks. Autoencoders learn compact data representations, and can be trained specifically to reduce noise while enabling faithful data reproduction, whereas GANs are generative models which can be trained to produce denoised or reduced output conditioned on a particular input. AI-driven data denoising and reduction can be applied post hoc on data from existing sensors, or in a co-design approach with smart sensors. ML models can be trained and adapted to be run on embedded computational resources in novel smart sensors, and multiple models may also coexist and be selected dynamically depending on sensor or data conditions.

References

[1]Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018.

[2] Li, W., Kiaghadi, A. & Dawson, C. High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. Neural Comput & Applic (2020). https://doi.org/10.1007/s00521-020-05010-6

[3] Vaswani, Ashish, et al. "Attention is All you Need." NIPS. 2017.

[4] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[5] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." Journal of machine learning research 11.12 (2010).

[6] Goodfellow, Ian J., et al. "Generative adversarial networks." arXiv preprint arXiv:1406.2661 (2014).