Using Machine Learning to Expedite Data Cleaning for Oceanographic Models of Narragansett Bay

Jonathan Benoit, Alice Foster & Baylor Fox-Kemper

Earth, Environmental, & Planetary Sciences, Brown University, Providence, RI

Computational models of Narragansett Bay are in the early stages of development. To improve these models, historical oceanographic data from the Rhode Island Department of Environmental Management are compared to hindcast runs of the models. Before being used for comparison, the observational data is cleaned manually because fouling and other unexpected effects produce erroneous data. There is currently a multi-year long gap between the data that has been taken and the data that has been cleaned. Historical data over the past few years is needed to expand the scope of hindcast model runs, and near real-time data is needed to develop a forecasting model. This project investigated the usefulness of machine learning in expediting data cleaning. By training a classification algorithm with samples of clean and raw data, a preliminary tool was developed that could accelerate the cleaning process by flagging data spikes and timeframes where fouling calculations are necessary. The machine learning algorithm was trained using two methods. In one method, all nine variables were input for one time, while in a second method, a single variable was input along with neighboring values in time. Though neither of these algorithms produces publication-ready cleaned data, this project is a proof of concept and creates a tool for speeding up data cleaning and providing preliminary data for predictive model runs.