Environmental Science & Technology Letters

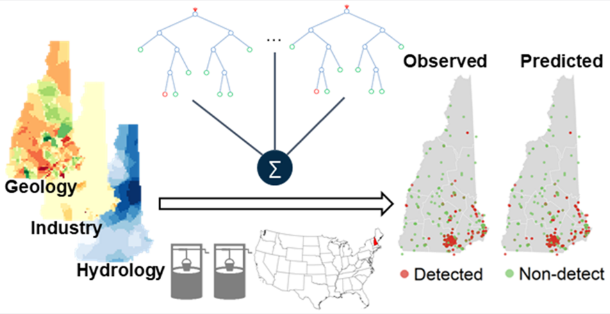
A Statistical Approach for Identifying Private Wells Susceptible to Perfluoroalkyl Substances (PFAS) Contamination

Xindi C. Hu,* Beverly Ge, Bridger J. Ruyle, Jennifer Sun, and Elsie M. Sunderland

Cite This: https://doi.org/10.1021/acs.estlett.1c00264 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Drinking water concentrations of per- and polyfluoroalkyl substances (PFAS) exceed provisional guidelines for millions of Americans. Data on private well PFAS concentrations are limited in many regions, and monitoring initiatives are costly and time-consuming. Here, we examine modeling approaches for predicting private wells likely to have detectable PFAS concentrations that could be used to prioritize monitoring initiatives. We used nationally available data on PFAS sources, and geologic, hydrologic and soil properties that affect PFAS transport as predictors, and trained and evaluated models using PFAS data ($n \sim 2300$ wells) collected by the state of New Hampshire between 2014 and 2017. Models were developed for the five most frequently detected PFAS: perfluoropentanoate, perfluorohexanoate, perfluoroheptanoate, perfluorooctanoate, and perfluorooctanesulfonate. Classification random forest models that allow nonlinearity in interactions among predictors performed the best (area under the receiver operating characteristics curve: 0.74−0.86). Point sources such as the plastics/rubber and textile industries accounted for the highest contribution to accuracy. Groundwater recharge, precipitation, soil sand content, and hydraulic conductivity were secondary predictors. Our study demonstrates the utility of machine learning models for predicting PFAS in private wells, and the classification random forest model based on nationally available predictors is readily extendable to other regions.



■ INTRODUCTION

Per- and polyfluoroalkyl substances (PFAS) are a diverse class of anthropogenic chemicals containing thousands of chemical structures that have been used by industry and in consumer products since the late 1950s.[1,2] Human exposure to PFAS has been linked to adverse health impacts such as kidney and testicular cancer, immunotoxicity, and endocrine disruption.[3−5] Probabilistic surveys show detectable levels of at least one PFAS in the serum of 98−99% of the United States (U.S.) general population.[6,7] Drinking water is widely recognized as the predominant human exposure source near PFAS contaminated sites.[8−10] National data on concentrations of six PFAS in large (serving >10000 individuals) public water supplies (PWS) have been collected by the U.S. Environmental Protection Agency's (U.S. EPA) third Unregulated Contaminant Monitoring Rule (UCMR3).[11] PFAS in smaller PWS will be comprehensively sampled in the next sampling cycle between 2023 and 2025 (UCMR5).[12] However, more than 40 million U.S. individuals obtain their drinking water from private wells and this represents a large remaining data gap for many states.[13] In UCMR3 data, the detection frequency for PFAS in water sourced from groundwater was more than twice that from surface water.[8]

Universal screening of PFAS in U.S. domestic wells would be both costly and logistically difficult to accomplish.[14] Modeling analyses can help prioritize testing in regions that are most likely to contain detectable concentrations of PFAS. For example, spatial modeling approaches have been successfully used to predict inorganic contaminant concentrations (especially arsenic and nitrate) in well water at the local, regional and national scales.[15−19] A Bayesian network model was applied to predict the occurrence of a novel PFAS, GenX, in private wells around a fluorochemical manufacturing facility.[20] These studies have identified potentially important predictors based on understanding of the sources and transport of chemical contaminants in groundwater.

Limited national and state-level data on the locations and magnitudes of PFAS point sources and environmental predictors present a challenge for the development of reliable models. Past work has shown that the probability of detecting PFAS in public water supplies can be linked to the number of wastewater treatment plants (WWTPs), industrial sites, military fire-training areas, and airports certified for aqueous film-forming foam use within a watershed.[8] Other work used

Received: April 9, 2021
Revised: May 7, 2021
Accepted: May 7, 2021

Table 1. Summary Statistics for PFAS Concentrations in New Hampshire Well Water

PFAS	n[a]	Uniform DL[b] (ng/L)	Percent Detectable (%)	Q1[c] (ng/L)	Median (ng/L)	Q3 (ng/L)	98th Percentile (ng/L)	Max (ng/L)
PFPeA	1617	5.0	29	3.5	3.5	6.0	92	7500
PFHxA	1725	8.0	29	5.7	5.7	10	110	17000
PFHpA	2253	5.0	24	3.5	3.5	5.0	65	9200
PFOA	2373	8.0	48	5.7	5.7	30	320	52000
PFOS	2376	5.0	17	3.5	3.5	3.5	75	11000
PFAS	2379	31	40	22	24	58	670	86000

[a] *n* = Number of unique wells sampled. [b] Uniform detection limit (DL): this value was calculated as the 98th percentile of all DL reported across batches after extreme outliers were removed. We removed 254 (1.6%) of total samples where the DL was more than five times the median DL across batches. [c] When calculating summary statistics (Q1:25th percentile, median, Q3:75th percentile, 98th percentile), samples below detected were imputed with uniform DL divided by square root of 2.

Facility Registration Service (FRS) codes to identify diverse potential PFAS sources within watersheds in the Northeastern U.S. and to explain differences in observed PFAS profiles in surface waters.[21] Near point sources, local groundwater fluxes are often poorly constrained but other environmental predictors such as precipitation, groundwater recharge, bedrock geology, and soil type that are relevant for hydraulic conductivity and groundwater flow are available and can used for spatial modeling.[22,23]

Here, we examine the performance of different machine learning approaches for predicting PFAS in private wells based on a case study of PFAS data collected by the New Hampshire (NH) Department of Environmental Services[24] between 2014 and 2017. All of the predictors used in the modeling presented are available at the national scale, enabling generalizability of the approach to other regions. We compare the performance of a relatively simple functional model form (logistic regression) to a classification random forest model. On the basis of results of this case study, we discuss the strengths and limitations of different modeling approaches and steps toward developing national scale models for predicting the locations where PFAS concentrations in private wells may exceed health-based thresholds.

■ MATERIALS AND METHODS

New Hampshire Domestic Well PFAS Data. Approximately half of NH's population (∼520000) obtains their drinking water from private wells.[25] The state sampled more than 2300 unique domestic wells (3900 individual samples) between 2014 and 2017, creating a useful data set for training and evaluating statistical models (Table 1, Supporting Information (SI) Figure S1, Section S1).[24] Sampling locations were prioritized based on proximity to known contaminated sites and as part of investigations into sites with potential PFAS uses. Low detection frequencies can create bias issues that interfere with the training of predictive models.[26] The sampling design that was weighted toward contaminated sites and point sources therefore created a more balanced data set for statistical model training.[27,28] A limitation of this design is that few wells were sampled in the vicinity of some known PFAS contamination sources such as airports that use aqueous film forming foams (AFFF) (n = 36 wells) and military bases (n = 51 wells), limiting their predictive power in the final model.

Concentrations of 35 PFAS were analyzed by several laboratories following EPA Method 537 or a modified version of Method 537 that includes isotope dilution.[29] We developed predictive models for the five PFAS that were most frequently detected in NH wells: perfluoropentanoate (PFPeA), perfluorohexanoate (PFHxA), perfluoroheptanoate (PFHpA), perfluorooctanoate (PFOA), and perfluorooctanesulfonate (PFOS). PFPeA, PFHxA, PFHpA, and PFOA were all detected in more than 20% of the water samples, and PFOS was detected in 17% of the samples (Table 1).

Point Sources and Environmental Predictors. Two categories of PFAS point sources were used as model predictors: confirmed and potential sources. Confirmed sources included known industrial point sources (one plastics manufacturer and one textile manufacturer), military sites contaminated by aqueous film forming foams (AFFF), airports certified for AFFF use, and wastewater treatment plants (WWTP).[8,10,30,31] Potential sources were identified following the approach used in prior work.[21,32] Specifically, we used the North American Industrial Classification System (NAICS) codes and the US EPA Facility Registry Service (FRS) codes that correspond to industries that are known to use and release PFAS such as the metal plating, petroleum manufacturing, and semiconductor industries, and textile mills (Table S2). The potential sources were lumped into a single category because they are more uncertain and have not been confirmed.

Since no data on the magnitudes of PFAS releases from various sources are available, we developed a unitless relative impact score for each point source (both confirmed and potential sources) based on hydrological distance to the well location, adapted from the method applied in prior work.[21,32] Specifically, we calculated an impact score as an exponential decay function with distance from the sampling site (i.e., $1/e^{d}$, where d (km) is the Haversine distance between the point source and well). Only industries with elevation above a well and within the same 12-digit Hydrological Unit Code were considered.[33] We used elevation as a proxy for hydrological flow direction because topography is known to control groundwater flow direction for most of NH.[34] Impacts from multiple individual point sources within the same group (e.g., WWTPs) were summed for each well (i.e., $\sum 1/e^{d}$).

Two major industrial PFAS sources in New Hampshire are known to release atmospheric PFAS emissions.[35] Other work suggests this is a plausible transport pathway for PFAS detected in private wells.[36] On the basis of air deposition modeling studies,[37,38] we estimated predominant impacts within a circular buffer with a 10-km radius. Our modeling analysis was insensitive to selection of a larger buffer size (30 km) more reflective of air emissions from a fluoropolymer manufacturing location (SI Figure S2).[39]

Environmental predictors were selected based on their potential relevance for transport of PFAS in groundwater. These included soil properties (e.g., percent clay, bulk density, organic carbon content),[17] precipitation, and groundwater

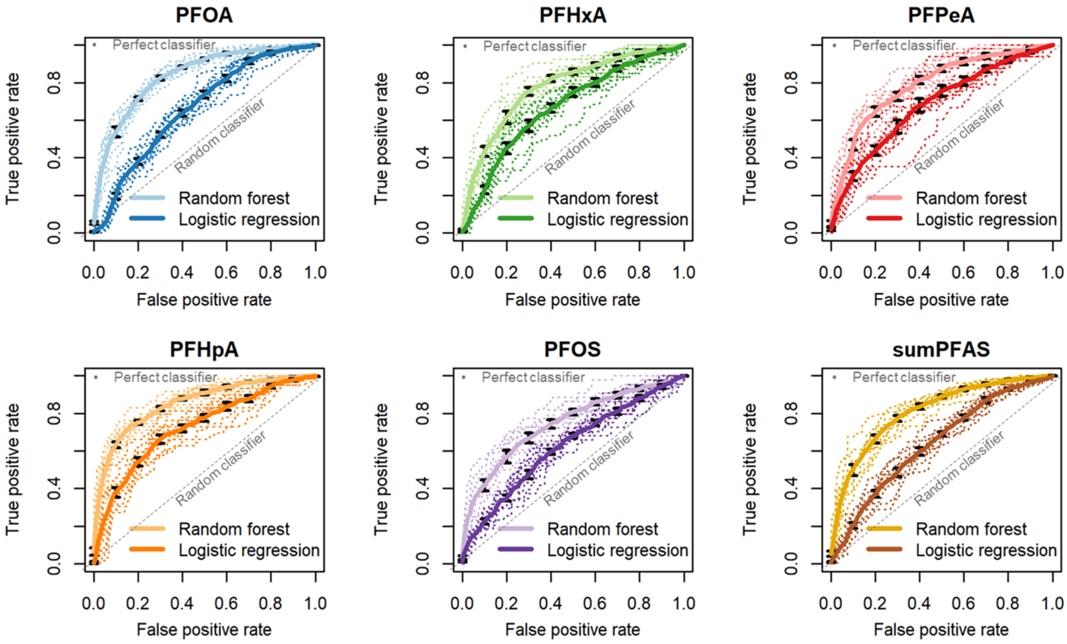


Figure 1. Receiver Operating Characteristic (ROC) curves for comparing the performance of classification random forest models (lighter lines) and logistic regression models (darker lines) for the five PFAS and the detection of any of the five ("sumPFAS"). The perfect classifier (0.0, 1.0) would be located in the upper left corner, while a random classifier would be the diagonal line going through (0.5, 0.5). Discrimination thresholds for the true positive rate (probability of correctly identifying a well with detectable PFAS) and probability of false alarm (false positive: incorrectly flagging a well as likely to have detectable PFAS when concentrations are below detection) can be tuned in each model. For screening the potential presence of PFAS in private wells, maximizing the true positive rate and collecting field samples from flagged locations would be most protective of public health.

recharge,[15] and geological factors that can affect groundwater flow (see SI, Section S2 and Table S3 for additional details).[40−42]

Statistical Models. For both the logistic regression and random forest classification models, the predictive threshold was set as the probability of PFAS concentrations exceeding a uniform detection limit inferred from the observations (Table 1).[15,43] Well water samples from NH were analyzed in different batches and across laboratories with variable detection limits (0.015−40 ng/L). Samples with a detection limit higher than five times the median detection limit (n = 254, 1.6% of samples) were omitted from our analysis. For each of the five frequently detected PFAS, we chose the threshold for the categorical models (detectable/nondetectable) as the upper 98th percentile of detection limits for all samples (see SI section S1 and Table S1 for details).

For the logistic regression (classification model), we removed highly collinear predictors (Spearman correlation coefficients greater than 0.7) before fitting the model. Residual multicollinearity problems were addressed by removing predictors with a large variance inflation factor (greater than 10). We performed stepwise selection of predictors by minimizing Akaike's Information Criteria (AIC). Model performance was evaluated by model discrimination (C-statistics) and the area under the Receiver Operating Characteristics curve (AUROC), which both range from 0.5 (no predictive power) to 1 (perfect predictive power).[44] The C-statistic indicates the probability a randomly selected well with a positive detection has a higher risk score than a well with a nondetect. True positive rate, true negative rate, false positive rate, and false negative rate were calculated during the stratified 10-fold cross-validation using the confusion matrix shown in Figure S3. A true positive is defined as a positive model prediction that is observed above detection limit. A true negative is defined a predicted nondetect that is also below detection in the measured data. A false negative is a modeled concentration that is below detection but the observed valued is detectable. False positives are modeled detectable concentrations that are below detection in the observations. We calculated the 95% confidence interval (CI) around the AUROC to show the uncertainty associated with model performance metrics. This was calculated by performing 10-fold stratified cross-validation, in which the observations were partitioned into ten subsets that maintained the ratio of wells with detectable and nondetectable concentrations.[45,46] We treated one subset as the test data (n ∼ 230) and trained a model on the remaining nine subsets (n ∼ 2070). This procedure was performed 10 times such that each subset was used once as the test data. We calculated standardized coefficients to assess the relative importance of predictor variables. Statistical analyses were conducted using *stats* package, *MASS* package, and *arm* package in R 4.0.0.[47]

The classification random forest model represents an ensemble of individual classification trees. These models tend to have higher prediction accuracy than individual tree-based methods.[48] Predictions on new data are obtained by the "majority vote" of all of the trees in the ensemble. Model hyperparameters influence the speed and quality of the machine learning and include the number of trees, node size, and number of variables randomly sampled at each split. We tuned hyperparameters using grid search to maximize predictive accuracy in the classification random forest. Similar to the logistic regression model, we calculated the AUROC and its 95% CI using stratified 10-fold cross-validation. Statistical analyses were conducted using *randomForest* package in R 4.0.0.[49] We also developed a continuous model

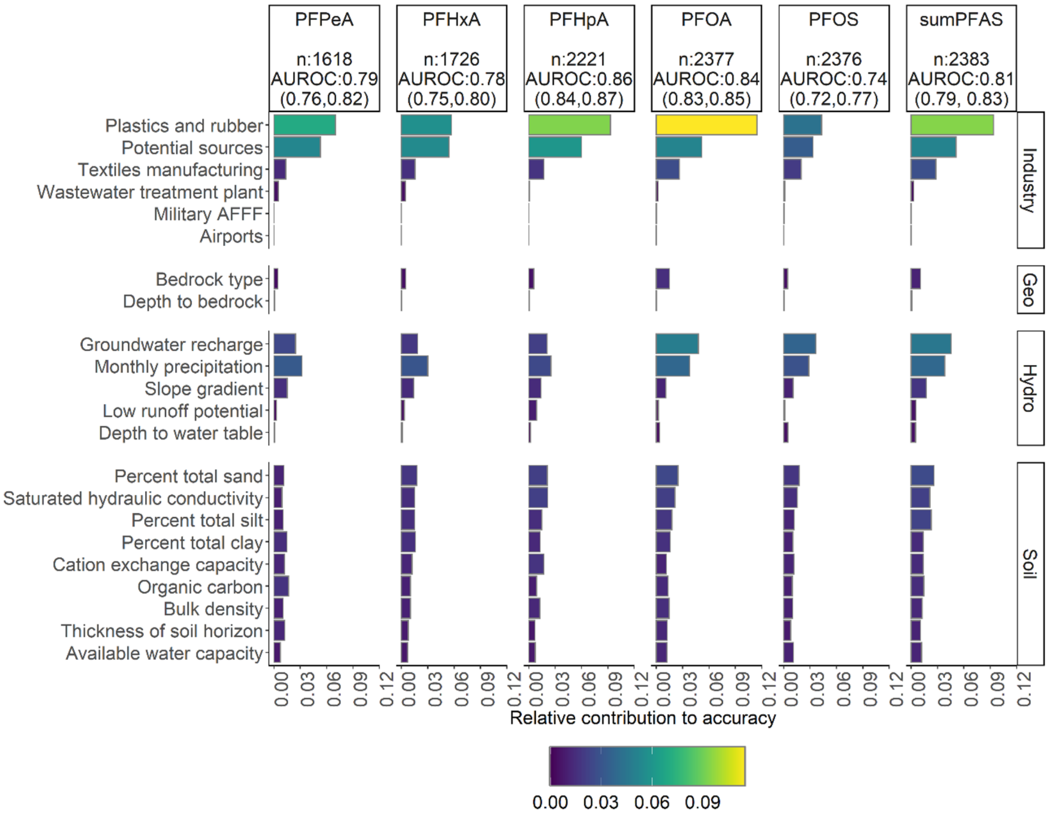


Figure 2. Relative contributions of all variables in classification random forest for five PFAS that maximized the overall prediction accuracy. The relative contribution of a variable was measured by the decrease in prediction accuracy if this variable was permuted. "sumPFAS" refers to the detection of any of the five PFAS. Headers provide performance information for the classification random forest models (n = sample size, Area under the Receiver Operating Characteristics curve (AUROC) indicates the classification model's performance). The mean AUROC and 95% confidence intervals (CI) were calculated by 10-fold cross-validation, where the entire data set was split into 10 folds and each fold was used as the testing set with the rest of the data used as the training set. The process was repeated for 10 times until each fold was used as the testing set.

(regression random forest) that is described in the SI Section S3 and Table S4. The R codes for data management and analyses in this article can be accessed at https://github.com/SunderlandLab/pfas_nh_model.

RESULTS AND DISCUSSION

PFAS Concentrations in New Hampshire Wells. Summed PFOA and PFOS concentrations exceeded the US EPA provisional lifetime advisory of 70 ng/L level for 14% of the wells and maximum concentrations of the sum of five PFAS were hundreds to thousands of times higher than the U.S. EPA health-advisory level and the state of NH's proposed maximum contaminant levels (MCLs): 12 ng/L PFOA, 15 ng/L PFOS, 18 ng/L PFHxS, 11 ng/L PFNA.[24] The highest concentrations of individual PFAS detected in NH were in the ug/L range (Table 1) and were mainly concentrated in the southern part of the state close to the two main industrial sources (SI Figure S1). Private wells in the northern parts of the state were more sparsely sampled but generally appear to have lower concentrations (see TOC art).

Model Evaluation and Comparison. For the classification models, the AUROC for the classification random forest was higher than the logistic regression for all five PFAS modeled and the detection of any of the five PFAS (sumPFAS) (Figure 1). Up to 13 significant predictors (p < 0.0001) were selected in the logistic regression models for the five individual PFAS and sumPFAS models (Table S5). The AUROC reflects the trade-off in model performance to optimize the true positive rate (detectable levels of PFAS predicted and observed) or the false positive rate (levels of PFAS predicted to be above detection but observed below detection) and ranges from 0.5 (no predictive power) to 1 (perfect predictive power). The best performing model (classification random forest) is the one with the largest integral (Figure 1).

The AUROC for the classification random forest model ranged from 0.74 (95% CI: 0.72, 0.77 for PFOS to 0.86 (95% CI: 0.84, 0.87) for PFHpA (Figure 2). Improvements in model performance expressed as AUROC relative to the logistic regression ranged from 0.1 for PFOS to 0.15 for PFOA. Tuning of hyperparameters in the classification random forest had a small impact on the AUROC (SI Table S6).

For a screening model, the false negative rate (predicting a well is below detection when it is above detection) is more important than false positives. Setting the false negative rate at 20%, the false positive rates for the four perfluoroalkyl carboxylic acids (PFCA, which refers to PFPeA, PFHxA, PFHpA, and PFOA) (29% to 38%) and PFOS (50%) in classification random forest are lower than those for the logistic regression (50% to 60%, Figure 1). Better performance of the classification random forest models for PFAS in private wells suggests they are most useful for prioritizing sampling and identifying susceptible regions.

Important Predictors of PFAS in Private Wells. PFAS point sources were the most important predictors in the classification random forest model for NH private wells (Figure 2). The top ranked predictor for all PFAS except

PFOS in terms of relative contribution to model accuracy was the plastics and rubber industry, followed by the potential PFAS source category identified by screening industry codes (Figure 2). For PFOS, point sources contribute relatively less to model accuracy than the four PFCA, likely because the confirmed sources included in the model such as a plastics manufacturer released relatively more PFCA than PFOS.[50,51] In addition, AFFF contaminated military bases and airports were not as strong predictors as might be expected, given the low number of wells (n = 51 and 36, respectively) sampled that could be affected by these sources in the training data set (Figure S4).

Soil properties (bottom panel of Figure 2) exert a greater influence on model accuracy for PFOA and PFOS compared to the shorter chain PFCA. This may reflect the propensity for longer chain PFAS to partition more strongly to soil organic carbon, thus allowing a greater influence of soil properties on transport.[52] Groundwater recharge and the plastics and rubber industry are particularly important for PFOA, likely due to a large releases and subsequent transport away from several well established point sources (e.g., manufacturing industry in Merrimack, NH).[53,54] For the shorter chain PFCA, monthly precipitation stands out as a relatively important contributor to model performance and may reflect an atmospheric contribution to some private wells, as noted in other regions of the Northeastern US.[36]

Toward a National Model for PFAS in Private Wells. A large number of domestic well users (43 million residents) across the U.S. lack access to well testing but may be exposed to PFAS.[55] Many states are developing monitoring programs for PFAS in private wells, and sampling is often skewed toward known industrial sources and other contaminated sites. Results of this work show preliminary data collected by states can be used to develop statistical models that help screen for PFAS occurrence. Classification random forest models, in particular, performed well (AUROC ~ 0.8) for identifying locations likely to have detectable PFAS concentrations in private wells in this NH case study.

Logistic regression has been used in previous modeling of groundwater contamination by inorganic contaminants because it can handle left-censored concentration data, it is straightforward to develop, and often produces a parsimonious model after variable selection (SI Section S4, Table S7).[17,55] However, logistic regression assumes a linear relationship between predictors and the log odds of detection, while in fact this relationship can be highly nonlinear.[56,57] Logistic regression also has limited capacity for handling potential interactions among predictors. Classification random forest models provide a preferred alternative because they do not contain assumptions related to data distribution, they are well-suited for handling nonlinear relationships between predictors and outcomes,[48] and they are less sensitive to collinearity among predictors.[58] A limitation of both logistic regression and classification random forest models is that they predict the probability of exceeding a threshold PFAS concentration rather than an absolute value. However, threshold concentrations may be of more interest to environmental regulators because model results can be compared directly with health advisory levels or water-quality standards.

This study shows the results of model evaluation and performance for the state of NH, but the approach for developing a classification random forest model for PFAS in private wells, using nationally available predictors, is easily extendable to other regions with available private well testing data for model construction and testing. The relative importance of predictors is likely to vary substantially across states, and correct identification of PFAS sources is extremely important for model performance. For NH, impacts from industrial sources were the most important contributors to model performance. Thus, the greatest improvements in model performance may be obtained by local ground-truthing exercises for the "potential sources" identified in this study that correctly identify additional confirmed PFAS sources, as well as the better characterization of the impacted areas around the potential contaminated sources. Overall, this study demonstrates the utility of machine learning models for screening private-wells likely to have detectable PFAS concentrations that could be prioritized for additional monitoring.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.1c00264.

(PDF)

■ AUTHOR INFORMATION

Corresponding Author

Xindi C. Hu − *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States; Mathematica, Inc., Oakland, California 94612-1475, United States; Department of Environmental Health, Harvard T.H Chan School of Public Health, Boston, Massachusetts 02115, United States;* orcid.org/0000-0002-4299-3931; Phone: 5102854675; Email: chu@mathematica-mpr.com

Authors

Beverly Ge − *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States*

Bridger J. Ruyle − *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States;* orcid.org/0000-0003-1941-4732

Jennifer Sun − *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States*

Elsie M. Sunderland − *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States; Department of Environmental Health, Harvard T.H Chan School of Public Health, Boston, Massachusetts 02115, United States;* orcid.org/0000-0003-0386-9548

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.1c00264

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the NIEHS Superfund Research Center grant P42ES02770. The authors thank Brandon Kernen from New Hampshire Department of Environmental Services for sharing data from the NH domestic well monitoring program. BG acknowledges support from the Harvard College Research Program.

■ REFERENCES

(1) Buck, R. C.; Franklin, J.; Berger, U.; Conder, J. M.; Cousins, I. T.; de Voogt, P.; Jensen, A. A.; Kannan, K.; Mabury, S. A.; van Leeuwen, S. P. Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. *Integr. Environ. Assess. Manage.* **2011**, *7* (4), 513−41.

(2) Organisation for Economic Co-operation Development, Toward a New Comprehensive Global Database of Per-and Polyfluoroalkyl Substances (PFASs): Summary Report on Updating the OECD 2007 List of per-and Polyfluoroalkyl Substances (PFASs). 2018.

(3) Rappazzo, K.; Coffman, E.; Hines, E. Exposure to perfluorinated alkyl substances and health outcomes in children: a systematic review of the epidemiologic literature. *Int. J. Environ. Res. Public Health* **2017**, *14* (7), 691.

(4) Vieira, V. M.; Hoffman, K.; Shin, H.-M.; Weinberg, J. M.; Webster, T. F.; Fletcher, T. Perfluorooctanoic acid exposure and cancer outcomes in a contaminated community: a geographic analysis. *Environ. Health Perspect.* **2013**, *121* (3), 318−323.

(5) Mogensen, U. B.; Grandjean, P.; Heilmann, C.; Nielsen, F.; Weihe, P.; Budtz-Jørgensen, E. Structural equation modeling of immunotoxicity associated with exposure to perfluorinated alkylates. *Environ. Health* **2015**, *14* (1), 47.

(6) Calafat, A. M.; Wong, L. Y.; Kuklenyik, Z.; Reidy, J. A.; Needham, L. L. Polyfluoroalkyl chemicals in the U.S. population: data from the National Health and Nutrition Examination Survey (NHANES) 2003−2004 and comparisons with NHANES 1999−2000. *Environ. Health Perspect.* **2007**, *115* (11), 1596−602.

(7) Lewis, R. C.; Johns, L. E.; Meeker, J. D. Serum Biomarkers of Exposure to Perfluoroalkyl Substances in Relation to Serum Testosterone and Measures of Thyroid Function among Adults and Adolescents from NHANES 2011−2012. *Int. J. Environ. Res. Public Health* **2015**, *12* (6), 6098−114.

(8) Hu, X. C.; Andrews, D. Q.; Lindstrom, A. B.; Bruton, T. A.; Schaider, L. A.; Grandjean, P.; Lohmann, R.; Carignan, C. C.; Blum, A.; Balan, S. A.; Higgins, C. P.; Sunderland, E. M. Detection of Poly- and Perfluoroalkyl Substances (PFASs) in U.S. Drinking Water Linked to Industrial Sites, Military Fire Training Areas, and Wastewater Treatment Plants. *Environ. Sci. Technol. Lett.* **2016**, *3* (10), 344−350.

(9) Li, Y.; Fletcher, T.; Mucs, D.; Scott, K.; Lindh, C. H.; Tallving, P.; Jakobsson, K. Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occup. Environ. Med.* **2018**, *75* (1), 46−51.

(10) Andrews, D. Q.; Naidenko, O. V. Population-Wide Exposure to Per- and Polyfluoroalkyl Substances from Drinking Water in the United States. *Environ. Sci. Technol. Lett.* **2020**, *7*, 931.

(11) U.S. Environmental Protection Agency Third Unregulated Contaminant Monitoring Rule. https://www.epa.gov/dwucmr/third-unregulated-contaminant-monitoring-rule (04−08),.

(12) U.S. Environmental Protection Agency Fifth Unregulated Contaminant Monitoring Rule. https://www.epa.gov/dwucmr/fifth-unregulated-contaminant-monitoring-rule (04−08),.

(13) Dieter, C. A.; Maupin, M. A.; Caldwell, R. R.; Harris, M. A.; Ivahnenko, T. I.; Lovelace, J. K.; Barber, N. L.; Linsey, K. S. *Estimated use of water in the United States in 2015*; 1441; Reston, VA, 2018; p 76.

(14) Zheng, Y.; Flanagan, S. V. The case for universal screening of private well water quality in the US and testing requirements to achieve it: evidence from arsenic. *Environ. Health Perspect.* **2017**, *125* (8), 085002.

(15) Ayotte, J. D.; Medalie, L.; Qi, S. L.; Backer, L. C.; Nolan, B. T. Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States. *Environ. Sci. Technol.* **2017**, *51* (21), 12443−12454.

(16) Nolan, B. T.; Fienen, M. N.; Lorenz, D. L. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *J. Hydrol.* **2015**, *531*, 902−911.

(17) Nolan, B. T.; Gronberg, J. M.; Faunt, C. C.; Eberts, S. M.; Belitz, K. Modeling Nitrate at Domestic and Public-Supply Well Depths in the Central Valley, California. *Environ. Sci. Technol.* **2014**, *48* (10), 5643−5651.

(18) Lombard, M. A.; Bryan, M. S.; Jones, D. K.; Bulka, C.; Bradley, P. M.; Backer, L. C.; Focazio, M. J.; Silverman, D. T.; Toccalino, P.; Argos, M. Machine Learning Models of Arsenic in Private Wells Throughout the Conterminous United States As a Tool for Exposure Assessment in Human Health Studies. *Environ. Sci. Technol.* **2021**, *55*, 5012.

(19) Erickson, M. L.; Elliott, S. M.; Brown, C. J.; Stackelberg, P. E.; Ransom, K. M.; Reddy, J. E.; Cravotta, C. A. Machine-Learning Predictions of High Arsenic and High Manganese at Drinking Water Depths of the Glacial Aquifer System, Northern Continental United States. *Environ. Sci. Technol.* **2021**, *55*, 5791.

(20) Roostaei, J.; Colley, S.; Mulhern, R.; May, A. A.; Gibson, J. M. Predicting the risk of GenX contamination in private well water using a machine-learned Bayesian network model. *J. Hazard. Mater.* **2021**, *411*, 125075.

(21) Zhang, X.; Lohmann, R.; Dassuncao, C.; Hu, X. C.; Weber, A. K.; Vecitis, C. D.; Sunderland, E. M. Source attribution of poly- and perfluoroalkyl substances (PFASs) in surface waters from Rhode Island and the New York Metropolitan Area. *Environ. Sci. Technol. Lett.* **2016**, *3* (9), 316−321.

(22) Staff, S., Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States. In United States Department of Agriculture, Natural Resources Conservation Service.

(23) Wolock, D. M. *Estimated mean annual natural ground-water recharge in the conterminous United States*; 2003−311; Reston, VA, 2003.

(24) New Hampshire Department of Environmental Services NH PFAS Investigation. https://www4.des.state.nh.us/nh-pfas-investigation/ (Accessed 05−19−2019),.

(25) New Hampshire Department of Environmental Services Drinking Water. https://www.des.nh.gov/water/drinking-water (04−08),.

(26) Ruiz

(27) Ramyachitra, D.; Manikandan, P., Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)* **2014**, *5*, (4).

(28) Yap, B. W.; Abd Rani, K.; Abd Rahman, H. A.; Fong, S.; Khairudin, Z.; Abdullah, N. N. In *An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets*, Proceedings of the first international conference on advanced data and information engineering (DaEng-2013), 2014; Springer: 2014; pp 13−22.

(29) Shoemaker, J., Method 537. Determination of selected perfluorinated alkyl acids in drinking water by solid phase extraction and liquid chromatography/tandem mass spectrometry (LC/MS/MS). 2009.

(30) U.S. Department of Transportation Federal Aviation Administration Programs for Training of Aircraft Rescue and Firefighting Personnel. https://www.faa.gov/airports/resources/advisory_circulars/index.cfm/go/document.current/documentNumber/150_5210-17 (04−08),.

(31) U.S. Environmental Protection Agency Clean watersheds needs survey. https://www.epa.gov/cwns/clean-watersheds-needs-survey-cwns-2012-report-and-data#access (04−08),.

(32) Ruyle, B. J.; Pickard, H. M.; LeBlanc, D. R.; Tokranov, A. K.; Thackray, C. P.; Hu, X. C.; Vecitis, C. D.; Sunderland, E. M. Isolating the AFFF Signature in Coastal Watersheds Using Oxidizable PFAS Precursors and Unexplained Organofluorine. *Environ. Sci. Technol.* **2021**, *55* (6), 3686−3695.

(33) Carswell, W. J. *National Geospatial Program*. US Department of the Interior, US Geological Survey: 2011.

(34) Gleeson, T.; Marklund, L.; Smith, L.; Manning, A. H., Classifying the water table at regional to continental scales. *Geophys. Res. Lett.* **2011**, *38*, (5). DOI: 10.1029/2010GL046427

(35) Panikkar, B.; Lemmond, B.; Allen, L.; DiPirro, C.; Kasper, S. Making the invisible visible: results of a community-led health survey following PFAS contamination of drinking water in Merrimack, New Hampshire. *Environ. Health* **2019**, *18* (1), 1−16.

(36) Schroeder, T.; Bond, D.; Foley, J. 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* **2021**, *23* (2), 291−301.

(37) Barr Engineering Co. *Air Deposition Modeling Report. Air Permit Application for Installation of a Regenerative Thermal Oxidizer.*; Merrimack, NH, 2019.

(38) Shin, H.-M.; Vieira, V. M.; Ryan, P. B.; Detwiler, R.; Sanders, B.; Steenland, K.; Bartell, S. M. Environmental Fate and Transport Modeling for Perfluorooctanoic Acid Emitted from the Washington Works Facility in West Virginia. *Environ. Sci. Technol.* **2011**, *45* (4), 1435−1442.

(39) Galloway, J. E.; Moreno, A. V. P.; Lindstrom, A. B.; Strynar, M. J.; Newton, S.; May, A. A.; Weavers, L. K. Evidence of Air Dispersion: HFPO−DA and PFOA in Ohio and West Virginia Surface Water and Soil near a Fluoropolymer Production Facility. *Environ. Sci. Technol.* **2020**, *54* (12), 7175−7184.

(40) Weber, A. K.; Barber, L. B.; LeBlanc, D. R.; Sunderland, E. M.; Vecitis, C. D. Geochemical and hydrologic factors controlling subsurface transport of poly-and perfluoroalkyl substances, Cape Cod, Massachusetts. *Environ. Sci. Technol.* **2017**, *51* (8), 4269−4279.

(41) Umeh, A. C.; Naidu, R.; Shilpi, S.; Boateng, E. B.; Rahman, A.; Cousins, I. T.; Chadalavada, S.; Lamb, D.; Bowman, M. Sorption of PFOS in 114 Well-Characterized Tropical and Temperate Soils: Application of Multivariate and Artificial Neural Network Analyses. *Environ. Sci. Technol.* **2021**, *55* (3), 1779−1789.

(42) Sharifan, H.; Bagheri, M.; Wang, D.; Burken, J. G.; Higgins, C. P.; Liang, Y.; Liu, J.; Schaefer, C. E.; Blotevogel, J. Fate and transport of per-and polyfluoroalkyl substances (PFASs) in the vadose zone. *Sci. Total Environ.* **2021**, *771*, 145427.

(43) Amini, M.; Abbaspour, K. C.; Berg, M.; Winkel, L.; Hug, S. J.; Hoehn, E.; Yang, H.; Johnson, C. A. Statistical modeling of global geogenic arsenic contamination in groundwater. *Environ. Sci. Technol.* **2008**, *42* (10), 3669−3675.

(44) Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **2006**, *27* (8), 861−874.

(45) Zeng, X.; Martinez, T. R. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **2000**, *12* (1), 1−12.

(46) Purushotham, S.; Tripathy, B. In *Evaluation of classifier models using stratified tenfold cross validation techniques*, International Conference on Computing and Communication Systems, 2011; Springer: *2011*; pp 680−690.

(47) Ripley, B.; Venables, B.; Bates, D. M.; Hornik, K.; Gebhardt, A.; Firth, D.; Ripley, M. B., Package 'mass'. *CRAN R* **2013**, *538*.

(48) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*. Springer: New York City, 2013.

(49) Liaw, A.; Wiener, M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression* **2018**, *4*, 6−14.

(50) U.S. Environmental Protection Agency Saint Gobain Performance Plastics site, village of Hoosick Falls, New York. https://cumulis.epa.gov/supercpad/SiteProfiles/index.cfm?fuseaction=second.cleanup&id=0202702 (04−08),.

(51) New Hampshire Department of Environmental Services Elevated Levels of PFAS in Stormwater at the Saint-Gobain Facility https://www4.des.state.nh.us/nh-pfas-investigation/?p=769 (04−08),.

(52) Guelfo, J. L.; Higgins, C. P. Subsurface Transport Potential of Perfluoroalkyl Acids at Aqueous Film-Forming Foam (AFFF)-Impacted Sites. *Environ. Sci. Technol.* **2013**, *47* (9), 4164−4171.

(53) Chovancova, A. Profiles of Poly-and Perfluoroalkyl Substances (PFASs) in Surface Waters and Biota From Merrimack River Subwatersheds, New Hampshire. 2018.

(54) Daly, E. R.; Chan, B. P.; Talbot, E. A.; Nassif, J.; Bean, C.; Cavallo, S. J.; Metcalf, E.; Simone, K.; Woolf, A. D. Per-and polyfluoroalkyl substance (PFAS) exposure assessment in a community exposed to contaminated drinking water, New Hampshire, 2015. *Int. J. Hyg. Environ. Health* **2018**, *221* (3), 569−577.

(55) Ayotte, J.; Medalie, L.; Qi, S.; Backer, L. F.; Nolan, B. T. In *Estimating the probability of arsenic occurrence in domestic wells in the United States*, AGU Fall Meeting Abstracts, 2016; *2016*; pp PA21B−2204.

(56) Nolan, B. T.; Hitt, K. J. Vulnerability of Shallow Groundwater and Drinking-Water Wells to Nitrate in the United States. *Environ. Sci. Technol.* **2006**, *40* (24), 7834−7840.

(57) Ouedraogo, I.; Defourny, P.; Vanclooster, M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeol. J.* **2019**, *27* (3), 1081−1098.

(58) Anning, D. W.; Paul, A. P.; McKinney, T. S.; Huntington, J. M.; Bexfield, L. M.; Thiros, S. A. *Predicted nitrate and arsenic concentrations in basin-fill aquifers of the southwestern United States*. US Department of the Interior, US Geological Survey: 2012.