

A Machine Learning Approach to Identifying the Key Factors Influencing Global Water Quality

Razi Sheikholeslami¹ & Jim Hall

School of Geography and the Environment, University of Oxford, Oxford, UK

Due to its substantial role on the Earth’s biogeochemical cycles and human health, nitrogen is recognized as one of the major water quality indicators of Sustainable Development Goal 6.3.2. Quantifying these potential impacts in large spatial scales still appears to be a grand challenge because of the high computational demand required by the distributed physically based global models and their intensive data requirements for calibration and validation. The former prevents a comprehensive analysis of the full spectrum of the model behavior under different conditions, and the latter impinges on the reliability of model-based inference. To tackle this problem, we developed a data-driven model using a spatio-temporal Random Forest algorithm to predict levels of nitrogen at 0.5-degree spatial resolution from 1992 to 2010 across the world. Several variables representing livestock, climate, hydrology, topography, etc. have been selected as predictors. The response variable of interest was nitrate–nitrite, which is responsible for the high risk of infant methemoglobinemia. Our results indicate that changes in the nitrogen concentration is mainly driven by cattle and sheep population, fertilizer application, precipitation, and temperature variability, implying livestock population, climate change, and anthropogenic forces can be important risk factors for global water quality deterioration. Furthermore, using the predicted levels of nitrogen, we characterized large-scale water quality patterns, and thus identified a few major ‘hot spots’ of water quality. The proposed model can also help assess potential impacts of future scenarios on water quality conditions (e.g., livestock production or land use change) for better development of effective policy strategies.

¹*razi.sheikholeslami@ouce.ox.ac.uk*