Machine Learning Model Prediction of Streamflow in Walker Branch Watershed

Dan Lu,[1,*] Natalie Griffiths,[2] and Eric M. Pierce[2]

[1]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN
[2]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

(lud1@ornl.gov)

Project Lead Principal Investigator (PI): Eric M. Pierce

BER Program: ESS

Project: Critical Interfaces Science Focus Area (Oak Ridge National Laboratory)

Project Website: https://www.esd.ornl.gov/programs/rsfa/

Project Abstract: How will the hydro-biogeochemical function of watersheds respond to hydrologic intensification and land use/land cover change from local to regional scales? Addressing this question is crucial but also challenging. It requires large data, comprehensive model representation, and sophisticated data-model integration. Although a broad collection of diverse observations is increasing, current hydro-biogeochemical models have inadequate watershed process representation, and existing data assimilation methods are not powerful enough to incorporate diverse data for prediction.

Here we highlight the use of machine learning (ML) methods for hydro-biogeochemical simulations. ML models are powerful in extracting patterns, discovering new knowledge from multi-scale, multi-types of data, and identifying underlying cause-effect relationships for predictive understanding. We applied a ML approach, called Long Short-Term Memory (LSTM) network, to understand rainfall-runoff processes using data collected in Walker Branch Watershed (WBW). WBW is a 97.5-ha, temperate deciduous forest watershed located in East Tennessee, USA. Hydrological, biogeochemical, and ecological process studies have been carried out in this seminal research catchment since 1967. The multiple, long-term datasets that have been collected in WBW provide the opportunity to apply ML methods to predict hydrological dynamics. These long-term datasets include daily climate and soil temperature (1993-2010), precipitation (1969-2012), streamflow (1969-2014), and stream chemistry (1989-2013). We use LSTM to simulate streamflow based on meteorological and environmental observations including precipitation, air temperature, relative humidity, and soil temperature. A total of 14 years of daily data (1993-2006) were used, with the first 10 years for network training and the remaining 4 years for out-of-sample prediction. Results indicate that LSTM performed well in the training period where its predictions closely matched observed streamflow with the Nash-Sutcliffe efficiency of 0.97, although the predictions were relatively poor in 2004 and 2006. After hypothesis testing, we found that the poor performance in these two hot and dry years resulted from the lack of ET-related observations that were not included in the training and thus the relevant processes were not learned.

Here we used an interpretable LSTM which not only calculates variable importance but can also guide data collection and model development. We are working on ML model uncertainty quantification and will use the uncertainty bounds to indicate when the ML results can be trusted in the projection of future streamflow. Additionally, we are developing hybrid ML models to sufficiently leverage the data information and our domain knowledge for improving predictions.