A Bayesian Neural Network Ensemble Approach for Improving Large-Scale Streamflow Predictability

Dan Lu (lud1@ornl.gov)[1], Daniel Ricciuto[1], Marcia L. Branstetter[1], Eric Pierce[1], Seung-Hwan Lim[1], and David Womble[1]

[1]*Oak Ridge National Laboratory, Oak Ridge, TN*

Focal Area

(2) Predictive modeling through the use of AI techniques and AI-derived model components; the use of AI and other tools to design a prediction system comprising of a hierarchy of models (e.g., AI driven model/component/parameterization selection). We propose a Bayesian neural network ensemble approach for improving large-scale streamflow predictability and understanding in a changing environment that combines multiple Earth system land model predictions by calculating spatiotemporally varying model weights and biases while accounting for various types of observations at multiple scales with uncertainty.

Science Challenge

A better predictive understanding of streamflow changes in large river basins over continental to global scales is a critical need for addressing scientific and societal challenges. Streamflow affects and is affected by many processes at the land-atmosphere interface, influences ecosystem structure and functioning, and provides freshwater for the society. Improving large-scale streamflow predictability is a complex problem and contingent on improving understanding of large-scale fluxes of water, energy, and biogeochemical cycles. Current Earth system land model (ELM) simulations of streamflow have a significant uncertainty among models and at different locations. Key challenges are (1) how to effectively leverage each ELM's spatiotemporally varying predictive skill, (2) how to efficiently incorporate a variety of observations at multiple scales to constrain the models (e.g., observations from the ARM facility and NGEEs and SFA observatories), and (3) how to improve the representation of hydrologic processes in ELMs (e.g., the Energy Exascale Earth System Model [E3SM]). These challenges must be addressed to improve large-scale streamflow predictability, and they will be addressed here by using advanced AI and uncertainty quantification methods combined with high performance computing and edge computing.

Rationale

The ability to accurately simulate streamflow at large river basin, continent, and global scales is crucial to understanding changes in the hydrological cycle and water availability. Many ELMs have been applied for large-scale streamflow simulations. They have large uncertainties in hydrologic process representations, show varying prediction skills at different locations and times, and are usually not constrained by observations. Improving streamflow prediction requires a comprehensive multi-model ensemble approach that leverages each individual ELM's spatiotemporally varying predictive skill and integrates various types of observations at multiple scales to reduce uncertainty using a calibration framework [1]. Existing ensemble approaches usually assume model independence and model democracy in which each model is weighted equally, but neither of these assumptions is true. Many ELMs in Coupled Model Intercomparison Projects (CMIPs) share components or are variants of another model in the ensemble, and these models have large inconsistency in their skill at a given location and time. Even an individual ELM shows considerably inconsistent skills at different locations and times. Additionally, with the

DOI: https://doi.org/10.2172/1769641
DOE BER, AI4ESP 2021 #AI4ESP1081 (https://www.ai4esp.org)

increase of resolution and complexity in ELMs, the models are usually uncalibrated and uniformly weighted across space and time, and the same weights are applied for future projection. This may result in an inaccurate and overconfident streamflow prediction and thus misguide water management strategies.

We propose a novel Bayesian neural network (BNN) ensemble approach to improve large-scale streamflow predictability. This approach leverages AI power in data analytics and predictive analytics to calculate spatiotemporally varying model weights and biases by taking advantage of ELMs' diverse performance in heterogeneous catchments and different seasons. The approach will use a variety of observations to constrain the ensemble prediction while considering heteroscedastic data uncertainty. It will also consider extrapolation uncertainty as we project to the future climate. Additionally, the proposed approach will provide insights of the model and the data to inform the model and data development.

Narrative

We assume that observations $y(\boldsymbol{x}, t)$ at a given location $\boldsymbol{x}$ and time t can be modeled as a sum over an ensemble of m ELM predictions $M_i(\boldsymbol{x}, t)$ weighted by their respective weights $\omega_i(\boldsymbol{x}, t)$, a bias term $\beta(\boldsymbol{x}, t)$, and a data noise term $\sigma(\boldsymbol{x}, t)$. The model ensemble is designed to capture a large range of structural and parametric uncertainty for hydrologic processes related to streamflow. The proposed BNN will optimize the model weights, biases, and data noises simultaneously as probabilistic functions against observations to provide accurate and uncertainty aware predictions of the streamflow (Figure 1).

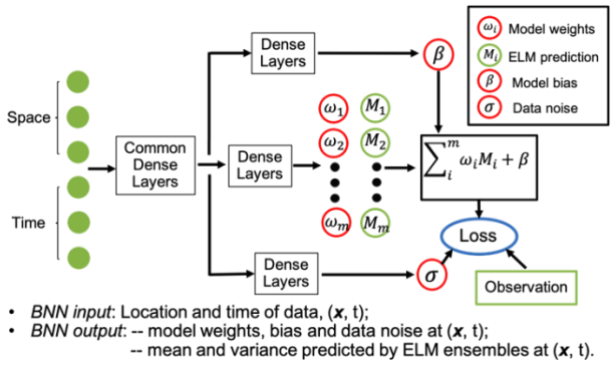

Figure 1. Schematic of the proposed BNN ensemble approach which combines multiple ELM predictions to improve large-scale streamflow predictability. The approach calculates spatiotemporally varying model weights and biases, use observations to constrain ensemble predictions with considering data uncertainty, as well as provide insights of the model and the data to inform the model and observation development.

BNN Ensemble Leverages Domain Knowledge and Observations for Streamflow Prediction

A BNN can learn a complex relationship between inputs and outputs and estimate the output distribution using Bayesian inference [2],[3]. The proposed BNN reads the data location ($\boldsymbol{x}$) and time (t) as inputs and estimates the model weights, biases, and data noises at the given $(\boldsymbol{x}, t)$. It first uses a set of dense layers to extract common information for the output learning, and then it designs three dense layers to learn the information specific to the model weights, biases, and data noises, respectively, according to their physical constraints and by leveraging domain knowledge. Next, the BNN incorporates the multiple ELM predictions $M_i(\boldsymbol{x}, t)$ and combines them with the calculated model weights, biases, and data noises in the loss function to compare the ensemble prediction with observations. After training, the BNN will produce the weighted ensemble mean as the streamflow prediction and a variance quantifying predictive uncertainty. We will incorporate

our physical and historical data understanding into the BNN design to generate weights and biases reflecting streamflow annual periodicity and the influence of extreme climate events on the streamflow. The BNN design will also consider that model skills and biases are likely to vary over the typical length scales spanned by climatic or geographic regions.

BNN Ensemble can Improve Predictability of Streamflow with Uncertainty Quantification

The proposed BNN ensemble approach calculates spatiotemporally varying model weights, biases, and data noises, which leverages each individual ELM's skill in space and time and also constrains the ensemble prediction using observations with considering data quality. This will result in a data-validated prediction and also reduce the predictive uncertainty caused by the diverse ELMs. The produced prediction variance not only quantifies data uncertainty but also calculates epistemic uncertainty due to the limited data and knowledge. This enables reasonable predictability quantification and prevents overconfident prediction in the changing environment. Additionally, the proposed approach is designed under the Bayesian framework, so it is robust to data noise and small data set and can encode domain knowledge into the prior to ensure a realistic prediction for regions with sparse and poor data.

BNN Ensemble can Advance Data and Model Understanding about the Streamflow Prediction

The proposed BNN ensemble approach can provide interpretability about which models with which particular processes contribute more to the ensemble prediction at which locations and times. This insight of detailed model analysis will advance our understanding of the ELMs across the space and time and inform the model development (e.g., implement the identified key processes in E3SM). Besides analyzing streamflow, we will also analyze other water availability–related quantities such as soil moisture and evaporation to investigate whether the same model gives consistent good prediction of these variables and explore the reasons to further improve our understanding. Additionally, we will use explainable AI techniques to analyze which data features have more important impact on which individual ELM's skill at which regions and seasons. The insights from data analysis will inform the best observations for the streamflow and other water availability characteristics and help observations campaign [4] (e.g., guide the placement of new ARM sites).

Models and Data

We will consider both process-based ELMs (including ELMs in CMIP and the E3SM) and data-driven machine learning models (such as long short-term memory networks trained from meteorological forcing observations to predict streamflow [5],[6]), and incorporate both point observations from the ARM facility (e.g., meteorological forcing data), the NGEEs and SFA observatories (e.g., soil property, streamflow), USGS gauge stations (e.g., streamflow), and large-scale satellite and radar products from NASA and NOAA (e.g., soil moisture map). The model simulation and analysis results and used observation data and associated metadata will have a detailed data management to ensure they are findable, accessible, interoperable, and reusable. We will use randomized maximum a posteriori sampling for the BNN training, which can leverage the exascale computing for computational efficiency. We will also use transfer learning to quickly integrate new observation streams by leveraging edge computing to refine the BNN, update the results, and thus improve the prediction. Code products from this project will be open-sourced, tested on various hardware, and reusable to address other Earth system problems.

Abbreviation

ARM: Atmospheric Radiation Measurement

NGEE: Next Generation Ecosystem Experiments

NASA: National Aeronautics and Space Administration

NOAA: National Oceanic and Atmospheric Administration

SFA: Science Focus Area

USGS: United States Geological Survey

References

[1] Lu, D., and Ricciuto, D. 2019. Learning-based inversion-free model-data integration to advance ecosystem model prediction. DOI 10.1109/ICDMW.2019.00049.

[2] Lu, D., and Ricciuto, D. 2019. Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques, *Geosci Model Dev*, 12(5), 1791-1807.

[3] Lu, D., Liu, S., and Ricciuto, D. 2019. An efficient Bayesian method for advancing the application of deep learning in earth science. DOI 10.1109/ICDMW.2019.00048.

[4] Lu, D., Ricciuto, D., and Evans, K. 2018. Efficient Bayesian data-worth analysis using a multilevel Monte Carlo method, *Advances in Water Resources*, 113, 223-235.

[5] Konapala, G., Kao, S., Painter, S. L., and Lu, D. 2020. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environmental Research Letters, 15(10).

[6] Lu, D., Konapala, G., Painter, S., and Kao, S. In review. Streamflow simulation in data-scarce basins using Bayesian and physics-informed machine learning models, *Journal of Hydrometeorology*.