Data-driven discovery of governing differential equations for hydrologic systems utilizing stochastic optimization

Jinwoo Im[1], Sami Masri[1], and Felipe de Barros[1]

[1]University of Southern California

November 22, 2022

Abstract

There has been progress in machine learning (ML) to produce accurate and robust model predictions for hydrologic systems. This progress opens a new opportunity to enhance our fundamental understanding of the underlying physics of a given attribute in the subsurface environment. In order to achieve such understanding through ML in addition to reliable model predictions, we develop a general framework of system identification, named by GPfSI, which provides interpretable models representing system dynamics embedded in data. This approach aims to discover multiple governing differential equations for a target multi-physics system by combining user's prior knowledge about the system dynamics and a data set of system excitation and responses. In efforts to identify governing equations in the infinite model space in an effective way, one of the machine learning methods, Genetic Programming (GP), is employed. As a stochastic optimization method, GP is utilized to optimize differential equations to a given data set through evolutionary processes. We significantly enhance the effectiveness and computational efficiency of the GP-based identification process, by incorporating a multi-purpose loss function and stochastic sampling into the parallelized fitness test, and bloat control techniques into the evolution process. We demonstrate the proposed framework, GPfSI, against simulated and experimental data sets. In the demonstration case with the simulated data, the reference canonical model, i.e., the advection-dispersion equation (ADE), was successfully identified with a wide range of system characteristics and data noise. In the application to the experimental data from a soil column tracer study, GPfSI provided several nonlinear ADEs that are capable of capturing the anomalous (i.e., non-Fickian) tailing behavior observed in the solute breakthrough data. The inspection of the identified models provides some insights into the underlying physics of non-Fickian transport. Consequently, our results indicate that GPfSI is an effective and robust gray box modeling tool to produce accurate model predictions and enhance our fundamental understanding of hydrologic systems.

Data-driven discovery of governing differential equations for hydrologic systems utilizing stochastic optimization

Jinwoo Im, Sami Masri, and Felipe P. J. de Barros

August 2021

Abstract

There has been progress in machine learning (ML) to produce accurate and robust model predictions for hydrologic systems. This progress opens a new opportunity to enhance our fundamental understanding of the underlying physics of a given attribute in the subsurface environment. In order to achieve such understanding through ML in addition to reliable model predictions, we develop a general framework of system identification, named by GPfSI, which provides interpretable models representing system dynamics embedded in data. This approach aims to discover multiple governing differential equations for a target multi-physics system by combining user's prior knowledge about the system dynamics and a data set of system excitation and responses. In efforts to identify governing equations in the infinite model space in an effective way, one of the machine learning methods, Genetic Programming (GP), is employed. As a stochastic optimization method, GP is utilized to optimize differential equations to a given data set through evolutionary processes. We significantly enhance the effectiveness and computational efficiency of the GP-based identification process, by incorporating a multi-purpose loss function and stochastic sampling into the parallelized fitness test, and bloat control techniques into the evolution process. We demonstrate the proposed framework, GPfSI, against simulated and experimental data sets. In the demonstration case with the simulated data, the reference canonical model, i.e., the advection-dispersion equation (ADE), was successfully identified with a wide range of system characteristics and data noise. In the application to the experimental data from a soil column tracer study, GPfSI provided several nonlinear ADEs that are capable of capturing the anomalous (i.e., non-Fickian) tailing behavior observed in the solute breakthrough data. The inspection of the identified models provides some insights into the underlying physics of non-Fickian transport. Consequently, our results indicate that GPfSI is an effective and robust gray box modeling tool to produce accurate model predictions and enhance our fundamental understanding of hydrologic systems.