APPLICATION OF MACHINE LEARNING AND AI FOR PREDICTION IN UNGAUGED BASINS (PUB)

Pin-Ching Li

ABSTRACT

Streamflow prediction in ungauged basins (PUB) generates streamflow time series at ungauged reaches in a river network, facilitating various engineering tasks such as managing stormwater, water resources, and water-related environmental impacts. Machine Learning (ML) has emerged as a powerful tool for PUB by applying their generalization process to capture the streamflow generation processes from hydrological datasets (observations). The generalization process of ML involves splitting the observations into training and testing datasets, training ML models using training data, and evaluating the performance of ML models using testing data. To unveil the potential limitations of ML's generalization process, this dissertation explores its robustness and associated uncertainty. More precisely, this dissertation has three objectives: (1) analyze the potential uncertainty caused by the data split process for ML modeling, (2) investigate the improvement of ML models' performance by incorporating hydrological processes within their architectures, and (3) identify the potential biases in ML's generalization regarding the trend and periodicity of streamflow simulations.

The first objective of this dissertation is to assess the sensitivity and uncertainty caused by the regular data split process for ML modeling. The regular data split process in ML was initially designed for homogeneous and stationary datasets, but it may not be suitable for hydrological datasets in the context of PUB studies. Hydrological datasets usually consist of data collected from diverse watersheds with distinct streamflow generation regimes influenced by varying meteorological forcing and watershed characteristics. To address the potential inconsistency in the data split process, multiple data split scenarios are generated using the Monte Carlo method. The scenario with inconsistent data split results accounts for the covariate shift and tends to add uncertainty and biases to ML's generalization process. The findings in this objective suggest the importance of avoiding the covariate shift during the data split process when developing ML models for PUB to enhance the robustness and reliability of ML's performance.

The second objective of this dissertation is to investigate the improvement of ML models' performance brought by Physics-Guided Architecture (PGA), which incorporates ML with the rainfall abstraction process. PGA is a theory-guided machine learning framework integrating conceptual tutors (CTs) with ML models. In this study, CTs correspond to the modeling results and input parameters of rainfall abstraction processes from Green-Ampt (GA) and SCS-CN models, which provide different rainfall abstraction estimations. Integrating the GA model's CTs, which involves information on dynamic soil properties, into PGA models leads to better performance than a regular ML model. On the contrary, PGA models integrating the SCS-CN model's CTs yield no significant improvement of ML model's performance. The results of this objective demonstrate that the ML's generalization process can be improved by incorporating CTs involving dynamic soil properties.

The third objective of this dissertation is to explore the limitations of ML's generalization process in capturing trend and periodicity for streamflow simulations. Trend and periodicity are essential components of streamflow time series, representing the long-term correlations and periodic patterns, respectively. When the ML models generate streamflow simulations, they tend to have relatively strong long-term periodic components, such as yearly and multiyear periodic patterns. In addition, compared to the observed streamflow data, the ML models display relatively weak short-term periodic components, such as daily and weekly periodic patterns. As a result, the ML's generalization process may struggle to capture the short-term periodic patterns in the streamflow simulations. The biases in ML's generalization emphasize the demands for external knowledge to improve the representation of the short-term periodic components in simulating streamflow.