FOCUS: A NOISE-AWARE GEOSPATIAL LEARNING FRAMEWORK FOR PFAS CONTAMINATION MAPPING

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

Alexa Friedman
Environmental Working Group
Washington, DC, USA
alexa.friedman@ewg.org

Sydney Evans
Environmental Working Group
Washington, DC, USA
sydney.evans@ewg.org

Rachel Klein
University of Michigan
Ann Arbor, MI, USA
raklein@umich.edu

Runzi Wang
University of California, Davis
Davis, CA, USA
mrwang@ucdavis.edu

Katherine E. Manz
University of Michigan
Ann Arbor, MI, USA
katmanz@umich.edu

Kaley Beins
Environmental Working Group
Washington, DC, USA
kaley.beins@ewg.org

David Q. Andrews
Environmental Working Group
Washington, DC, USA
dandrews@ewg.org

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

ABSTRACT

Per- and polyfluoroalkyl substances (PFAS) are persistent environmental contaminants with significant public-health impacts, yet large-scale monitoring remains severely limited due to the high cost and logistical challenges of field sampling, and the difficulty of simulating their spread. As a result, scientific understanding of PFAS transport in surface waters is incomplete. At the same time, rich geospatial and satellite-derived data describing land cover, hydrology, and industrial activity are widely available, creating an opportunity for AI to integrate sparse observations with large-scale environmental context. We introduce FOCUS, a geospatial deep learning framework for PFAS contamination mapping that learns from sparse point measurements propagated over satellite-based raster data while explicitly accounting for the resulting label noise. Rather than assuming known governing equations, FOCUS incorporates priors derived from hydrological connectivity, land cover, source proximity, and sampling distance to model uncertainty in supervision. These priors are integrated into a principled noise-aware loss, yielding a robust training objective under label noise. Across extensive ablations, robustness analyses, and real-world validation, FOCUS consistently outperforms baselines including sparse segmentation, Kriging, and pollutant transport simulations, while preserving spatial coherence and scalability over large regions. Our results demonstrate how AI can support environmental science by combining large-scale geospatial data with sparse, uncertain measurements to enable reliable PFAS contamination screening in the absence of complete physical models.

1 INTRODUCTION

PFAS or per- and polyfluoroalkyl substances are persistent "forever chemicals" widely used in industrial and consumer products such as non-stick cookware, waterproof textiles, and firefighting foams National Institute of Environmental Health Sciences (n.d.). Unfortunately, these substances resist degradation and accumulate in water Langenbach & Wilson (2021); Schroeder et al. (2021), soil Crone et al. (2019), and living organisms Environmental Working Group (2023), with 97% of Americans exhibiting detectable levels in their blood CDC (2024). Such pervasive exposure is linked to severe health risks including cancers, liver damage, and developmental disorders Manz (2024). With the high cost of measuring PFAS concentrations in the environment, which leads to sparse ground truth data, we are left uncertain about targeted remediation to best protect health.

Artificial intelligence (AI) presents a promising avenue for addressing complex health and environmental issues, including identifying PFAS hotspots. In various domains, AI has proven effective at automating labor-intensive tasks and synthesizing domain expertise, whether in land cover mapping Robinson et al. (2020), agricultural optimization Kerner et al. (2024), or disease prediction Bondi-Kelly et al. (2023). In the domain of PFAS contamination prediction, scientists have predominantly applied machine learning (ML) models such as random forests Breiman (2001) and XGBoost Chen & Guestrin (2016). For example, DeLuca et al. (2023) uses a random forest model to predict PFAS contamination via a tabular approach that aggregates environmental features within a 5 km buffer around sample points (Fig. 1). Building on these efforts, we frame PFAS prediction as a geospatial deep learning (DL) task. Our approach can take satellite image formats (i.e., rasters) as input directly, preserving spatial dependencies, reducing the need for extensive feature engineering and aggregation, and leading to more efficiency for ongoing monitoring efforts.

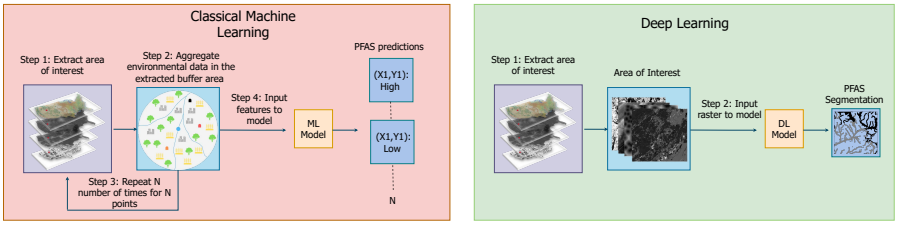

Figure 1: Left: Classical ML techniques aggregate surrounding pixel data to predict contamination at specific points (depicted as x, y coordinates); Right: DL methods process raster images directly to generate dense PFAS contamination maps in a single pass.

Our contributions in this work are summarized as follows:

- We formulate PFAS surface-water mapping as learning under structured, spatially dependent label noise induced by point-to-pixel supervision.
- We propose `FOCUS`[1], a principled noise-aware geospatial learning framework comprising a hydrology process-driven noise-robust objective that combines hydrological priors with focal loss, and provide a theoretical lower-bound guarantee showing that `FOCUS` yields a valid surrogate to the clean contamination likelihood under asymmetric pixelwise noise.
- We construct pixelwise correctness priors from hydrological flow, industrial discharge proximity, land cover, and sampling distance, linking environmental processes to robust learning.
- We evaluate on real PFAS field data across the United States and benchmark against geostatistical methods, pollutant transport simulations, and tabular ML baselines, demonstrating improved predictive performance and scalability.

We develop this framework together with an environmental nonprofit organization, academic researchers specializing in water quality models and environmental chemistry, and an advisory group with representatives from other nonprofits and agencies, ensuring that our approach is grounded in the latest expertise and real-world environmental concerns.

2 METHODOLOGY

PFAS mapping presents unique challenges: labels exist only at sparse monitoring locations, yet contamination spreads via hydrological and land-use processes. Expanding point-measurements into spatial masks introduces *structured, spatially-dependent label noise* rather than independent noise assumptions common in vision and NLP. To address this, we propose `FOCUS` as a noise-aware geospatial learning objective grounded in a latent contamination model and hydrology-informed pixel correctness priors. We perform segmentation on surface water area, classifying them as having PFAS concentrations above or below established health advisory thresholds using geospatial data.

[1]Code available at `https://github.com/jowariak/FOCUS`.

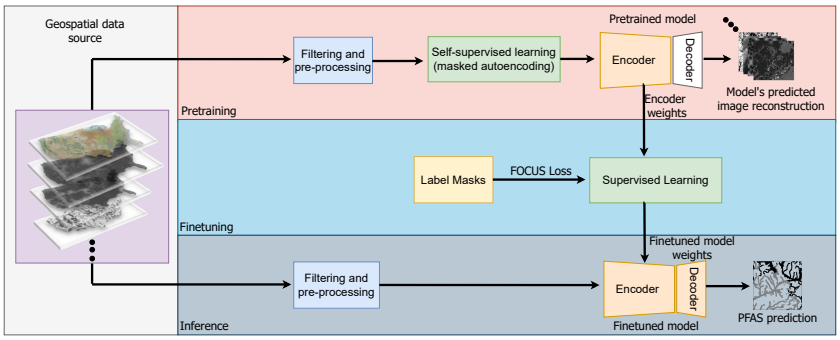

Figure 2: Model overview: high-level representation of FOCUS

2.1 Data, Labels, and Model Overview

We use PFAS measurements from U.S. government sources for fish tissue and surface water U.S. EPA, OW (2015); US EPA (2024); Michigan Department of Environment, Great Lakes, and Energy (2025; 2023), with fish samples providing nationwide coverage and water samples used for additional validation. Continuous PFAS concentrations are binarized into above/below health advisory thresholds using the Hazard Index U.S. EPA (2024). Based on a Prithvi-style masked autoencoder Blumenfeld (2023), we extract multi-channel raster patches centered at each sample location, encoding land cover U.S. Geological Survey (2024), hydrology Esri (2024), and distance-to-source features U.S. Environmental Protection Agency (n.d.), and pretrain the model on these geospatial rasters rather than using off-the-shelf satellite imagery. Because ground truth is only available at sparse point locations, we generate dense training masks by assigning the sample's label to all surface-water pixels within the patch, while non-water pixels are ignored. This point-to-pixel expansion introduces spatially structured label noise, since not all pixels in the patch are equally reliable. Our noise-aware objective is designed to account for this effect. Full dataset construction details are provided in the appendix.

2.2 Noise Model and FOCUS Loss

Expanding sparse point measurements to pixel-level supervision induces spatially structured, pixel-dependent label noise. Let x_i denote pixel features, $y_i \in \{0, 1\}$ the expanded (noisy) label, and $\pi_\theta(x_i)$ the latent contamination probability. We model this as asymmetric flip noise with pixel-specific rate $\eta_i < \frac{1}{2}$, yielding

$$\Pr(y_i = 1 \mid x_i) = \eta_i + (1 - 2\eta_i)\, \pi_\theta(x_i),$$

We estimate a pixelwise confidence score $M_i \in [0, 1]$ from hydrological connectivity, land cover, source proximity, and sampling distance, with higher M_i indicating more reliable supervision (details in Appendix). We incorporate this into training by weighting focal loss Lin et al. (2018):

$$\mathcal{L}_{\text{FOCUS}} = \frac{1}{N} \sum_{i=1}^{N} M_i (1 - p_i)^\gamma \left[-y_i \log p_i - (1 - y_i) \log(1 - p_i) \right].$$

where p_i is the model's predicted probability, γ is the focal parameter, and N is the total number of pixels. Intuitively, M_i down-weights unreliable pixels while focal modulation emphasizes hard examples. We show that this objective forms a valid noise-aware surrogate that upper-bounds the noisy risk and serves as a lower-bound surrogate to the clean likelihood under pixelwise asymmetric noise. A formal statement and proof are provided in the appendix.

3 Experiments

3.1 Baselines

We compare FOCUS against representative baselines spanning process-based, geostatistical, and learning-based approaches: (i) a pollutant transport simulation model based on hydrological principles SWAT (n.d.), (ii) a Landsat-based Prithvi model using raw multispectral imagery U.S. Geological Survey (n.d.), (iii) FESTA, a sparse-segmentation method for point supervision Hua et al. (2022), (iv) Kriging, a geostatistical interpolation method Esri (n.d.), (v) a random forest following DeLuca et al.

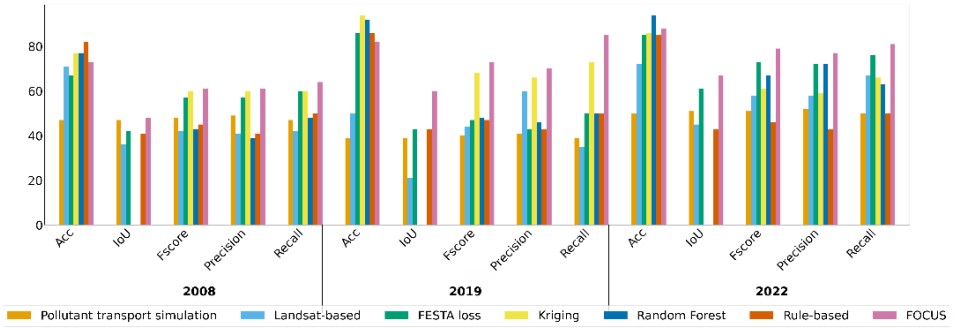

Figure 3: Performance comparison across methods and years. Results are averaged over three random seeds where applicable. IoU is omitted for methods that do not produce dense pixel-level predictions.

(2023) with features aggregated in a 5 km buffer, and (vi) a rule-based heuristic that thresholds our environmental priors to yield binary labels. Implementation details are provided in the appendix.

3.2 Training Configuration

We train and evaluate models separately per year using 80/20 geographically disjoint train/test splits, allowing us to assess **the model's ability to fill spatial data gaps, which is also a critical stakeholder need, rather than forecast across time.** Full training details are provided in the appendix. **Ablations:** We evaluate the impact of noise-aware weighting (reducing to standard focal loss when removed) and the effect of raster patch size; full ablation results are reported in the appendix.

3.3 Metrics

We report accuracy, IoU, F-score, precision, and recall, **evaluated only at ground-truth sample locations**. Decision thresholds are selected via cross-validation on the training set to maximize F-score and applied to the held-out test set. We use $K = 5$ folds and sweep 200 evenly spaced thresholds in $[0, 1]$. Thresholds are selected independently for each method.

4 Results

4.1 FOCUS Outperforms Existing Baselines

Fig. 3 compares FOCUS against six baselines across years and metrics. **Overall, FOCUS achieves the most consistent performance, with improved precision–recall trade-offs in this highly imbalanced setting**. While several baselines achieve competitive accuracy, accuracy alone is insufficient in this highly imbalanced setting, making precision–recall trade-offs more informative. Kriging and FESTA perform competitively in some years, highlighting the value of spatial structure, while the Landsat-only model underperforms, suggesting that satellite features alone are insufficient. Despite acceptable accuracy, the random forest showed lower F-score, precision, and recall, likely due to limited spatial context from per-point feature aggregation. The pollutant transport simulation and rule-based baselines provide useful references but lag behind learning-based approaches.

4.2 Qualitative Analysis of Noise-Aware Predictions

Fig. 4 shows grayscale probability maps with red overlays indicating confident positive predictions in low-trust regions; while the focal baseline produces many such responses, FOCUS largely suppresses them, yielding conservative behavior under unreliable supervision.

5 Conclusion and Future Work

We introduced FOCUS, a geospatial deep learning framework with a principled noise-aware objective for spatially structured label noise arising from point-to-pixel supervision. By integrating hydrology-informed priors into training, FOCUS improves predictive performance while preserving spatial coherence, enabling large-scale PFAS contamination screening. Future work will explore uncertainty-aware modeling to guide targeted sampling and extend the framework to temporal data for lifelong

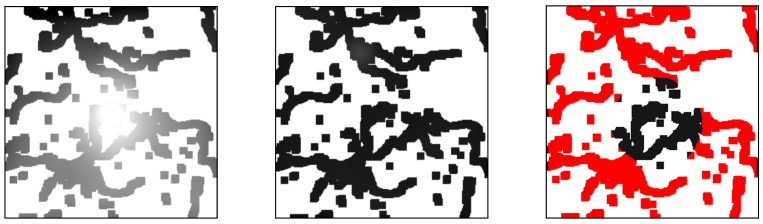

Figure 4: (a) Pixelwise confidence map Mi (brighter = higher trust). (b) FOCUS and (c) focal loss predictions shown as grayscale PFAS probabilities (brighter = higher). Red overlays indicate overconfident positives in low-trust regions ($p_i \geq 0.50,\ M_i \leq 0.30$). FOCUS suppresses such responses, while the focal baseline produces many.

learning. **We have also begun deploying FOCUS in a public-facing PFAS risk mapping interface, with a visualization provided in the appendix.**

REFERENCES

Josh Blumenfeld. Nasa and ibm openly release geospatial ai foundation model for nasa earth observation data — earthdata, August 2023. URL https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model.

Elizabeth Bondi-Kelly, Haipeng Chen, Christopher D. Golden, Nikhil Behari, and Milind Tambe. Predicting micronutrient deficiency with publicly available satellite data. *AI Magazine*, 44(1): 30–40, 2023. ISSN 2371-9621. doi: 10.1002/aaai.12080.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

CDC. Centers for disease control and prevention, August 2024. URL https://www.cdc.gov/index.html.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, August 2016. doi: 10.1145/2939672.2939785. URL http://arxiv.org/abs/1603.02754. arXiv:1603.02754 [cs].

Brian C. Crone, Thomas F. Speth, David G. Wahman, Samantha J. Smith, Gulizhaer Abulikemu, Eric J. Kleiner, and Jonathan G. Pressman. Occurrence of per- and polyfluoroalkyl substances (pfas) in source water and their treatment in drinking water. *Critical reviews in environmental science and technology*, 49(24):2359–2396, June 2019. ISSN 1064-3389. doi: 10.1080/10643389.2019.1614848.

Nicole M. DeLuca, Ashley Mullikin, Peter Brumm, Ana G. Rappold, and Elaine 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*, 57(37):14024–14035, sep 2023. ISSN 0013-936X. doi: 10.1021/acs.est.3c03670.

Environmental Working Group. Forever chemicals in freshwater fish: Mapping a growing environmental justice crisis, January 2023. URL https://www.ewg.org/news-insights/news/2023/01/forever-chemicals-freshwater-fish-mapping-growing-environmental-justice. Accessed: 2025-07-27.

Esri. Flow direction (spatial analyst) — arcgis pro documentation. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/flow-direction.htm, 2024. Accessed: 2025-07-17.

Esri. How kriging works — arcgis pro documentation, n.d. URL https://pro.arcgis.com/en/pro-app/latest/tool-reference/3d-analyst/how-kriging-works.htm. Accessed: 2025-07-27.

Yuansheng Hua, Diego Marcos, Lichao Mou, Xiao Xiang Zhu, and Devis Tuia. Semantic segmentation of remote sensing images with sparse annotations. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. ISSN 1545-598X, 1558-0571. doi: 10.1109/LGRS.2021.3051053. arXiv:2101.03492 [cs].

Hannah Kerner, Catherine Nakalembe, Adam Yang, Ivan Zvonkov, Ryan McWeeny, Gabriel Tseng, and Inbal Becker-Reshef. How accurate are existing land cover maps for agriculture in sub-saharan africa? *Scientific Data*, 11(1):486, May 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03306-z.

Blake Langenbach and Mark Wilson. Per- and polyfluoroalkyl substances (pfas): Significance and considerations within the regulatory framework of the usa. *International Journal of Environmental Research and Public Health*, 18(21):11142, October 2021. ISSN 1660-4601. doi: 10.3390/ijerph182111142.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. (arXiv:1708.02002), February 2018. doi: 10.48550/arXiv.1708.02002. URL http://arxiv.org/abs/1708.02002. arXiv:1708.02002 [cs].

Katherine E. Manz. Considerations for measurements of aggregate pfas exposure in precision environmental health. *ACS Measurement Science Au*, 4(6):620–628, December 2024. doi: 10.1021/acsmeasuresciau.4c00052.

Michigan Department of Environment, Great Lakes, and Energy. Pfas surface water sampling, June 2023. URL https://gis-egle.hub.arcgis.com/datasets/391cca4f364845829abcd5a92093c631_1/about. Accessed: 2025-07-27.

Michigan Department of Environment, Great Lakes, and Energy. Michigan fish contaminant monitoring sampling sites and select results (consolidated), January 2025. URL https://gis-egle.hub.arcgis.com/maps/d4bbb519842c44638eeb9a15461c441f/about. Accessed: 2025-07-27.

National Institute of Environmental Health Sciences. Perfluoroalkyl and polyfluoroalkyl substances (pfas). https://www.niehs.nih.gov/health/topics/agents/pfc, n.d. URL https://www.niehs.nih.gov/health/topics/agents/pfc. Accessed: 2025-07-27.

Caleb Robinson, Anthony Ortiz, Kolya Malkin, Blake Elias, Andi Peng, Dan Morris, Bistra Dilkina, and Nebojsa Jojic. Human-machine collaboration for fast land cover mapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(0303):2509–2517, April 2020. ISSN 2374-3468. doi: 10.1609/aaai.v34i03.5633.

Tim Schroeder, David Bond, and Janet Foley. Pfas soil and groundwater contamination via industrial airborne emission and land deposition in sw vermont and eastern new york state, usa. *Environmental Science. Processes & Impacts*, 23(2):291–301, March 2021. ISSN 2050-7895. doi: 10.1039/d0em00427h.

SWAT. Soil & water assessment tool, n.d. URL https://swat.tamu.edu/.

U.S. Environmental Protection Agency. Water pollution search — echo — us epa, n.d. URL https://echo.epa.gov/trends/loading-tool/water-pollution-search.

U.S. EPA. Contaminants to monitor in fish and shellfish advisory programs: Compilation of peer review-related information. 2024.

OW US EPA. 2022 national lakes assessment - fish tissue study, August 2024. URL https://www.epa.gov/choose-fish-and-shellfish-wisely/2022-national-lakes-assessment-fish-tissue-study.

U.S. EPA, OW. National rivers and streams assessment, June 2015. URL https://www.epa.gov/national-aquatic-resource-surveys/nrsa.

U.S. Geological Survey. National land cover database — u.s. geological survey, 2024. URL https://www.usgs.gov/centers/eros/science/annual-national-land-cover-database.

U.S. Geological Survey. Landsat 7 level 2, collection 2, tier 1 — earth engine data catalog, n.d. URL https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C02_T1_L2.

6 APPENDIX

To help navigate the appendix, Table 1 provides an overview linking each appendix section to the corresponding sections in the main paper where they are mentioned.

Table 1: Mapping between appendix sections and corresponding references in the main paper.

Appendix Subsection	**Referenced in Main Paper Section(s)**
A. Hydrology-Informed Pixel Confidence M_i Details	Section 2.2
B. Baselines	Section 3.1
C. Kriging	Section 3.1
D. Pollutant Transport Simulation	Section 3.1
E. Dataset	Section 2.1
F. FOCUS Per-Class Results	-
G. Statistical Significance of Performance Gains	-
H. Distance-to-Industry Rasters	Section 2.1
I. Ablation Results for Image Resolution	Section 3.2
J. Hyperparameter Tuning and Selection	Section 3.2
K. FOCUS Loss Derivation and Assumptions	Section 2.2
L. Web Map Interface	Section 5
M. Time Efficiency	-
N. FOCUS Noise Masks Improve Performance	Section 3.2
O. FOCUS has Robust Results	-
P. Model	-

A HYDROLOGY-INFORMED PIXEL CONFIDENCE M_i DETAILS

we generate a domain expert-informed probability of the label correctness for each surface water pixel i:

$$p_{\text{final}} = \alpha_1 \cdot p_{\text{dischargers}} + \alpha_2 \cdot p_{\text{landcover}} + \alpha_3 \cdot p_{\text{sample\_dist}} + \alpha_4 \cdot p_{\text{downstream}} \tag{1}$$

where $p_{\text{dischargers}}$, $p_{\text{landcover}}$, $p_{\text{sample\_dist}}$, and $p_{\text{downstream}}$ represent probabilities derived from (i) proximity to PFAS dischargers, (ii) land cover type, (iii) distance to other sample points, and (iv) downstream flow, respectively. In detail:

- **Proximity to PFAS Dischargers ($p_{\text{dischargers}}$)** Proximity to known PFAS dischargers is a strong indicator of contamination risk, as shown in scientific analyses [2]. If a pixel's label is 1, we apply an exponential decay function of distance, assigning higher probabilities to pixels closer to known dischargers. For pixels labeled 0, we invert this logic so that being near a discharger lowers confidence in the 0 label.
- **Land Cover Type ($p_{\text{landcover}}$)** For pixels labeled 1, surrounding urban or built-up areas received higher probabilities, consistent with research linking PFAS to industrial zones [3]. Conversely, for 0-labeled pixels, surrounding undeveloped or forested land cover increased confidence in low contamination.
- **Proximity to Other Sample Points ($p_{\text{sample\_dist}}$)** Pixels closer to ground truth sample points were assigned higher probabilities using an exponential decay function based on distance to each sample point in the patch.
- **Downstream Flow ($p_{\text{downstream}}$)** PFAS may also travel downstream in aquatic systems, making hydrologic flow another critical factor in contamination spread [4]. If a pixel lies

[2] PFAS dischargers
[3] Landcover
[4] Downstream Flow

downstream of a sample point labeled 1, its probability for label 1 increases, and similarly for label 0. This flow direction channel was generated using ArcGIS Pro [5] based on digital elevation models (DEMs) [6].

We perform an ablation study on the 2008 PFAS dataset using a grid search over 24 noise-mask configurations derived from four inputs: dischargers, land cover, flow direction, and distance to sample points. We vary their relative weights and evaluate segmentation performance across multiple metrics. The best configuration assigns 40% to dischargers, 20% to land cover, 10% to sample distance, and 30% to flow direction. These results align with domain expert guidance that emphasizes the importance of industrial activity in driving contamination patterns. Full performance results for all configurations are reported in Table 2, demonstrating that our selected weighting scheme most effectively balances the different environmental signals for robust PFAS contamination prediction. Pixels at ground truth sample points labeled 1 are assigned a probability of 1, reflecting complete confidence. However, for sample points labeled 0, we account for additional uncertainty due to certain PFAS compounds having method detection limits (MDLs) that exceed advisory thresholds[7].

Table 2: Performance across all noise mask weight configurations using the 2008 PFAS dataset.

Dischargers (%)	Landcover (%)	Flow Dir. (%)	Sample Dist. (%)	IoU (%)	F-score (%)	Precision (%)	Recall (%)	Accuracy (%)
40	**20**	**30**	**10**	**48**	**61**	**61**	**64**	**73**
40	30	20	10	44	55	58	52	70
40	10	30	20	46	59	60	56	72
40	10	20	30	42	53	55	51	69
40	30	10	20	45	57	59	54	71
40	20	10	30	43	54	57	52	70
30	40	10	20	46	58	60	55	72
30	40	20	10	43	56	57	53	68
30	10	40	20	44	55	56	53	69
30	10	20	40	45	57	58	55	70
30	20	40	40	47	62	60	58	72
30	20	10	40	45	58	59	56	71
30	10	30	30	41	52	54	49	69
10	40	30	20	47	60	60	58	72
10	40	20	40	44	56	57	54	69
10	30	40	20	42	53	55	51	70
10	30	20	40	43	55	56	52	68
10	20	40	10	46	59	60	57	72
10	20	30	40	44	56	57	53	70
20	40	30	40	45	58	59	56	69
20	40	10	30	46	59	60	57	72
20	30	40	10	44	57	58	55	69
20	30	20	30	43	56	57	54	68
20	10	40	30	40	52	54	50	66

B BASELINES

To contextualize our proposed framework for PFAS contamination prediction, we introduce several baseline approaches commonly employed in the field:

- **Pollutant transport simulation:** A scientific, process-based approach that models contaminant movement using hydrological principles [8].
- **Landsat-based method:** Uses the original Prithvi weights to predict PFAS from raw satellite imagery (Landsat 7 multispectral [9]).
- **FESTA loss approach:** An alternative state-of-the-art technique designed for sparse segmentation using sparse point data [10].
- **Kriging:** A geostatistical interpolation technique [11] that predicts contamination based on spatial relationships between points. Further details on implementation in appendix.

[5] ArcGIS Pro

[6] DEM

[7] In such cases, a compound may go undetected despite being present at a high-risk level. To address this, the corresponding noise mask value is down-weighted in proportion to the ratio of the advisory threshold to the MDL

[8] Pollutant transport

[9] Landsat-based

[10] FESTA

[11] Kriging

- **Random forest:** Closely following the methodology of [12], this model aggregates environmental features within a 5 km buffer around each sample (e.g., calculating the percentage of various land cover types) to predict contamination at individual points.
- **Rule-based approach:** Assigns contamination labels based on predefined environmental heuristics only, replicating our noise mask approach to compute a contamination probability (but without ($p_{\text{sample\_dist}}$)), which is then thresholded to yield binary labels.

C KRIGING

We employed Kriging as a spatial interpolation technique to estimate PFAS. First, we examined the spatial structure of our data by calculating an empirical semivariogram. Using latitude and longitude coordinates alongside the binary target variable, we computed pairwise distances and differences to produce a semivariogram plot, enabling us to observe how variance changed with increasing distance. This step informed the selection of appropriate variogram models (e.g., linear or spherical) for Kriging. Based on this analysis, we selected a spherical variogram model, which best captured the spatial dependencies observed in our data.

Next, we split the data on a state-by-state basis, ensuring geographic diversity between the training and testing subsets. We explored both Ordinary Kriging (assuming a stationary mean across the study area) and Universal Kriging (introducing a drift term to account for regional linear trends), but the two approaches produced comparable results on our dataset. After training each Kriging model on the geographic regions in the training set, we predicted PFAS presence at test set locations. Since we treated PFAS presence as a binary outcome, the resulting continuous Kriging predictions were thresholded at 0.5. We then computed standard classification metrics (such as precision, recall, and F1-score) to evaluate each model's ability to distinguish between high contamination versus low contamination sites.

While Kriging provided a practical means of interpolating PFAS presence, its reliance on variogram assumptions and spatial stationarity can limit accuracy in areas with highly heterogeneous conditions.

D POLLUTANT TRANSPORT SIMULATION

Instead of running a simulation like SWAT directly, which is designed for general watershed hydrology modeling but lacks established parameterization for PFAS transport, we implemented a streamlined Python-based workflow more tailored to PFAS-related surface water contamination. This approach allowed us to integrate domain-relevant features without relying on assumptions unsupported by current PFAS science.

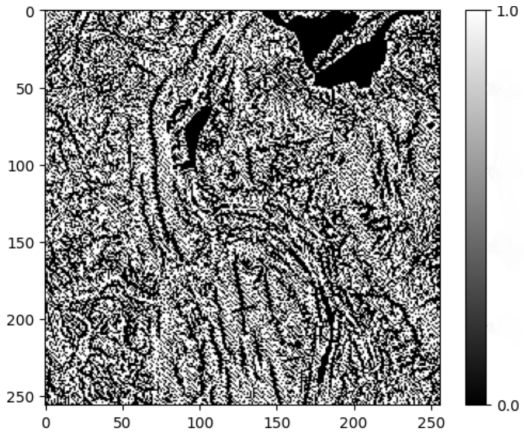

Figure 5: Example binary output from the pollutant transport simulation: 0 represents low contamination, and 1 represents high contamination. The simulation approximates the distribution of PFAS contamination based on hydrological and environmental parameters, providing a practical alternative to full-scale SWAT simulations.

[12] Random forest

First, we assign initial pollutant concentrations in each patch by setting cells flagged as "dischargers" to a high baseline value (e.g., 100), while land cover–based default concentrations provide moderately-low to low initial contamination levels for other cells. We then construct a Hydrologic Response Unit (HRU) parameter table that assigns infiltration and runoff values based on a cell's land cover, soil type, and slope. Using these parameters, each patch's land cover and soil rasters define infiltration/runoff ratios for every grid cell. We incorporate two key rasters; flow direction and flow accumulation, both exported from System for Automated Geoscientific Analyses (SAGA); a GIS and geospatial analysis tool focused on geospatial data processing, analysis, and visualization, to specify how water and pollutant mass move downstream. In each iteration, a fraction of the pollutant mass in each cell is transferred to its downstream neighbor according to the cell's infiltration/runoff factors, guided by the flow direction raster and scaled by the flow accumulation values. Repeating this process causes pollutant mass to concentrate in lower-lying or higher-flow cells, thereby modeling how contamination evolves over time within the patch.

After the simulation converges, we produce a final pollutant concentration raster, thresholded by the median value of concentrations across all patches for the year to generate a binary contamination map, as illustrated in Fig. 5. The binary outputs, with values of 0 and 1 representing low and high contamination respectively, visualizes the simulation's predictions of contamination distribution. These outputs are compared against test set patches containing actual observed PFAS presence *in surface water* to evaluate how accurately the simulation captures observed contamination patterns. Although this standalone approach remains computationally non-trivial, it is more tractable for batch processing of multiple patches than running a full SWAT project repeatedly. Consequently, it provides a practical baseline for pollutant transport modeling within our broader framework, allowing us to compare simulated outputs to both real-world data and other modeling approaches.

E Dataset

The dataset was curated to ensure that patches included in the training and testing sets were geographically disjoint, avoiding any overlap. In cases where two or more patches overlapped, as illustrated in Fig. 7, all such patches were assigned to the same set (either training or testing). This ensured that no part of a test patch had been seen during training, preserving the integrity of the evaluation.

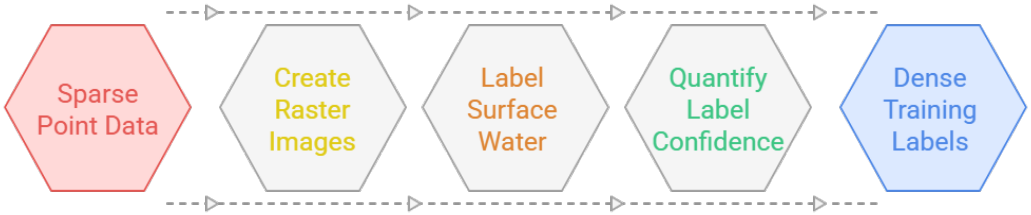

Figure 6: Overview of dataset curation pipeline

To contextualize our data preparation pipeline, Fig. 6 illustrates how ground truth samples are integrated with geospatial data to create the final dataset.

E.1 Ground Truth Samples

We use a combination of datasets and health advisories from government agencies in the United States, for both water and fish tissue samples. These are summarized in Table 3

We use both fish and water samples: fish reflect longer-term accumulation relevant to exposure risk, while water samples capture localized contamination at a given time [13]. Our analysis primarily uses fish data due to the nationally representative extent. Furthermore, PFAS concentrations are measured as continuous values in these samples, with individual measurements for different species of PFAS. We binarize PFAS concentration measurements into above (1) or below (0) health advisory thresholds, such as the EPA's Fish and Shellfish Advisory Program health thresholds [14]. In particular, we use the cumulative metric called the Hazard Index, which assesses combined risk from individual

[13] EWG study
[14] Advisory Program

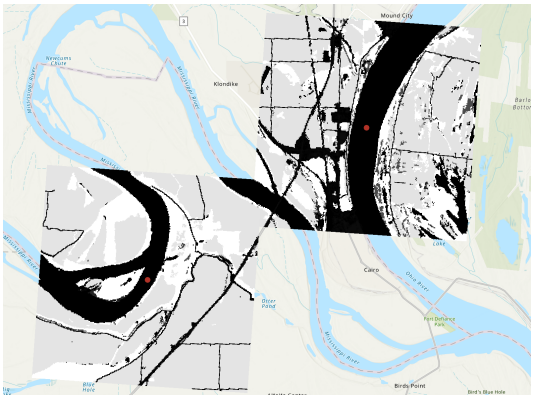

Figure 7: Example of overlapping patch assignment. The figure illustrates two slightly overlapping patches overlaid on the U.S. map, each containing distinct sample points. To maintain disjoint training and testing sets, both patches are assigned to the same set (either training or testing), ensuring no part of a test patch overlaps with any patch seen during training.

Table 3: Datasets used for training, testing, and real-world validation, including total size and class balance.

Type	Use	Data Source	Total Size	% Pos / Neg
Fish	Train/Test	EPA[a]	866	89.5% / 10.5%
Fish	Validation	MPART[b]	114	77.2% / 22.8%
Water	Train/Test	MPART[c]	293	66.9% / 33.1%
Water	Real-World Validation	Sampled by Authors	8	100% / 0%

[a] https://www.epa.gov/national-aquatic-resource-surveys/nrsa; https://www.epa.gov/choose-fish-and-shellfish-wisely/2022-national-lakes-assessment-fish-tissue-study
[b] https://gis-egle.hub.arcgis.com/maps/d4bbb519842c44638eeb9a15461c441f/about
[c] https://gis-egle.hub.arcgis.com/datasets/391cca4f364845829abcd5a92093c631_1/about

PFAS compounds by summing their concentration-to-threshold ratios [15] . Note that for fish train/test data, this yields 775 above-threshold (89.5%) and 91 below-threshold (10.5%) samples, reflecting a critical real-world imbalance. There is a significant class imbalance that we observe in the real world: unfortunately, a majority of water bodies have PFAS contamination that exceed health guidelines, as found in prior scientific work on PFAS prediction as well [16] . To affirm that there is in fact signal to our predictions, we provide class-specific results in Section I. For instance, we find approximately 80% F-score for class 0 for the year 2024, showing we are truly predicting class 0, and better than random. We also add a ternary classification in Section J to provide additional levels of risk.

E.2 Raster Image Generation

To spatially integrate the data for training and testing, multi-channel raster images were generated around each sample point (see Fig 8). Each image is a $P \times P$ pixels patch with 30-meter resolution, centered on the geographic coordinates of a sample point, where the optimal value of P is determined in Section M. Once trained, the model can make predictions at any location simply by extracting and processing the corresponding raster, which is done for real-world validation.

These raster images integrate several key features. First, we incorporate readily available rasters such as the National Land Cover Database (NLCD) [17], and flow direction rasters that capture hydrological connectivity using the D8 algorithm in ArcGIS Pro [18]. In addition, we include discharger location data

[15] Hazard Index

[16] See: *PFAS pollute 83% of U.S. waterways*, E&E News by Politico (2023). Available at: https://www.eenews.net/articles/pfas-pollute-83-of-u-s-waterways/

[17] NLCD

[18] Downstream flow

obtained from the U.S. EPA Enforcement and Compliance History Online (ECHO) [19] and convert them into distance rasters using a distance transform [20].

E.3 Ground Truth Masks

Our goal is to perform segmentation to predict PFAS contamination in surface water, but we have sparse point-level contamination measurements rather than dense labels across rasters. To generate dense training labels, every surface water pixel in a patch is labeled according to the point-level sample measurement in that patch (i.e., if the sample point indicates PFAS above the safety thresholds, all surface water pixels are labeled 1; if below, they are labeled 0). Non-surface water pixels are assigned a value of 2. This approach extends point-level data across the patch, assuming that nearby areas have similar contamination.

F FOCUS Per-Class Results

To better understand model behavior across classes, Table 4 reports per-class performance metrics for FOCUS across each year, including results on the 2024 MPART surface water dataset.

Table 4: Per-class performance of our model using FOCUS across four years (MPART surface water results at 2024). Results averaged over 3 random seeds.

	2008 (%)		**2019 (%)**		**2022 (%)**		**2024 (%)**	
Metric	**Class 0**	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Accuracy	73 ± 1	73 ± 1	82 ± 1	82 ± 1	88 ± 2	88 ± 2	79 ± 2	79 ± 1
IoU	25 ± 2	70 ± 2	40 ± 2	79 ± 2	47 ± 3	87 ± 2	67 ± 2	63 ± 2
F-score	40 ± 2	82 ± 2	57 ± 2	88 ± 2	64 ± 3	93 ± 2	80 ± 2	77 ± 2
Precision	33 ± 3	88 ± 2	42 ± 2	98 ± 2	60 ± 3	94 ± 1	75 ± 1	83 ± 2
Recall	50 ± 2	78 ± 2	89 ± 1	80 ± 1	69 ± 1	92 ± 2	86 ± 1	71 ± 1

Table 5: Performance of FOCUS under ternary classification (0: low, 1: medium, 2: high). Results averaged over 3 random seeds.

Metric	**2008 (%)**	**2019 (%)**	**2022 (%)**
Accuracy	55 ± 1	72 ± 1	65 ± 1
IoU	38 ± 1	48 ± 2	44 ± 2
F-score	54 ± 1	63 ± 1	61 ± 1
Precision	66 ± 2	70 ± 2	60 ± 2
Recall	65 ± 2	61 ± 1	62 ± 2

Results demonstrate signal, but are less strong compared to FOCUS binary classification performance, given the added complexity of the ternary task.

G Statistical Significance of Performance Gains

To evaluate the robustness of our improvements over the focal-only baseline, we conducted statistical significance testing using the Wilcoxon signed-rank test, a non-parametric paired test appropriate for comparing performance across random seeds.

For each year (2008, 2019, 2022), we collected the F1-scores from multiple runs (5 seeds per year) and compared the FOCUS results against focal-only using a two-sided Wilcoxon signed-rank test. As shown in Table 6, FOCUS achieved statistically significant improvements in all three years ($p = 0.03125$), despite the relatively small number of seeds. Notably, for every random seed, the F1-score achieved with FOCUS was consistently higher than the focal-only counterpart.

[19] PFAS dischargers

[20] Distance transform

These results confirm that the performance gains from our noise-aware loss are not due to chance, but are statistically reliable across different random initializations.

Table 6: Wilcoxon signed-rank test comparing F1-scores of FOCUS and focal-only across 5 random seeds for each year.

Year	Wilcoxon Statistic	p-value
2008	0.0	0.03125
2019	0.0	0.03125
2022	0.0	0.03125

H Distance-to-Industry Rasters

To capture potential industrial sources of PFAS contamination, we constructed distance rasters for each relevant facility type using data from the EPA's ECHO database. For each year (e.g., 2008, 2019, 2022), we identified all facilities that were active *on or before* that year, ensuring historical context was preserved. Facility types were grouped by industrial category (e.g., airports, landfills, chemical manufacturers), and a separate Euclidean distance raster was generated for each category, where pixel values represent the distance to the nearest facility of that type.

The total number of distance raster channels therefore varies by year, depending on which facility types were present in the ECHO dataset up to that point. For example, a given year may include distance rasters for 15 industry types, while another may include 41.

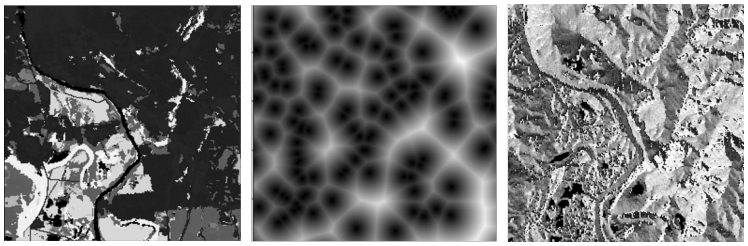

Figure 8: Example raster channels in an image: (left) land cover, (center) distances from chemical manufacturing industries, and (right) flow direction.

I Ablation Results for Image Resolution

Table 7 presents the results of our image resolution ablation study. We compare the model's performance when trained on 256×256 versus 512×512 input resolutions, without the use of noise-aware training. The smaller resolution consistently performed better, likely due to reduced label noise propagation across larger contexts. We did not evaluate patch sizes smaller than 256×256, as excessively small patches can compromise accuracy and increase variability [21].

Table 7: Comparison of performance between 256×256 and 512×512 image resolutions without noise-aware training. Bold values indicate the better result.

	2008 (%)		2019 (%)		2022 (%)	
Metric	**256**	**512**	**256**	**512**	**256**	**512**
Accuracy	**73 ± 2**	70 ± 3	**82 ± 1**	80 ± 2	**88 ± 3**	78 ± 3
IoU	**48 ± 3**	40 ± 2	**60 ± 2**	56 ± 2	**67 ± 3**	56 ± 2
F-score	**61 ± 3**	55 ± 3	**73 ± 1**	70 ± 2	**79 ± 3**	71 ± 3
Precision	**61 ± 2**	58 ± 3	**70 ± 2**	68 ± 2	**77 ± 3**	70 ± 3
Recall	**64 ± 2**	54 ± 2	**85 ± 1**	83 ± 2	**81 ± 1**	80 ± 2

[21] Ablation

J HYPERPARAMETER TUNING AND SELECTION

A batch size of 4 and a learning rate of 5×10^{-4} were adopted to preserve the stability of our pretrained backbone, minimizing the risk of large, destabilizing updates. In addition, we employed AdamW [22] as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We tuned the learning rate and batch size empirically by monitoring training stability. The final values were selected based on consistent convergence and favorable test-time metrics across years. Furthermore, we used a custom learning rate scheduler with an initial warmup phase followed by polynomial decay, which helped stabilize convergence and ensured that the model gradually adapted to the limited data. We tuned the learning rate and batch size empirically by monitoring training stability. Specifically, we experimented with learning rates in $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$ and batch sizes in $\{2, 4, 8\}$. We selected a batch size of 4 and a learning rate of 5×10^{-4} as they provided the best stability while preserving the benefits of our pretrained backbone, minimizing the risk of large, destabilizing updates. Notably, these final hyperparameter values are consistent with those adopted in the original Prithvi paper, further supporting their suitability in this context.

K FOCUS LOSS DERIVATION AND ASSUMPTIONS

Setup (latent asymmetric flip noise). Let $z_i \in \{0, 1\}$ be a latent clean label and $y_i \in \{0, 1\}$ the observed noisy label. Assume per-pixel flip noise with rate $\eta_i \in [0, \frac{1}{2})$ (can be asymmetric and spatially varying):

$$\Pr(y_i = z_i \mid x_i) = 1 - \eta_i, \qquad \Pr(y_i \neq z_i \mid x_i) = \eta_i.$$

Let $\pi_\theta(x_i) = \Pr_\theta(z_i = 1 \mid x_i)$ and define $\pi_\theta^{(y_i)}(x_i) = \pi_\theta(x_i)$ if $y_i = 1$ and $1 - \pi_\theta(x_i)$ if $y_i = 0$. Then the noisy likelihood is

$$\Pr_\theta(y_i \mid x_i) = \eta_i + (1 - 2\eta_i)\, \pi_\theta^{(y_i)}(x_i), \tag{2}$$

and the per-example noisy negative log-likelihood (NLL) is

$$\ell_{\text{noisy}}(\theta; x_i, y_i) = -\log\Big(\eta_i + (1 - 2\eta_i)\, \pi_\theta^{(y_i)}(x_i)\Big). \tag{3}$$

Lemma K.1 (Local weighted-CE upper bound)**.** *Fix $\eta_i \in [0, \frac{1}{2})$ and an anchor $p_{0i} \in (0, 1)$. Let $g_i(p) = -\log(\eta_i + (1 - 2\eta_i)p)$. Then for all p in a neighborhood of p_{0i},*

$$g_i(p) \;\leq\; g_i(p_{0i}) + g_i'(p_{0i})(p - p_{0i}) \;=\; -w_i \log p + C_i, \tag{4}$$

where

$$w_i = \frac{(1 - 2\eta_i)p_{0i}}{\eta_i + (1 - 2\eta_i)p_{0i}} \in [0, 1], \qquad C_i = g_i(p_{0i}) + w_i \log p_{0i}. \tag{5}$$

Interpretation. Lemma K.1 implies the noisy NLL behaves locally like a *weighted cross-entropy* term $-w_i \log p$, and the weight w_i decreases as η_i increases (labels become less informative).

Assumption on M_i (role of the environmental priors-informed confidence). We use $M_i \in [0, 1]$ as a practical surrogate for the unknown reliability weight. Concretely, we assume M_i is *aligned with label cleanliness*:

(Alignment assumption). We assume that $\mathbb{E}[M_i \mid \eta_i]$ is non-increasing in η_i, and that M_i approximates w_i up to a monotone re-scaling:

$$\mathbb{E}[M_i \mid \eta_i] \downarrow \text{ in } \eta_i, \qquad M_i \approx w_i. \tag{6}$$

Intuitively, M_i is larger near locations where supervision is more trustworthy (e.g., near sampling points or along plausible hydrologic pathways) and smaller in regions where point-to-pixel propagation is likely wrong.

We next show that multiplying by a bounded focal factor preserves a local surrogate form.

Lemma K.2 (Focal modulation preserves local surrogate form)**.** *Fix $\gamma \geq 0$ and anchor $p_{0i} \in (0, 1)$. There exist constants $A_i > 0$ and B_i such that for all p near p_{0i},*

$$-\log(\eta_i + (1 - 2\eta_i)p) \;\leq\; A_i\,(1 - p)^\gamma(-\log p) + B_i. \tag{7}$$

[22] AdamW

Interpretation. Near a fixed operating point p_{0i}, $(1-p)^\gamma$ is a positive bounded scalar, so focal loss is a locally rescaled cross-entropy and retains classification-calibrated behavior.

We now connect the above lemmas to the proposed objective.

Theorem K.3 (FOCUS as a noisy-NLL surrogate (local))**.** *Under the model in equation 2–equation 3 and $\gamma \geq 0$, for each example i there exist constants $\tilde{A}_i > 0, \tilde{B}_i$ such that for all p near p_{0i},*

$$\ell_{\text{noisy}}(\theta; x_i, y_i) \;\leq\; \tilde{A}_i\, M_i\, (1-p)^\gamma(-\log p) + \tilde{B}_i, \quad \textit{where } p = \pi_\theta^{(y_i)}(x_i).$$

Equivalently, minimizing

$$\mathcal{L}_{FOCUS}(x_i, y_i) = M_i\, (1 - \pi_\theta^{(y_i)}(x_i))^\gamma \left(-\log \pi_\theta^{(y_i)}(x_i) \right) \tag{8}$$

minimizes a valid local upper bound on ℓ_{noisy} (up to constants), and thus maximizes a corresponding local lower bound on the noisy log-likelihood.

Connection to expected noisy risk and calibration. Aggregating Theorem K.3 over pixels yields a surrogate for the expected noisy risk $\mathbb{E}[\ell_{\text{noisy}}]$. Since the modulating factor is non-negative and bounded for $p \in (0,1)$ and $\gamma \geq 0$, the objective remains classification-calibrated in the standard sense (locally equivalent to CE).

Edge cases and sanity checks.

- **Clean-label limit.** If $\eta_i \to 0$, then $w_i \to 1$ in equation 5. If additionally $M_i \to 1$, then FOCUS reduces to standard focal loss; with $\gamma = 0$ it reduces to CE.
- **Uninformative-label limit.** If $\eta_i \to \frac{1}{2}$, then $w_i \to 0$. Under the alignment assumption equation 6, M_i also becomes small, so FOCUS naturally down-weights (or effectively ignores) pixels whose supervision is uninformative.
- **Calibration of M_i.** In practice M_i need not equal the true w_i; it suffices that M_i ranks pixels by reliability (monotone alignment). This is precisely the role of our environmental priors-informed construction.

L FOCUS WEB MAP INTERFACE

To support real-world use and stakeholder engagement, we have developed a web-based, interactive PFAS risk mapping interface that visualizes model predictions at national scale (Fig. 9). The interface displays predicted PFAS contamination risk across surface waters and fish tissue using discrete risk categories (e.g., low, medium, high), enabling intuitive exploration by non-technical users.

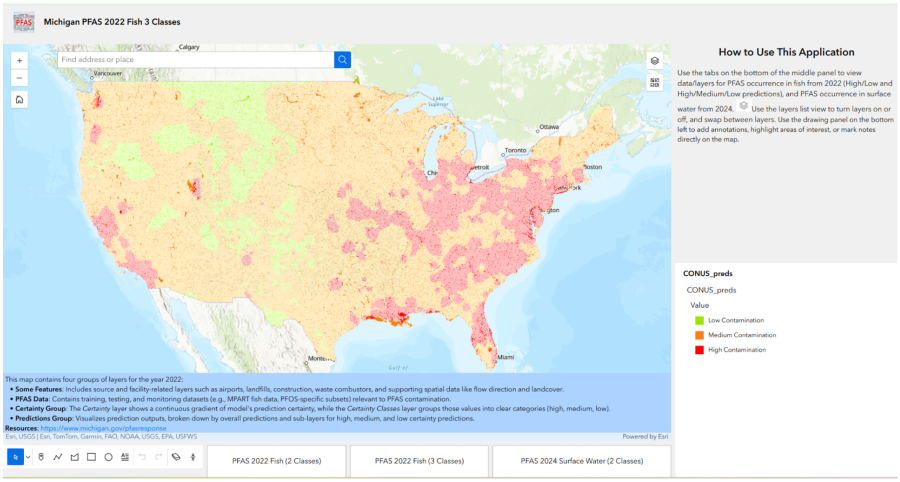

Figure 9: Our web-based PFAS risk mapping interface.

The map integrates FOCUS predictions with contextual geospatial layers, including surface water extents, land cover, hydrological features, and known discharge locations. Users can toggle layers,

zoom to regions of interest, and inspect spatial patterns across years and data modalities. The interface was developed in collaboration with policymakers and domain stakeholders to ensure interpretability, transparency, and responsible communication of model outputs.

This deployment illustrates how noise-aware geospatial predictions can be translated into actionable decision-support tools, while acknowledging uncertainty and the limitations of sparse environmental sampling.

M TIME EFFICIENCY

Following DeLuca et al. [23], we compare FOCUS against a random forest baseline using similar environmental features and datasets. Table 8 reports feature extraction and inference times over a 1 km^2 area and Northern Michigan (NM, $\sim$44,000 km^2).

The ML baseline requires aggregating features within a spatial buffer per point, resulting in significantly higher extraction costs at scale. In contrast, FOCUS processes full raster patches efficiently, reducing feature extraction time from days to hours over NM, while inference times remain comparable between methods. All times were averaged over 10 runs. Feature extraction was performed on an Intel Xeon Gold 6154 CPU (3.0 GHz), and inference used an AMD Ryzen Threadripper PRO 7985WX with an NVIDIA RTX 6000 Ada GPU.

Table 8: Comparison of performance for feature extraction and inference using random forest and FOCUS over 1 km^2 area and Northern Michigan (NM).

	Feature extraction		**Inference**	
	ML	**DL**	**ML**	**DL**
For 1 km^2 area	1.2 ± 0.2 mins	4.7 ± 0.5 sec	0.0003 ± 0.0002 sec	0.0005 ± 0.0001 sec
For NM	2 days	3.2 hrs $\pm$ 0.4 hrs	7.5 ± 1 sec	13 ± 2 sec

N FOCUS NOISE MASKS IMPROVE PERFORMANCE

Table 9: Ablation showing impact of noise-aware loss (FOCUS) vs. standard focal loss. Results averaged over 3 random seeds. Bold values denote the best result per year.

	2008 (%)		**2019 (%)**		**2022 (%)**	
Metric	**Focal only**	**FOCUS**	**Focal only**	**FOCUS**	**Focal only**	**FOCUS**
Accuracy	37 ± 2	$\mathbf{73 \pm 1}$	62 ± 1	$\mathbf{82 \pm 1}$	55 ± 3	$\mathbf{88 \pm 2}$
IoU	22 ± 3	$\mathbf{48 \pm 2}$	41 ± 2	$\mathbf{60 \pm 2}$	36 ± 3	$\mathbf{67 \pm 2}$
F-score	36 ± 3	$\mathbf{61 \pm 2}$	57 ± 1	$\mathbf{73 \pm 2}$	53 ± 3	$\mathbf{79 \pm 2}$
Precision	54 ± 2	$\mathbf{61 \pm 2}$	63 ± 2	$\mathbf{70 \pm 2}$	63 ± 3	$\mathbf{77 \pm 2}$
Recall	55 ± 2	$\mathbf{64 \pm 2}$	78 ± 1	$\mathbf{85 \pm 1}$	74 ± 1	$\mathbf{81 \pm 2}$

Table 9 evaluates the effect of incorporating noise masks in the FOCUS loss computation, compared to focal loss only (i.e., removing M_i from Equation 8). We find that the addition of noise masks consistently improves performance, though FOCUS may sometimes favor precision over sensitivity. This is due to emphasizing high-confidence pixels and down-weighting uncertain ones during training. Note that this ablation is conducted at the 256×256 resolution, which was selected based on the ablation study comparing it with 512×512 patches in Section M.

Table 10: Cross-state generalization results (mean ± std). Bold values denote the best result per year.

	2008 (%)		**2019 (%)**		**2022 (%)**	
Metric	**Focal Only**	**FOCUS**	**Focal Only**	**FOCUS**	**Focal Only**	**FOCUS**
Accuracy	58 ± 2	$\mathbf{70 \pm 1}$	76 ± 1	$\mathbf{89 \pm 1}$	78 ± 2	$\mathbf{92 \pm 1}$
IoU	34 ± 2	$\mathbf{48 \pm 1}$	40 ± 2	$\mathbf{55 \pm 1}$	56 ± 2	$\mathbf{75 \pm 1}$
F-score	47 ± 1	$\mathbf{63 \pm 1}$	46 ± 2	$\mathbf{61 \pm 1}$	68 ± 1	$\mathbf{85 \pm 1}$
Precision	49 ± 2	$\mathbf{62 \pm 1}$	49 ± 3	$\mathbf{59 \pm 2}$	69 ± 2	$\mathbf{82 \pm 1}$
Recall	49 ± 2	$\mathbf{72 \pm 1}$	50 ± 1	$\mathbf{71 \pm 2}$	85 ± 1	$\mathbf{91 \pm 1}$

O FOCUS HAS ROBUST RESULTS

To further assess the generalizability of our model to geographically unseen regions, we performed a spatial cross-validation experiment by partitioning data based on U.S. states in Table 10. Specifically, we selected three distinct sets of five states each to serve as held-out test sets. The remaining states in each case were used exclusively for training. This ensures that no spatial overlap exists between training and testing regions, mimicking the real-world scenario where models are deployed in previously unsampled areas. FOCUS again consistently outperforms focal loss alone across all years, highlighting its ability to generalize spatially to unseen regions and its robustness in distinguishing contaminated areas even under strong distribution shifts.

P MODEL

Our framework, FOCUS, is inspired by the Prithvi architecture Blumenfeld (2023) and employs a masked autoencoder (MAE) approach for pretraining (Fig. 2). Instead of relying on Prithvi's pretrained weights, we pretrain our model on derived geospatial data products (e.g., land cover and distance rasters), which offer more contextual information for capturing the environmental nuances critical to PFAS contamination prediction (see Section 4).

[23]Random forest