EGU21-16388
https://doi.org/10.5194/egusphere-egu21-16388
EGU General Assembly 2021

Groundwater pollution monitoring and the inverse problem of source identification. Evaluation of various Machine Learning methods

Yiannis Kontos[1,2], Theodosios Kassandros[2], Konstantinos Katsifarakis[1], and Kostas Karatzas[2]
[1]Lab. of Water Resources Engineering and Management, Div. of Hydraulics and Environmental Engineering. School of Civil Engineering, Aristotle Univ. of Thessaloniki, GR-54124, Thessaloniki, Greece
[2]Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle Univ. of Thessaloniki, GR-54124, Thessaloniki, Greece

Groundwater pollution numerical simulations coupled with Genetic Algorithms (GAs) lead to vast computational load, while flow fields' simplification can compensate in design, but not real-time/operational, applications. Various Machine Learning/Deep Learning (ML/DL) methods/problem-formulations were tested/evaluated for real-time inverse problems of aquifer pollution source identification. Aim: investigate data-driven approaches towards replacing flow simulation with ML/DL trained models identifying the source, faster but efficiently enough.

Steady flow in a 1500mx1500m theoretical confined, isotropic aquifer of known characteristics is studied. Two pumping wells (PWs) near the southern boundary provide irrigation/drinking water, defining the flow together with a varying North-South natural flow. Six suspected possible sources, capable of instantaneous leakage, may spread a conservative pollutant. Particle tracking simulates advective mass transport, in a 2D flow-field for 2500 1-day timesteps. The 14x14 inner field grid nodes serve as locations of sources, PWs and monitoring wells (MWs; for simple daily yes/no pollution detection and/or drawdown measuring). 15,246 combinations of 6 Source Nrs, 21 N-S hydraulic gradients, 11+11 PW1,2 flow-rates were simulated with existing own software, providing the necessary data-sets for ML training/evaluation.

Two basic ML/DL approaches were implemented: Classification (CL) and Computer Vision (CV). In CL, every source is a discrete class, while each MW is a discrete variable. The target variable Y can equal 1 to 6, while input variables X can be: a) 0/1 (MW_i polluted or not), b) the first day of MW_i's pollution, c) the duration of MW_i's pollution, d) hydraulic drawdown of MW_i. For a bit more realism, the two southern rows of 28 MWs, and the MWs on/around PWs are concealed. CL features the advantage of facilitating Correlation-based Feature Subset Selection (CFSS), indirectly leading to a pseudo-optimization of the monitoring network, minimizing the number of MWs (not the sampling frequency though), based solely on the efficiency in identifying the source criterion. As a downside, time dimension and spatial correlation of MWs are not considered. Approach (b) being the best scheme, Random Forests (RFs; 86.5576% accuracy), Multi-Layer Perceptron (MLP; 77.5%), and Nearest Neighbors (NN; 86.5%) were tested. CFSS led to 8 only MWs being important, so training with the optimal subsets gave promising results: RF=85.4%, MLP=73.1%, NN=85.4%. In CV, MW_is' pollution input data on a 10-day basis (0-60, 800-on concealed) were formulated into 14x14-pixel black/white images, that is 14x14 binary (0,1) matrices, the t=0 image being the desideratum. A Convolutional Neural Network (CNN; U-Net architecture for image segmentation) achieved 97.1% accuracy. A Convolutional Long/Short-Term Memory Neural Network (CLSTM), training a model to back-propagate predicting each given time step, with unchanged data formulation (60-800d, step 10), exhibits 82.3% accuracy. CLSTM's performance is timestep-sensitive, best results yielded (98% accuracy) using configuration 5-800d, step 6.

Concluding, CL's CFSS minimizes the input space, while CV approaches yield more promising results in terms of accuracy. Each approach has certain constraints in operational applicability, concerning the number of MWs, the sampling resolution and the total elapsed time. This process paves the way for realistic inverse problem solutions, ML-GAs monitoring network optimization, and real-time pollution detection operational systems.