Identifying PFAS Hotspots Influenced by Hydrologic & Climatic Stressors Using Remote Sensing and Geospatial Analysis

Kathleen I. Shank[1], Himmat Basnet, Eunsang Cho, Sangchul S. Hwang
Department of Civil Engineering | Ingram School of Engineering | Texas State University | [1]nat98@txstate.edu

Poster #448

INTRODUCTION

- Per- and polyfluoroalkyl substances (PFAS) are persistent "forever chemicals" that resist environmental degradation and accumulate in natural systems
- Due to their strong carbon-fluorine bonds, PFAS have been detected throughout the hydrologic cycle, raising concerns about widespread exposure (Tokranov et al., 2024)
- PFAS exposure risks are elevated in communities near known point sources, including industrial facilities, military installations, airports, wastewater treatments plants, agricultural areas with irrigation or biosolids application
- **Objective:** Evaluate how climate variables influence PFAS concentrations and identify potential contamination hotspots across the conterminous United States (CONUS)

DATA & METHODS

- **Data Integration:** PFAS occurrence and concentration data from Environmental Protection Agency's 3rd and 5th Unregulated Contaminant Monitoring Rule (UCMR), covering years 2013-15 and 2023-25, respectively. Concentrations of multiple compounds were averaged to generate a single composite PFAS concentration, allowing comparisons across locations
- U.S. Census tract-level demographic data, population, and military site locations within a ZIP code
- **Vulnerability Index:** Equally weighted %Population below poverty line, %Black, %Hispanic or Other, %Population without health insurance, %Renter occupied households, Normalized median household income
- **Remote Sensing Products:** Annual dominate land cover type of each ZIP code from National Landcover Database (NLCD)
- **Climate & Hydrologic Variables:** Annual averaged precipitation and temperature from Parameter-elevation Regressions on Independent Slopes Model (PRISM), Annual maximum and minimum drought indices from the Standardized Precipitation Evapotranspiration Index (SPEI)

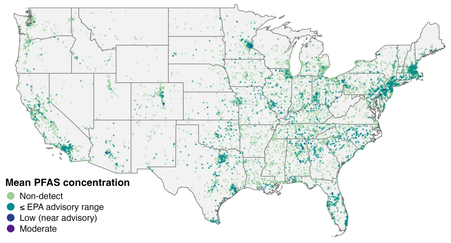

Figure 1: Mean PFAS concentrations across ZIP codes, values categorized using EPA-informed screening thresholds

Table 1: Categorical variables used in K-Means clustering, RF, and XGBoost modeling

Category	Variables
Temporal	year
Climate/Hydrologic Variables	precip_ann, temp_ann, spei_min, spei_max
Location	lat, lon
Land Cover	lc_code
Socio-Demographic	population
Potential PFAS Sources	military_site_count
Identifier (excluded from training)	zip_code

- **Geospatial & Statistical Analysis:** Spatial joining and aggregation of 5878 ZIP codes to census tract scale
- **K-Means Clustering:** Identify spatial PFAS risk groupings utilizing 3 optimal clusters, determined by Elbow and Silhouette methods
- **Machine Learning Models:** Random Forest (RF) (number of trees = 200 to 300, 80/20 split for training and testing and Extreme Gradient Boosting (XGBoost)

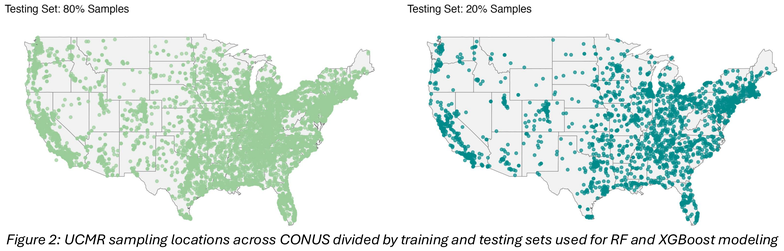

Figure 2: UCMR sampling locations across CONUS divided by training and testing sets used for RF and XGBoost modeling

RESULT 1

- **K-Means Clustering:** Moderate PFAS risk (Cluster 1) are ZIP codes with average PFAS levels, climate stress, and mid-level population; High PFAS risk (Cluster 2) have elevated PFAS levels, higher population, and more military sites; Low PFAS risk (Cluster 3) have low PFAS levels, lower populations, potentially rural or less industrialized areas
- Clustering pattern generally follows national climate patterns as low risk dominates the northeast and Midwest, high risk is distributed amongst the west, and moderate is concentrated around the geopolitical south

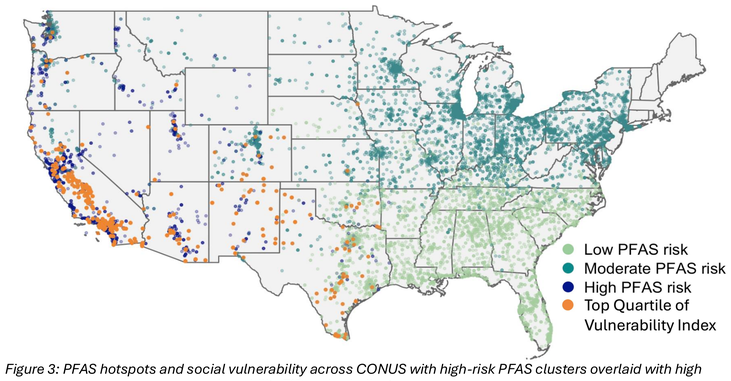

Figure 3: PFAS hotspots and social vulnerability across CONUS with high-risk PFAS clusters overlaid with high vulnerability ZIP codes indicate areas of overlap

- **Vulnerability Index:** Each indicator is a standardized z-score and averaged to produce a single Vulnerability Index value
- **Higher values** indicate populations that may be more socially vulnerable to environmental hazards, including PFAS exposure
- **Lower values** indicate relatively less vulnerable populations

RESULT 2

- **Random Forest:** Tested 200 to 300 trees with 80/20 stratified train/test split, and 5-fold cross-validation (CV) showed low but stable predictive skill, indicating PFAS stability is dominated by localized sources beyond regional environmental predictors
- RF model explains 5% (R^2 = 0.05) of the variance in PFAS concentrations, indicative of limited but consistent predictive skill across folds
- RMSE is small (0.0025 µg/L) in absolute terms but remains large relative to the narrow PFAS concentration range, suggesting that most fine-scale variability is not captured by the available predictors

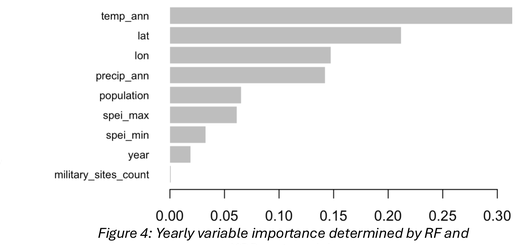

Figure 4: Yearly variable importance determined by RF and XGBoost models

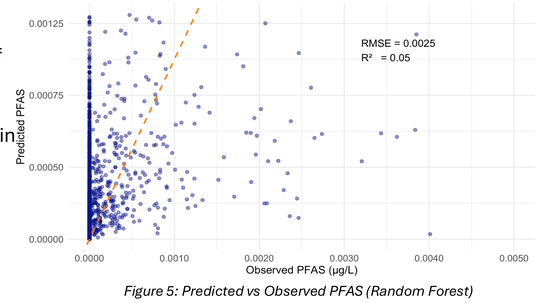

Figure 5: Predicted vs Observed PFAS (Random Forest)

Table 2: Comparative model performance parameters

	Averaged Variable Results			Yearly Variable Results		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.25	0.0025	0.0005	0.05	0.002	0.0004
XGBoost	0.01	0.0026	0.0005	0.01	0.003	0.0005

RESULT 3

- **XGBoost:** Trained on numeric variables with 80/20 stratified train-test split, 5-fold CV metrics indicated limited predictive skill, highlighting challenges associated with sparse PFAS observations, extremely low concentration variability, and reliance on indirect proxy variables.
- Predicted PFAS concentrations closely track observed values (Figure 6), but explained variance remains negligible compared to extremely low target concentrations
- Figure 7 shows substantial predictive uncertainty across CONUS, with many ZIP codes exhibiting >100% error relative to observed PFAS

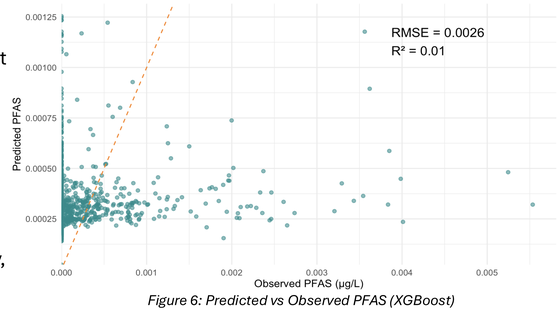

Figure 6: Predicted vs Observed PFAS (XGBoost)

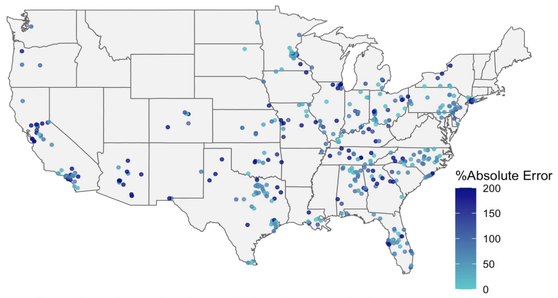

Figure 7: Spatial distribution of XGBoost %Absolute Error with error values clipped at 200% for visual clarity

CONCLUSIONS

- Generally averaged variables (Table 2) performed better than the yearly variables due to added noise in the data, despite hyper-parameter tuning
- PFAS risk is unevenly distributed across the U.S., disproportionately affecting communities near known contamination sources
- Despite hyperparameter tuning and cross-validation, validation metrics (RMSE, MAE, R^2) indicate moderate predictive performance, reflecting the challenges of sparse PFAS observations, spatial heterogeneity, and complex environmental drivers

FUTURE RESEARCH

- Expand the analysis to UCMR water quality monitoring locations to AK, HI, and PR
- Include additional climate, hydrologic, and PFAS point source variables to enhance model accuracy
- Enhance PFAS data availability with a PFAS-proxy consisting of nitrate, chloride, specific conductance, and tritium detections as this data is more robust and these contaminants tend to co-exist in PFAS contaminated waters
- Incorporate lagged climate and hydrologic variables to train models on previous-year conditions and predict future PFAS concentrations

ACKNOWLEDGMENTS

- This research builds off previous study: Tokranov et al., 2024
- The authors thank Jipeng Liu for processing climate datasets and the Cho Hydrology Lab for research direction, feedback, and technical assistance
- This study was funded, in part, by the Department of Energy (Award #DE-EM0005314)
- Go to QR code for references