Using big groundwater data to detect methane contamination in water within hydrocarbon production areas across the United States

T. WEN[1]*, M. LIU[2], Z. LI[2], S.L. BRANTLEY[1]

[1]Earth and Environmental Systems Institute, Penn State University, University Park, PA 16802, USA (*correspondence: twen08@syr.edu)

[2]College of Information Sciences and Technology, Penn State University, University Park, PA 16802, USA

Aqueous systems can be assessed by chemical equilibrium-solving programs, many of which are often proprietary and/or might require highly sophisticated domain knowledge to ensure proper use. Such assessments of thermodynamic condition are often essential for many environmental problems. In the face of increasing need for energy and water, the mostly widely reported environmental problem in hydrocarbon production areas is probably the leakage of methane (CH_4) into shallow groundwater and atmosphere. However, distinguishing between CH_4 leaked from hydrocarbon production wells and naturally-occurring CH_4 (biogenic or thermogenic) remains challenging.

Based on newly released data sets of large volumes of groundwater quality data (https://doi.org/10.4211/his-data-shalenetwork) within the Marcellus shale area, two models have been developed to detect sites that might show evidence of localized CH_4 contamination from hydrocarbon production. The first is a geospatial model [1] that considers spatial context of groundwater samples and the second is an empirical fingerprinting model [2] based on six chemical analytes. In this study, we propose a machin learning-based ensemble model to predict i) CH_4 concentration below or above prescribed thresholds, and ii) the likelihood of a sample being impacted by recently-invaded CH_4, based on the complete high-dimensional groundwater quality data (i.e., features). The likelihood is generated by synthesizing prediction results from the ensemble model with feature bagging. The learned models have also been tested on compiled data from hydrocarbon production areas in multiple states (e.g., PA, NY, CO, and TX) with different histories of energy extraction activities in the United States. The proposed algorithm is easy to implement in other hydrocarbon production areas in the world and is well complementing chemical equilibrium-solving programs.

[1] Wen *et al.* (2018) *Environ. Sci. Technol.* **52**, 7149-7159.
[2] Wen *et al.* (2019) *Environ. Sci. Technol.* **53**, 9317-9327.