AI-Driven Predictive Modeling of PFAS Contamination in Aquatic Ecosystems: Exploring A Geospatial Approach

Jowaria Khan
University of Michigan, Ann Arbor
jowaria@umich.edu

David Andrews
Environmental Working Group
dandrews@ewg.org

Kaley Beins
Environmental Working Group
kaley.beins@ewg.org

Sydney Evans
Environmental Working Group
sydney.evans@ewg.org

Alexa Friedman
Environmental Working Group
alexa.friedman@ewg.org

Elizabeth Bondi-Kelly
University of Michigan, Ann Arbor
ecbk@umich.edu

Abstract

Per- and polyfluoroalkyl substances (PFAS), a class of synthetic fluorinated compounds termed “forever chemicals”, have garnered significant attention due to their persistence, widespread environmental presence, bioaccumulative properties, and associated risks for human health. Their presence in aquatic ecosystems highlights the link between human activity and the hydrological cycle. They also disrupt aquatic life, interfere with gas exchange, and disturb the carbon cycle, contributing to greenhouse gas emissions and exacerbating climate change. Federal agencies, state governments and non-government research and public interest organizations have emphasized the need for documenting the sites and the extent of PFAS contamination. However, the time-consuming and expensive nature of data collection and analysis poses challenges. It hinders the rapid identification of locations at high risk of PFAS contamination, which may then require further sampling or remediation. To address this data limitation, our study leverages a novel geospatial dataset, machine learning models including frameworks such as Random Forest, IBM-NASA’s Prithvi and UNet, and geospatial analysis to predict regions with high PFAS concentrations in surface water. Using fish data from the National Rivers and Streams Assessment (NRSA) dataset by the Environmental Protection Agency (EPA), our analysis suggests the potential value of machine learning based models for targeted deployment of sampling investigations and remediation efforts.

1 Background and Motivation

PFAS are a group of over 9,000 synthetic chemicals first manufactured in the late 1940s. Known for their exceptional stability and resistance to degradation, PFAS are often termed "forever chemicals" due to their persistence in the environment and tendency to bioaccumulate in living organisms. These chemicals are found in the blood of 97% of Americans, according to the CDC, and have contaminated the drinking water of over 110 million people in the U.S. alone [1] - [2] . Human exposure to

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

PFAS occurs through various pathways, including contaminated food, drinking water, air, household products, and dust. Exposure to some PFAS is associated with serious health issues, including cancer. Additionally, some PFAS are classified as fluorinated greenhouse gases (F-GHGs), which are typically the most powerful and enduring greenhouse gases [3]. Inefficient waste management practices involving PFAS-containing materials have further contributed to their widespread pollution [4].

Given the complexity of environmental and health risks like those posed by PFAS, artificial intelligence (AI) has the potential to be a powerful tool to help identify the scope of contamination and areas of highest priority for remediation. There is precedence that AI has the potential to automate labor-intensive tasks, more strategically utilize and benefit from domain expertise, and derive valuable insights from complex datasets, in domains ranging from conservation, to health, to agriculture [5]. These domains also frequently use geospatial and remote sensing data for tasks such as land cover mapping [6], disease prediction [7], and agricultural productivity assessments [8].

In the context of PFAS contamination, spatial analysis has been employed to map pollution sources like industrial facilities and waste management areas [9], predict PFAS levels in fish tissue, and assess contamination risks at specific sites [10]. Recent work has also focused on predicting PFAS contamination in groundwater at a national scale [11]. Despite these advancements, many spatial analysis methods remain labor-intensive and require significant domain knowledge, which can hinder their scalability and efficiency.

Our study aims to address these limitations by comparing different machine learning models and data processing techniques. These include advanced models like IBM-NASA's Prithvi [12] and UNet [13], as well as traditional models like Random Forest [14], coupled with a novel geospatial dataset to predict PFAS concentrations in surface water. By integrating diverse data sources, including environmental and industrial factors, our predictive models identify areas at high risk of contamination. These models, once further validated by field sampling, can help guide environmental policy and intervention efforts in real time.

2 Methods

2.1 Dataset processing

Our study focuses on classifying high and low PFAS concentrations in the U.S. using fish tissue PFAS concentration data from the NRSA [15] dataset. Samples were classified as high (1) or low (0) concentration based on the U.S. EPA's Fish and Shellfish Advisory Program thresholds, determined for each type of PFAS in fish tissue. The individual assessments were combined to produce a final hazard quotient, where 1 indicates a high hazard and 0 indicates a low hazard. The NRSA dataset consists of 800 sample points across various years, with only 52 classified as 0, making it highly imbalanced.

We use a variety of datasets to predict PFAS contamination, including the U.S. Environmental Protection Agency (EPA) Enforcement and Compliance History Online (ECHO) PFAS dischargers' data [16], which provides point data in the form of latitude and longitude coordinates, and raster datasets like the National Land Cover Database (NLCD) [17], Digital Elevation Model (DEM) elevation data [18], and the Watershed Boundaries Dataset (WBD) [19]. These diverse data sources present unique challenges because point data like ECHO requires spatial context, while raster data like NLCD is inherently image-based.

To address this challenge, we analyzed two complementary approaches for data processing: tabular and image-based.

The tabular approach combined multiple data sources, extracting features within a 5 km radius around each sample point. These features included land cover percentages, mean elevation, median values of remote sensing indices like Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI), and distances to the nearest facility types.

In the image-based approach, we created multichannel raster images at a 30m resolution, each sized 512x512 pixels, centered on the sample points. Each image incorporated different data sources, including portions of rasters derived from EPA ECHO, NLCD, DEM, and other satellite data products, enabling a visual representation of the tabular features. Visualizations of some channels in a sample image are shown in Figure 1.

To generate labels for our images, given the point-level nature of our raw label data, we experimented with different configurations. Our most intuitive approach was to label all surface water areas in the image patch as 0 or 1, indicating high or low PFAS concentration, based on the known label at the center. This approach was chosen to prevent the imbalance that would result from labeling only a single pixel per image patch. It is important to note that this strategy is noisy due to the lack of ground truth data beyond the specific point locations, where only the points themselves provide verified data. We chose a BootStrapping cross-entropy loss function [20] to account for these noisy labels.

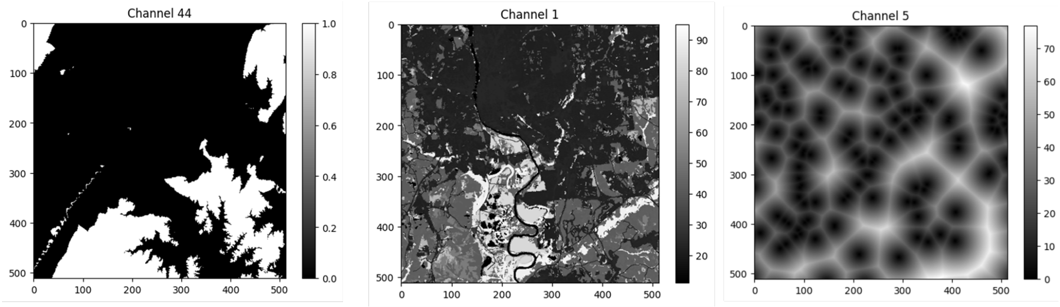

Figure 1: The figure shows three channels of a sample image from the dataset: (1) the watershed-elevation channel, (2) the NLCD channel, and (3) the discharger channel. The images are arranged from left to right.

2.2 Experimental design

For both the tabular and image-based data modalities, we applied a spatially stratified split to divide the data into training and test sets. This approach ensured that there was no geographic overlap between the regions surrounding the different points, allowing us to account for spatial variability.

We experimented with multiple models. First, for image-based data, we implemented the Prithvi model architecture, which features a transformer-based architecture adapted for geospatial data. Initially, the model was pre-trained using a Masked Autoencoder approach on our image data, which enabled the model to learn robust feature representations. We pre-trained our model here instead of directly using the Prithvi weights since raw satellite imagery does not fully capture the nuances needed to predict PFAS in aquatic ecosystems. Derived geospatial data products, like land cover, are not only more relevant but also enhance model explainability. The encoder was then initialized with these pre-trained weights and the decoder was subsequently finetuned using our label masks. In addition to the Prithvi model, a UNet architecture was utilized equipped with deep supervision, attention gates, and pyramidal pooling modules. The training was conducted using a BootStrapping cross-entropy loss function, which helped address label noise by blending the model's predictions with the ground truth, reducing the impact of potentially incorrect labels. We also used a Random Forest model with the tabular version of the data, following the approach outlined by Smith et al. [10]. This model relies on an ensemble of decision trees to make predictions. The direct data points (0 and 1) were used as labels in this model.

In this study, we evaluated model performance using precision, recall, F1-score, and accuracy. Among these, we prioritize F1-score because it balances precision and recall, making it especially relevant for distinguishing high or low PFAS concentration areas where both false positives and false negatives have significant consequences. This approach ensures a robust evaluation of the models' ability to guide targeted interventions. Performance validation will be further supported by field sampling. For a more comprehensive overview of the dataset creation process and training configurations please refer to the supplementary material.

3 Results

Table 1: Class-wise results on the test sets of NRSA across 2008 and 2019

Dataset	Method	Class 0			Class 1			
		F-score	Precision	Recall	Avg. Accuracy	F-score	Precision	Recall
NRSA - 2008	Random Forest	32.00	30.00	33.00	77.00	86.00	87.00	85.00
	Prithvi	**50.00**	**43.00**	60.00	**84.00**	**90.00**	**93.00**	**88.00**
	UNet	33.00	21.00	**80.00**	67.00	67.00	92.00	53.00
NRSA - 2019	Random Forest	**44.00**	67.00	33.00	**94.00**	**97.00**	95.00	**99.00**
	Prithvi	36.00	25.00	**67.00**	82.00	90.00	**97.00**	84.00
	UNet	13.00	**99.00**	17.00	52.00	84.00	92.00	78.00

We report the results in Table 1. The results from Prithvi are based solely on the verified actual points from the NRSA dataset, ensuring accurate validation. Overall, the Prithvi model seems to perform generally better across the NRSA data. We hypothesize that the Prithvi model, pre-trained using a masked autoencoder, provides a performance boost by extracting deep, meaningful features from geospatial data. The combination of initial feature pre-training and fine-tuning using label masks appears to enable the model to capture nuances in surface water contamination data, yielding generally better results overall. However, we also hypothesize that the results on the 2019 data may be more varied due to a more severe class imbalance compared to the 2008 data.

On the tabular dataset, we outperformed the baseline Random Forest results from [10], which achieved 81% accuracy on the NRSA dataset focused on the Columbia River Basin region. Using the same model on our train and test sets, this result underscores the importance of dataset context.

Due to its strong performance, our Prithvi model, trained on 2008 NRSA samples, was tested on a Michigan area near some ground truth points from the same dataset to assess generalization, as shown in Figure 2. The predicted high concentration areas (purple) in Figure 2 generally align well with the high concentration ground truth points (purple), particularly in closer proximity. While this is a simplified observation, it highlights the potential of the model's generalization capability and represents a promising first step, which will be further validated through field testing.

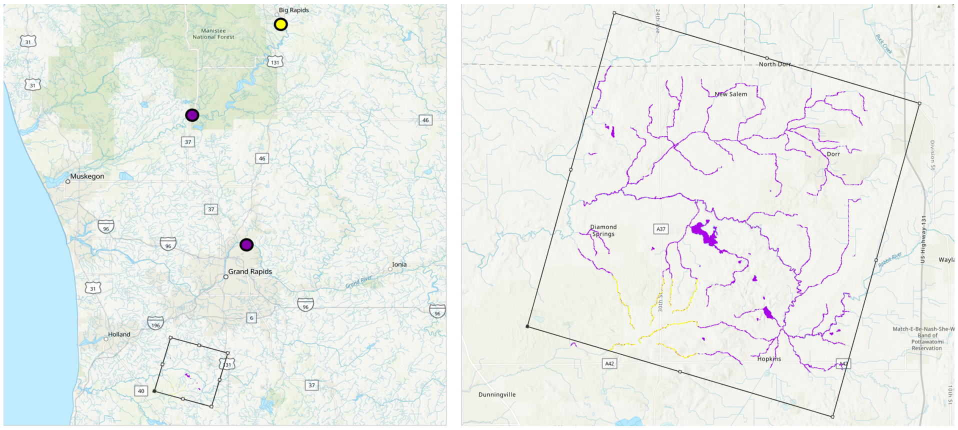

Figure 2: From left to right: (1) purple and yellow points denote sparse high and low concentration data from 2008 NRSA, respectively, with the black box highlighting the area of Prithvi's generalization result ; (2) The predicted mask with purple and yellow indicating high and low concentration regions.

4 Discussion

The main contribution of this work is to lay the groundwork for AI-generated maps of PFAS contamination in surface water, starting with estimated PFAS levels in fish tissue. In future work, we aim to quantify uncertainty in labels and in output predictions to both improve performance, and to prepare to communicate results to potential users and impacted communities. This will be important to inform decision-making, given the potential impact on public health and home property values [21], and environmental justice implications [22]. We aim to convey this uncertainty via maps [23].

To tackle this challenge, we aim to design loss functions to better quantify label uncertainty and incorporate PFAS domain knowledge, such as groundwater flow and proximity to toxic facilities. Additionally, a regression task could offer deeper insights than classification, if we can overcome the added sparse data challenges.

Community-driven mapping efforts, like those by the PFAS Project [24], West Plains Water Coalition [25], and the Environmental Working Group [26], have raised awareness of PFAS pollution. Moving forward, we plan to collaborate with local community partners in water pollution and sustainability to refine our predictive map. We will seek their input on its accuracy and usefulness, building on our existing partnership with a nonprofit organization.

References

[1] CDC, “Centers for disease control and prevention,” Aug. 2024. [Online]. Available: https://www.cdc.gov/index.html

[2] D. Q. Andrews and O. V. Naidenko, “Population-wide exposure to per- and polyfluoroalkyl substances from drinking water in the united states,” *Environmental Science Technology Letters*, vol. 7, no. 12, p. 931–936, Dec. 2020.

[3] U.S. Environmental Protection Agency, “Pfas analytic tools,” 2024, accessed: 2024-11-25. [Online]. Available: https://echo.epa.gov/trends/pfas-tools

[4] A. Mahmoudnia, “The role of pfas in unsettling ocean carbon sequestration,” *Environmental Monitoring and Assessment*, vol. 195, no. 2, p. 310, Jan. 2023.

[5] Z. R. Shi, C. Wang, and F. Fang, “Artificial intelligence for social good: A survey,” no. arXiv:2001.01818, Jan. 2020, arXiv:2001.01818 [cs]. [Online]. Available: http://arxiv.org/abs/2001.01818

[6] A. Poortinga, Q. Nguyen, K. Tenneson, A. Troy, D. Saah, B. Bhandari, W. L. Ellenburg, A. Aekakkararungroj, L. Ha, H. Pham, G. Nguyen, and F. Chishtie, “Linking earth observations for assessing the food security situation in vietnam: A landscape approach,” *Frontiers in Environmental Science*, vol. 7, dec 2019. [Online]. Available: https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2019.00186/full

[7] N. A. Rehman, U. Saif, and R. Chunara, “Deep landscape features for improving vector-borne disease prediction,” no. arXiv:1904.01994, Apr. 2019, arXiv:1904.01994 [cs]. [Online]. Available: http://arxiv.org/abs/1904.01994

[8] C. Nakalembe, “Urgent and critical need for sub-saharan african countries to invest in earth observation-based agricultural early warning and monitoring systems,” *Environmental Research Letters*, vol. 15, no. 12, p. 121002, Dec. 2020.

[9] D. Salvatore, K. Mok, K. K. Garrett, G. Poudrier, P. Brown, L. S. Birnbaum, G. Goldenman, M. F. Miller, S. Patton, M. Poehlein, J. Varshavsky, and A. Cordner, “Presumptive contamination: A new approach to pfas contamination based on likely sources,” *Environmental Science Technology Letters*, vol. 9, no. 11, p. 983–990, nov 2022.

[10] N. M. DeLuca, A. Mullikin, P. Brumm, A. G. Rappold, and E. Cohen Hubal, “Using geospatial data and random forest to predict pfas contamination in fish tissue in the columbia river basin, united states,” *Environmental Science Technology*, vol. 57, no. 37, p. 14024–14035, sep 2023.

[11] A. K. Tokranov, K. M. Ransom, L. M. Bexfield, B. D. Lindsey, E. Watson, D. I. Dupuy, P. E. Stackelberg, M. S. Fram, S. A. Voss, J. A. Kingsbury, B. C. Jurgens, K. L. Smalling, and P. M. Bradley, “Predictions of groundwater pfas occurrence at drinking water supply depths in the united states,” *Science*, vol. 386, no. 6723, pp. 748–755, Nov 2024.

[12] J. Blumenfeld, “Nasa and ibm openly release geospatial ai foundation model for nasa earth observation data | earthdata,” Aug. 2023. [Online]. Available: https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model

[13] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” no. arXiv:1505.04597, May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

[14] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, p. 5–32, Oct. 2001.

[15] US EPA, OW, “National rivers and streams assessment,” Jun. 2015. [Online]. Available: https://www.epa.gov/national-aquatic-resource-surveys/nrsa

[16] “Water pollution search | echo | us epa.” [Online]. Available: https://echo.epa.gov/trends/loading-tool/water-pollution-search

[17] “National land cover database | u.s. geological survey.” [Online]. Available: https://www.usgs.gov/centers/eros/science/national-land-cover-database

[18] “Usgs 3dep 10m national map seamless (1/3 arc-second) | earth engine data catalog.” [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/USGS_3DEP_10m

[19] “Watershed boundary dataset | u.s. geological survey.” [Online]. Available: https://www.usgs.gov/national-hydrography/watershed-boundary-dataset

[20] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, “Training deep neural networks on noisy labels with bootstrapping,” no. arXiv:1412.6596, Apr. 2015, arXiv:1412.6596 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6596

[21] “How pfas contamination is affecting area property values - varnum llp,” jun 2019. [Online]. Available: https://www.varnumlaw.com/insights/how-pfas-contamination-is-affecting-area-property-values/

[22] J. M. Liddie, L. A. Schaider, and E. M. Sunderland, “Sociodemographic factors are associated with the abundance of pfas sources and detection in u.s. community water systems,” *Environmental Science Technology*, vol. 57, no. 21, p. 7902–7912, May 2023.

[23] L. Padilla, “Understanding uncertainty on a map is harder than you think,” *interactions*, vol. 29, no. 3, p. 19–21, May 2022.

[24] “The pfas project lab.” [Online]. Available: https://pfasproject.com/

[25] “The situation – west plains water coalition.” [Online]. Available: https://westplainswater.org/the-situation/

[26] “Interactive map: Pfas contamination crisis: New data show 7,457 sites in 50 states.” [Online]. Available: https://www.ewg.org/interactive-maps/pfas_contamination/

Supplementary Material

A Dataset Processing

A.1 Image data

Initially, rasters were created for the entire contiguous U.S., including distance transform rasters that represent the distance to the nearest discharger of a specific type. The types of dischargers used in our data include aluminium forming, concentrated aquatic animal production, construction and development, copper forming, drinking water treatment, inorganic chemicals manufacturing, landfills, metal finishing, nonferrous metal forming, oil and gas extraction, petroleum refining, pharmaceutical manufacturing, pulp, paper, and paperboard, steam electric power generating, timber products processing, waste combustors, airports, military bases, and Aqueous Film Forming Foam (AFF) spills data as well.

Additionally, we created a raster combining slope data with watershed boundaries to account for potential upstream and downstream flows. For each sample point, we extracted 512x512 pixel patches for each of these rasters, covering approximately 236 km² area, centered around the sample point. These individual raster patches were then combined to an image with multiple channels.

B Training configuration

For the tabular dataset, the training set comprised 80% of the data, while the test set included 20%. The spatially stratified split involved assigning certain U.S. States to the training set and others to the test set, ensuring that the geographic distribution of the data was preserved. For the image dataset, patches were split into 70% for training and 30% for testing, ensuring no geographic overlap of patches between the sets. The UNet and Prithvi models were subsequently trained for 1000 epochs using the bootstrapping cross-entropy loss function

Training also utilized a custom learning rate scheduler, starting with a warmup phase and transitioning to polynomial decay, which helped stabilize the learning process.