Predicting PFAS Contamination in Drinking Water Sources Using Classical Machine Learning Approaches

Olha Khomenko
Cambridge Environmental Assessments (RSK ADAS), Boxworth, Cambridge, UK

Per- and polyfluoroalkyl substances (PFAS) contamination of surface waters poses a significant public health concern due to their persistence, bioaccumulation, and toxicity. While spatial patterns of PFAS contamination are increasingly documented, the role of meteorological events, particularly rainfall, in driving temporal variability remains poorly understood.

Objectives of this work:

1. **Identify sites** where rainfall significantly increases PFAS concentrations using rigorous statistical tests
2. **Quantify the relationship** between precipitation, surface/subsurface runoff, and PFAS mobilisation
3. **Develop predictive models** using machine learning to forecast PFAS spikes from weather data alone
4. **Characterise spatial patterns** of mobilisation to inform targeted monitoring and risk assessment

MATERIALS AND METHODS

Data

PFAS monitoring data (18,236 samples, 241 sites) from the Environment Agency Water Quality Archive (2016–2026) were matched with ERA5 (Copernicus Climate Change Service) reanalysis meteorological data (precipitation, surface runoff, subsurface runoff).

Stage 1: Statistical Identification

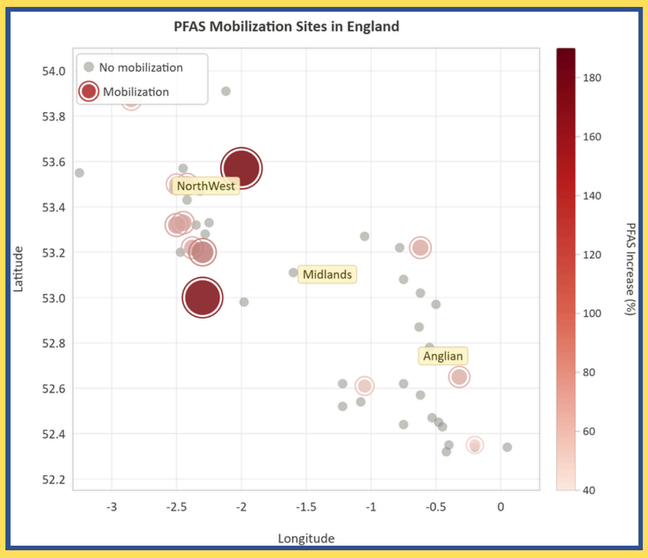
For each site, we compared PFAS concentrations between high-rain (> 15mm in 7 days) and low-rain (≤ 15mm) periods using the Mann-Whitney U test. Sites showing statistically significant increases ($p < 0.05$) exceeding 30% with adequate sample sizes ($n \geq 5$ per group) were classified as mobilisation sites.

Stage 2: Predictive Modelling

Samples from mobilisation sites were pooled ($n = 424$) and used to train machine learning models with 18 features, including precipitation metrics, surface/subsurface runoff, seasonal variables, and site identity. Three algorithms were tested (Linear Regression, Random Forest, Gradient Boosting) using 70/30 train-test splits and 5-fold cross-validation.

Figure 1. Spatial distribution of rainfall-driven PFAS mobilisation sites in England. Red circles indicate sites exhibiting statistically significant PFAS mobilisation during high-rainfall events ($n = 13$); grey circles indicate sites with no significant mobilisation ($n = 37$).

KEY FINDINGS



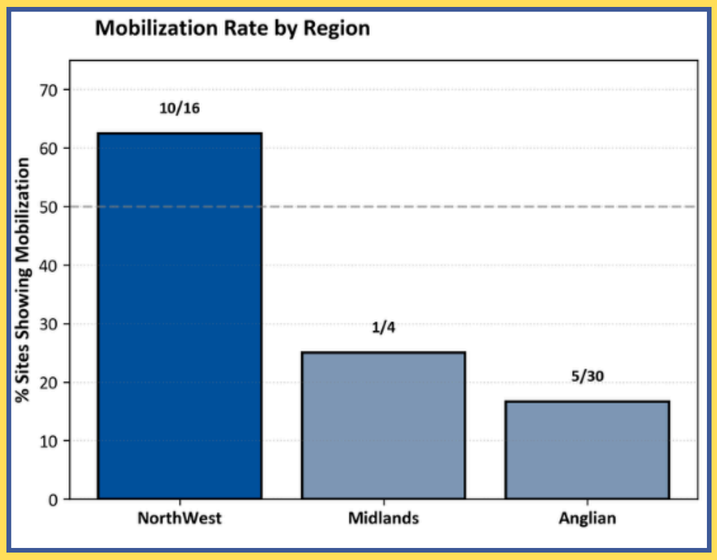


Figure 2. Geographic Clustering of Mobilisation

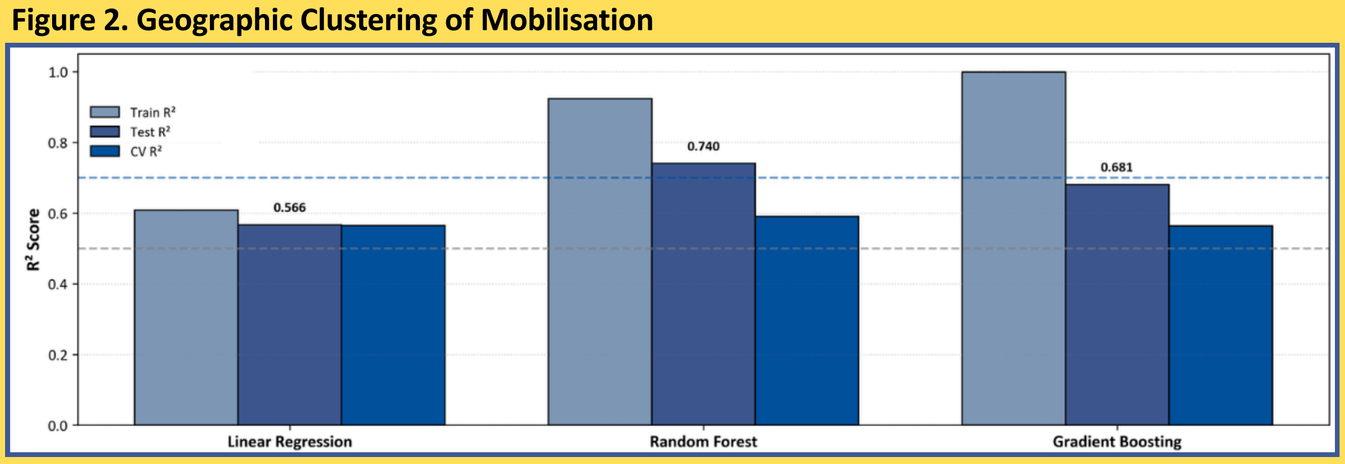


Figure 3. Predictive Model Performance

- Mobilisation is site-specific, not universal: 26% of sites (13/50) show statistically significant rainfall-driven PFAS spikes ($p < 0.05$)
- Highly predictable at contaminated sites: Random Forest model explains 74% of PFAS variance—excellent performance for environmental systems
- Industrial hotspot identified: NorthWest England (Manchester/Liverpool) accounts for 77% of mobilisation sites. Mobilisation is primarily an industrial/urban phenomenon
- Agricultural sites show minimal mobilisation: Only 17% of Anglian (agricultural) sites vs 63% of NorthWest (industrial) sites exhibit rain-driven spikes
- Multiple pathways matter: Surface runoff (25% importance) and subsurface runoff (14% importance) drive mobilisation at contaminated industrial sites
- Substantial increases: Mean PFAS spike of 78% during high-rain events (> 15mm in 7 days) at industrial sites, with individual sites up to 410%

CONCLUSION

Rainfall-driven PFAS mobilisation is a predictable, site-specific phenomenon primarily affecting industrial and urban catchments with legacy contamination. Machine-learning models achieved excellent predictive performance ($R^2 = 0.74$) using only hydro-meteorological data, demonstrating that weather-based early-warning systems are feasible for high-risk sites. Geographic clustering in Northwest England (77% of mobilisation sites) and low mobilisation rates in agricultural regions (17%) indicate that interventions should be geographically targeted rather than universally applied.

Because mobilisation risk varies sharply between regions and land-use types, organisations responsible for chemical stewardship, industrial sites and producers of regulated chemicals may face differing levels of operational and regulatory exposure. Spatially explicit, higher-tier modelling can help identify where proactive management or targeted monitoring would be most effective.

Copernicus Climate Change Service, Climate Data Store, (2024): ERA5 post-processed daily-statistics on single levels from 1940 to present . Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.4991cf48 (Accessed on DD-MMM-YYYY)