A hybrid random forest model for prediction of daily surface water nutrient concentrations

Benya Wang (1,2,3), Carolyn Oldham (1,3), Matthew Hipsey (2,3)

(1) School of Engineering, University of Western Australia, Crawley, Australia , (2) UWA School of Agriculture and Environment, University of Western Australia, Crawley, Australia, (3) Co-operative Research Centre for Water Sensitive Cities, Clayton, Australia

Nutrient concentration data are often sparse in space and time, compared to flow and rainfall data, and different historical investigations may also analyse different nutrient components. These issues make it difficult to apply the data for nutrient load assessments and analysis of trends. To fully utilize all available nutrient data, we apply a random forest model (RF) which is a widely-used machine learning method that has shown its superior performance in hydrological studies. We developed a hybrid RF model that first generates the missing nutrient components of the total nutrient pool using relationships with flow and rainfall. The generated nutrient data were then included as new variables for total nutrient prediction. This method was tested on TN, DOC, DON, TP, and TSS prediction in Ellen Brook (Western Australia). The results were compared with a RF model that did not pre-predict nutrient data, and a common multivariate regression model (EGRET), and results for the hybrid RF had significantly lower error and higher R2 for all predicted variables. The hybrid model provides a flexible method to combine data of variable resolution and quality, and is accurate for the prediction of surface water nutrient concentrations in response to hydrologic variability.