Supporting Information

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

Xindi C. Hu[a,b,c,†], Beverly Ge[a], Bridger J. Ruyle[a], Jennifer Sun[a], Elsie M. Sunderland[a, c]

Author Affiliations

[a]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

[b]Mathematica, Inc., Oakland, CA, USA

[c]Department of Environmental Health, Harvard T.H Chan School of Public Health, Boston, MA, USA

†*Correspondence to*: Xindi C. Hu, Mathematica, Inc., 505 14th street, 8th floor, Oakland, CA 94612 USA. Telephone: 5102854675. Email: chu@mathematica-mpr.com

Table of Contents

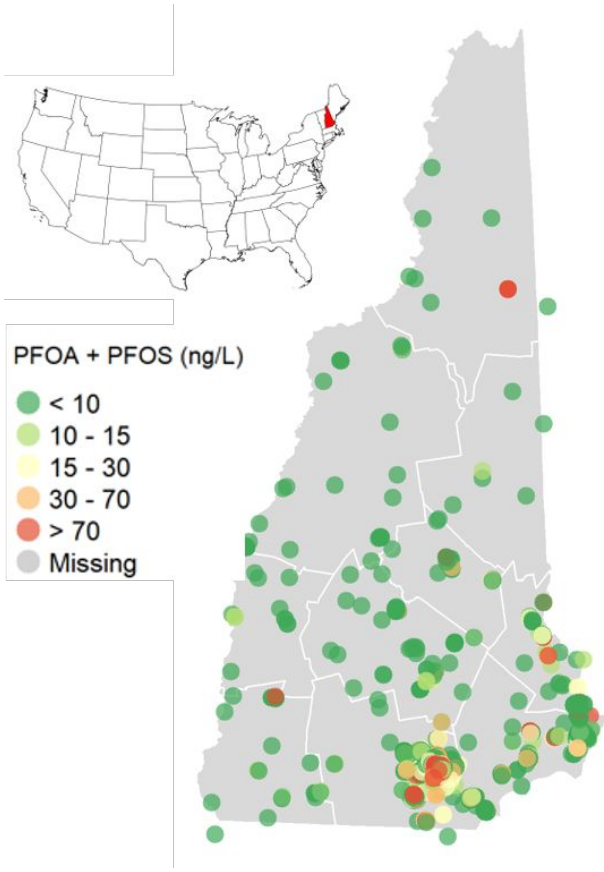

Figure S1. Observed PFOA and PFOS concentration in New Hampshire private wells (n = 2366). Data are from NHDES domestic well sampling campaign, 2014 - 2017.[1]

S1. Data processing steps for PFAS concentration data from New Hampshire (NH) domestic wells

One common limitation of secondary data is that samples measured at different times may have different detection limits, creating a multiple censoring problem (Table S1). We chose the median detection limit (DL) as a uniform DL and treated samples with DLs at or below this value as non-detects. The uniform DLs for PFPeA, PFHxA, PFHpA, PFOA, and PFOS were 5.0, 8.0, 5.0, 8.0 and 5.0 ng/L, respectively. The DL of most is very close to the uniform DL so this does not significantly skew the sample distribution. We removed 254 (1.6%) samples with DLs that are more than five times of the uniform DL due to data quality concerns. 790 wells were sampled multiple times (2 to 48 times), often due to those wells having concentrations approaching but not exceeding the standard. For these wells, we use the average PFAS concentration detected. The mean coefficient of variation across multiple samples for the same well was below 25% for all PFAS, suggesting relatively low temporal variability.

Table S1. Detection limit (DL) for PFAS measured in domestic well waters in NH

Compound Name	Acronym	DL (ng/L) [†]				
		min	median	75th percentile	98th percentile	max
Perfluoropentanoic acid	PFPeA	0.015	4.5	4.7	5.0	10
Perfluorohexanoic acid	PFHxA	0.015	4.5	4.6	8.0	10
Perfluoroheptanoic acid	PFHpA	0.015	3.4	4.5	5.0	16
Perfluorooctanoic acid	PFOA	0.015	2.0	2.0	8.0	8.0
Perfluorooctanesulfonic acid	PFOS	0.015	4.0	4.5	5.0	10

[†] The distribution of the limit of detection for PFAS was calculated across all 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.

Table S2. NAICS codes for identifying PFAS sources in EPA Facility Registry Service

NAICS code	Description of Industry	Number of unique sites in NH
313	Textile mills	28
322	Paper manufacturing	26
323	Printing and related support activities	100
324	Petroleum and coal products manufacturing	60
3255	Chemical manufacturing	24
32591	Printing ink manufacturing	7
3328	Metal coating, engraving, heat treating and allied activities	40
3344	Semiconductor and other electronic component manufacturing	124

Notes: 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 We identified the locations of these potential PFAS sources using for any time before October 2017, the year when the latest samples were collected.

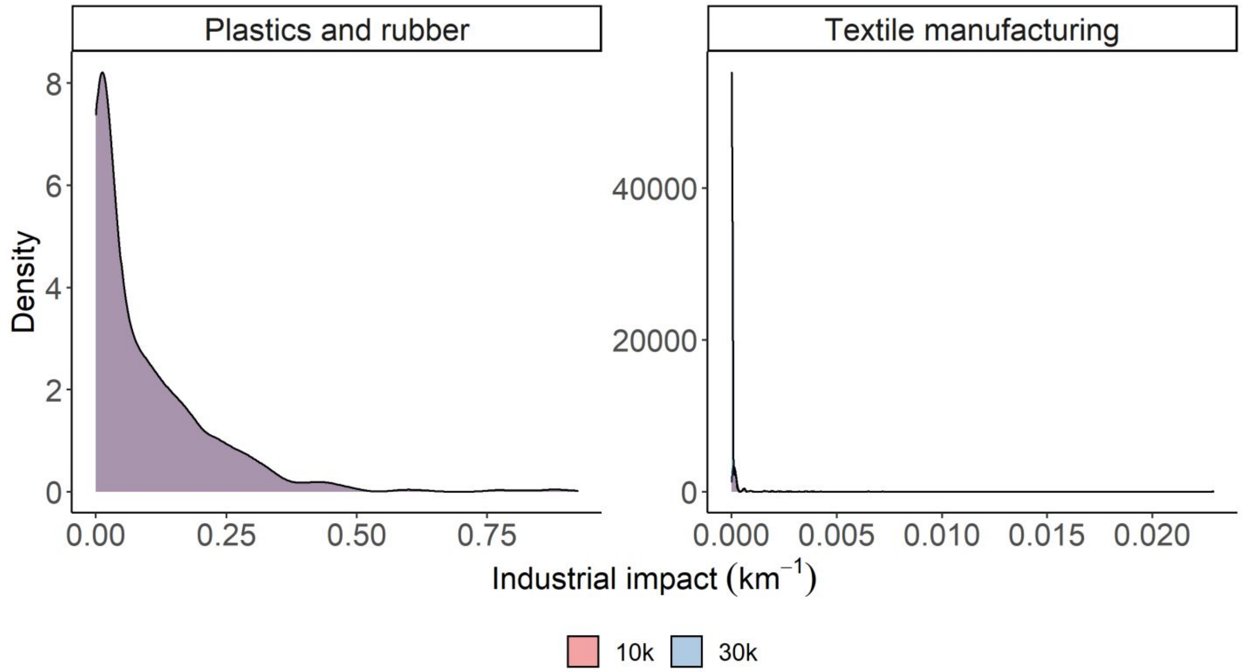

Figure S2. Sensitivity analysis of different atmospheric buffer distances on industrial impact scores for the plastics and rubber industry and textile manufacturing.

S2. Detailed description of environmental predictors

Environmental predictors considered in this work can be classified into four main categories: (1) geologic variables such as bedrock type, (2) variables reflecting soil geochemistry such as bulk density and sand/silt/clay content, (3) hydrologic variables such as precipitation and groundwater recharge, and (4) other features of the hydrologic landscape such as elevation, slope and land use. Well depth is often an important predictor for chemical concentrations in domestic wells,[2] but was not consistently collected in the NHDES sample campaign. We therefore used the annual minimum depth to water table in gSSURGO as a proxy.[3] No statewide data were available for the groundwater table. Thus, we used elevation as a proxy. Datasets for NH were accessed through the US Geologic Survey and NHDES websites in the format of raster files or spatial shapefiles (Table S2). To assign independent variables to each well, we overlaid the raster or shapefile that contained data for NH with the well location in a Geographic Information System (GIS). Some variables were available at high spatial resolution. For example, data on soil geochemistry were available at the 10 m x10 m scale. Other variables such as groundwater recharge were available at the 1 km x1 km scale, and precipitation was available at 4 km x 4 km scale. In the infrequent event (less than 20 wells out of the 2383 wells) where the well location was not covered by the variable layer, missing values were imputed by the arithmetic mean of the variable across all wells with available information. We chose this imputation method because it preserves the mean distribution of variables.[4]

Table S3. List of independent model variables (point sources and environmental factors)

Independent Variable	min	Q1	median	Q3	P98	Max	Units or Scale	Data Source
Point Source Impacts								
Impact: Plastics and Rubber Products Manufacturing	0	0	1.11×10^{-3}	6.22×10^{-2}	3.42×10^{-1}	9.23×10^{-1}	km^{-1} [1]	15
Impact: Textiles Manufacturing and Related Activities	0	0	0	0	6.33×10^{-1}	9.15×10^{-1}	km^{-1}	15
Impact: Airports	0	0	0	0	0	6.36×10^{-1}	km^{-1}	15
Impact: military sites	0	0	0	0	4.50×10^{-2}	9.63×10^{-1}	km^{-1}	15
Impact: wastewater treatment plants	0	0	0	0	1.25×10^{-1}	9.51×10^{-1}	km^{-1}	15
Impact: Potential sources	0	0	8.27×10^{-3}	7.57×10^{-1}	4.19	7.80	km^{-1}	15
Geologic Variables								
Bedrock Type: Metasedimentary and Metavolcanic Rocks of the Merrimack Trough	77.8% have value 1, 22.2% have value 0						Unitless[2]	16
Depth to bedrock	6.00×10^{-5}	30.8	30.8	30.8	41.0	76.0	meter [3]	17
Hydrologic Variables								
Total monthly precipitation in the year that each well sample was taken	29.0	53.7	70.1	103	167	205	mm [4]	18
Mean annual natural ground-water recharge, derived by multiplying a grid of base-flow index values by a grid of mean annual runoff values (from 1951-1980)	216	266	273	275	314	410	mm/year [5]	19
Depth to water table - annual minimum	1.00×10^{-5}	52.7	52.7	52.7	77.0	153	cm	17
Slope gradient – difference in elevation between two points as a percentage of the distance between those points	1.00×10^{-5}	2.90	3.30	9.90	25.0	44.0	%	17

Hydrologic Group Dominant Component: A – Low runoff potential	57.5% have value 1, 42.5% have value 0						unitless	17
Soil Geochemistry								
Silt: Total - Mineral particles 0.002 mm - 0.05 mm in equivalent diameter as a weight percentage of <2 mm fraction	2.00×10^{-1}	12.9	15.4	23.5	46.2	68.6	%	17
Clay: Total - Mineral particles less than .002 mm in equivalent diameter as a weight percentage of <2 mm fraction	3.00×10^{-2}	9.50×10^{-1}	1.85	4.00	8.71	31.6	%	17
Sand: Total - Mineral particles greater than 0.05 mm in equivalent diameter as a weight percentage of <2 mm fraction	6.20×10^{-1}	64.0	79.5	81.2	85.9	95.0	%	17
Bulk Density at a water tension of 1/3 bar	5.50×10^{-2}	1.14	1.31	1.47	1.63	1.79	g/mL	17
Available Water Capacity[6]	4.02×10^{-3}	8.80×10^{-2}	1.19×10^{-1}	1.32×10^{-1}	3.17×10^{-1}	5.30×10^{-1}	vol. water/vol. soil	17
Cation Exchange Capacity at pH 7.0, as estimated by the ammonium acetate method	5.01×10^{-3}	9.00×10^{-2}	7.65×10^{-1}	1.44	6.48	59.3	meq/g soil	17
Soil organic carbon stock estimate in total soil profile (0 cm to reported depth of soil profile)	312	1.04×10^{4}	1.23×10^{4}	1.31×10^{4}	2.50×10^{4}	1.59×10^{5}	g Carbon	17
Soil thickness	7.20×10^{-1}	14	20.4	45.6	105	149	cm	17

Notes:
[1] Impact is calculated as an exponential decay function of the Haversine distance between the point source and well. Only industries with elevation above a well and within the same 12-digit HUC were considered.
[2] 1:250000 scale.
[3] 10-meter resolution grid dataset.
[4] 4-kilometer resolution grid dataset.
[5] 1-kilometer resolution grid dataset
[6] The quantity of water that the soil is capable of storing

		Predicted		
		Detect	Non-detect	
Observed	Detect	True positive (TP)	False negative (FN)	Sensitivity $\frac{\text{TP}}{\text{TP} + \text{FN}}$
	Non-detect	False positive (FP)	True negative (TN)	Specificity $\frac{\text{TN}}{\text{TN} + \text{FP}}$
				Accuracy $\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$

Figure S3. Confusion matrix for categorical models (logistic regression and classification random forest).

S3. Regression random forest model

We tested the performance of both continuous (regression random forest) and categorical (logistic regression and classification random forest) models. Continuous models predict the magnitude of PFAS concentrations likely to be found in a well, while categorical models predict the likelihood that concentrations fall below or above a threshold level.

For the continuous model, we only considered wells with detectable PFAS due to the large number of measurements below detection. A natural log transformation was used to reduce impacts of extreme outliers on the model fitting process. Mean squared error (MSE) and a pseudo R^2 were used to assess the model performance. We evaluated the relative importance of predictors by random permutation and calculated the percent increase in MSE. Statistical analyses were conducted using the *randomForest* package in R 4.0.0.[5]

Performance of the continuous model (regression random forest) was moderate to poor across the five PFAS with pseudo-R^2 values ranging from 0.024 for PFOS to 0.52 for PFPeA (Table S6). This performance is similar to modeling studies for other toxicants in groundwater with a similar sample size (see SI Section S4 for more details). The lowest performance likely reflects the limited detectable concentration data available for PFOS (n = 465). The sample size for the regression random forest is much smaller than the categorical models due to the exclusion of samples below detection. The intended purpose of this type of statistical model is as a screening tool to prioritize field sampling. Thus, we conclude based on these results that classification models that can use all available data are preferable. Classification random forest models may be preferred over continuous models because they can use all data collected in monitoring programs, avoiding poor performance for PFAS like PFOS in this study that had a low overall frequency of detection.

Table S4. Model performance for regression random forest

	PFPeA	PFHxA	PFHpA	PFOA	PFOS
n*	499	749	750	1658	465
Mean Squared Error	0.56	0.60	0.53	1.2	1.2
pseudo-R^2	0.52	0.41	0.40	0.40	0.024

Note: *Regression random forest model has a smaller sample size than the other two methods because it was developed only on samples with detectable PFAS concentrations.

Table S5. Standardized[a] logistic regression model coefficients (± standard error).

	PFPeA	PFHxA	PFHpA	PFOA	PFOS	PFAS
			Industry¶			
Plastics and rubber	1.2 ± 0.13***	0.57 ± 0.12***	0.88 ± 0.11***	0.18 ± 0.09*	0.20 ± 0.11	0.33 ± 0.10***
Textile manufacturing	-0.60 ± 0.18**	-0.52 ± 0.13***		0.42 ± 0.10**	-0.29 ± 0.13*	0.28 ± 0.10**
Military sites				-0.22 ± 0.13		
WWTP[b]	0.53 ± 0.13***	0.44 ± 0.11***	0.28 ± 0.09**	0.19 ± 0.09*		0.40 ± 0.11***
Potential sources[€]		0.31 ± 0.11**	0.24 ± 0.10*		0.43± 0.10***	
			Geo			
Bedrock type[‡]	0.64 ± 0.16***	0.69 ± 0.17***	1.1 ± 0.15***	1.3 ± 0.12***		0.92 ± 0.11***
Depth to bedrock	0.19 ± 0.12***		-0.32 ± 0.12**	-0.37 ± 0.10***		-0.38 ± 0.11***
			Hydro			
Monthly precipitation		-0.29 ± 0.12*			0.32 ± 0.11**	
Low runoff potential	0.53 ± 0.16**	0.57 ± 0.18**	0.33 ± 0.20		0.45 ± 0.19*	0.34 ± 0.10**
Water table depth			0.26 ± 0.11*		-0.72 ± 0.12***	
Groundwater recharge				0.26 ± 0.09**		
Slope gradient	0.21 ± 0.13	0.19 ± 0.12		-0.19 ± 0.10	-0.20 ± 0.13*	-0.17 ± 0.10
			Soil			
Percent clay				-0.56 ± 0.12***	0.19 ± 0.11	-0.39 ± 0.12***
Percent silt		-0.43 ± 0.16**				
Percent sand	-0.25 ± 0.15	-0.69 ± 0.30		0.34 ± 0.10***		
Bulk density		0.58 ± 0.28*				
Available water capacity		0.37 ± 0.13**	0.34 ± 0.12**		-0.32 ± 0.12**	
Organic carbon content	0.40 ± 0.13**					
Soil thickness	-0.19 ± 0.13			-0.15 ± 0.09		-0.13 ± 0.09
Saturated hydraulic conductivity			0.57 ± 0.20***		-0.82 ± 0.20***	
Cation exchange capacity						0.20 ± 0.12
C-Statistics[‡]	0.70	0.69	0.74	0.68	0.65	0.66
AUROC (95% CI)[£]	0.68 (0.65, 0.72)	0.67 (0.65, 0.70)	0.72 (0.70, 0.75)	0.66 (0.64, 0.68)	0.63 (0.61, 0.64)	0.64 (0.62, 0.66)
n	1617	1725	2253	2373	2376	2379

[a] Standardized coefficients are unitless, normalized values so that the variances of dependent and independent variables are equal to 1 and can be compared because they reflect how many standard deviations PFAS concentrations will change per standard deviation in the predictor variable.
[b] WWTP = wastewater treatment plants.
Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Variables not selected by the logistic regression models are pH calculated by the 1:1 soil-water ratio, and available water storage from the surface to reported depth of soil profile.
¶ Industry impact is calculated as an exponential decay function of the Haversine distance between the point source and well. Only industries with elevation above a well and within the same 12-digit HUC were considered.
[‡]Bedrock type = Metasedimentary and Metavolcanic Rocks of the Merrimack Trough

[†] Coefficients that are not statistically significantly different from zero at $p = 0.05$ level are kept in the table because they were selected in the stepwise logistic regression.

[€] Potential sources: sources considered include semiconductor, printing, metal plating, textile mills, petroleum and coal products manufacturing, chemical manufacturing.

[‡]Concordance (C) statistics are used to assess model discrimination, which means how well the model can separate the wells with detectable concentrations from those with non-detect. C statistics ranges from 0.5 to 1, and values around 0.7 generally indicate a good model.

[£]Area under the Receiver Operating Characteristics curve (AUROC) is used to evaluate the classification model's performance. The mean AUROC and its 95% confidence interval (CI) is calculated by 10-fold cross validation.

Table S6. Tuning of hyperparameters in the random forest model and area under the Receiver Operating Characteristics curve (AUROC)

	Classification Random Forest[1]					
	Worst performance			Best performance		
	mtry[2]	*ns*[3]	AUROC	*mtry*	*ns*	AUROC
PFPeA	20	1	0.72	10	2	0.79
PFHxA	17	1	0.71	16	9	0.78
PFHpA	19	1	0.76	6	4	0.86
PFOA	20	8	0.78	10	4	0.84
PFOS	8	1	0.82	22	5	0.74
sumPFAS	20	1	0.74	17	6	0.81

Notes:
[1]The number of trees used was 1000 given that the out-of-bag error converged by then across all compounds.
[2]Number of features randomly sampled at each node out of an original 26 features.
[3]Minimal size of terminal nodes.

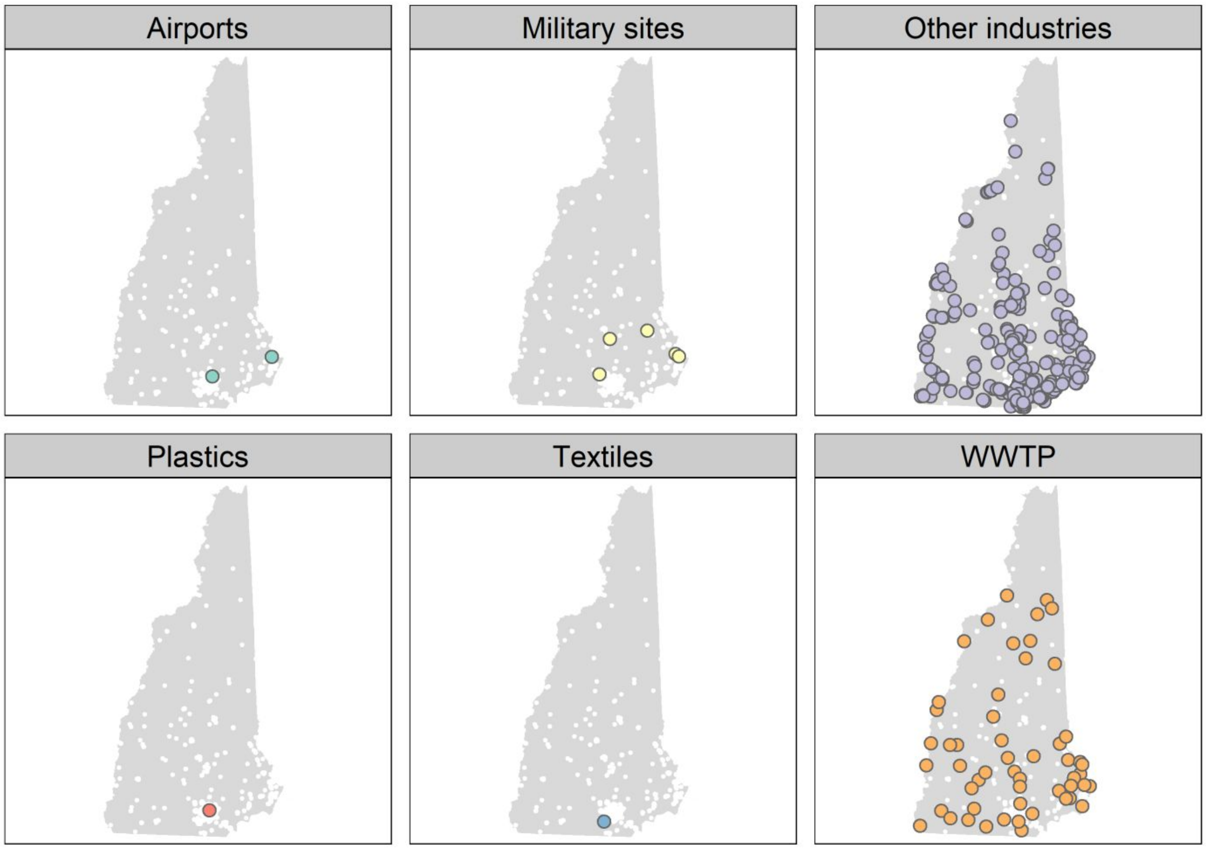

Figure S4. Locations of potential PFAS sources in NH. White dots represent the location of wells sampled in this study. The number of wells sampled that were influenced by each source are as follows: Airports (*n* = 36); Military Bases (*n* = 51); Other Industries (potential sources) (*n* = 1471); Plastics (*n* = 1320); Textiles (*n* = 573); Wastewater treatment plants (WWTP) (*n* = 203).

S4. Review of previous literature on machine learning models for drinking water contaminants.

We reviewed 18 peer-reviewed studies that used a similar methodology published between 2012 and 2019 (SI Table S5). Most studies have focused on geogenic and inorganic groundwater pollution such as arsenic, fluoride and anthropogenic pollution such as nitrate. Our study is the first to apply this methodology to predict PFAS in private wells. Model performance varied across location, compound, sample size, and the machine learning models used.

The performance of classification random forest models developed in this study is similar to previous efforts to model other toxicants in groundwater by developing machine learning models. Random forest models were used in 58% of the studies reviewed and achieved on average an accuracy rate of 79% (range: 37% - 92%), which is similar to the results shown here. For screening purpose, false negatives (missing wells with potentially high contamination) are more consequential than false positives. In this current work and most previous work, the classification threshold is set to maximize accuracy. In practice, this can be adjusted so that some true negative rate is sacrificed in order to reduce false negatives.

Performance of the regression random forest models in this study was comparable to those for predicting groundwater nitrate contamination from prior work.[2, 6, 7] Similar to the classification random forest model, groundwater recharge and monthly precipitation were consistently among the most important predictors for all five PFAS modeled, in addition to impacts from industrial sources such as plastics manufacturing, printing and textile manufacturing. A model for groundwater nitrate concentrations in Germany with a comparable sample size to ours (1890 wells) had an R^2 of 0.54.[6] In Iowa and North Carolina, continuous models similarly had low predictive performance (R^2 <0.33) for predicting groundwater nitrate concentrations in 22,000 private wells sampled.[7] Common challenges for such modeling include dealing with a low fraction of samples with detectable concentrations, and limited data on some important spatial predictors, particularly those relating to local groundwater flow conditions that are not always available at statewide or larger spatial scales.[8]

Table S7. Previously published predictive models for toxicants in private wells

No.	Author	Year	Location	Compound	Sample size	ML technique	Performance	Ref
1	Anning	2012	Arizona, California, Colorado, Nevada, New Mexico, and Utah	Nitrate	Not reported (NR)	Random forest	Correct 48.6%, Overpredicted 25.8%, Underpredicted, 25.6%	9
2	Anning	2012	Arizona, California, Colorado, Nevada, New Mexico, and Utah	Arsenic	NR	Random forest	Correct 36.7%, Overpredicted 33.5%, Underpredicted 29.8%	9
3	Nolan	2014	Central Valley, CA	Nitrate (shallow well)	314	Logistic regression	Predict nitrate>4mg/L; accuracy 69.7%, sensitivity 69.0%, specificity 70.4%	2
4	Nolan	2014	Central Valley, CA	Nitrate (shallow well)	318	Random forest classification	Predict nitrate>4mg/L; accuracy 71.7%, sensitivity 65.1%, specificity 77.3%	2
5	Nolan	2014	Central Valley, CA	Nitrate (shallow well)	318	Random forest regression	Predict nitrate>4mg/L; accuracy 68.9%, sensitivity 84.2%, specificity 55.8%	2
6	Nolan	2014	Central Valley, CA	Nitrate (deep well)	928	Logistic regression	Predict nitrate>4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9%	2
7	Nolan	2014	Central Valley, CA	Nitrate (deep well)	937	Random forest classification	Predict nitrate>4mg/L; accuracy 81.2%, sensitivity 25.1%, specificity 96.3%	2
8	Nolan	2014	Central Valley, CA	Nitrate (deep well)	937	Random forest regression	Predict nitrate>4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 89.7%	2
9	Rodriguez-Galiano	2014	Granada city, Spain	Nitrate	175	Random forest	Predict nitrate >50mg/L; 80.46%	10
10	Rodriguez-Galiano	2014	Granada city, Spain	Nitrate	175	Logistic regression	Predict nitrate >50mg/L; accuracy is 73.56%	10
11	Singh	2014	Indo-Gangetic	Chemical oxygen	409	Decision tree boost	Test dataset, R^2 = 0.918	11

			plains of north India	demand (COD)				
12	Nolan	2015	Central Valley, CA	Nitrate (shallow well)	318	Boosted regression trees	Hold-out data, R^2=0.26	12
13	Nolan	2015	Central Valley, CA	Nitrate (shallow well)	318	Artificial neural networks	Hold-out data, R^2=0.12	12
14	Nolan	2015	Central Valley, CA	Nitrate (shallow well)	318	Bayesian networks	Hold-out data, R^2=0.18	12
15	Wheeler	2015	Iowa	Nitrate	34,084	Random forest	Test dataset, R^2 = 0.38	13
16	Wheeler	2015	Iowa	Nitrate	34,084	Random forest	Predict nitrate >5mg/L; Accuracy is 0.92, sensitivity is 0.75, specificity is 0.96	13
17	Ayotte	2016	Central Valley, CA	Arsenic	1,180	Boosted regression trees	Predict arsenic >10 µg/L; Accuracy is 0.91, sensitivity is 0.39, specificity is 0.96	8
18	Ayotte	2016	Central Valley, CA	Arsenic	1,180	Logistic regression	Predict arsenic >10 µg/L; Accuracy is 0.90, sensitivity is 0.18, specificity is 0.98	8
19	Ayotte	2017	Contiguous US	Arsenic	20,450	Logistic regression	Predict arsenic >10 µg/L; Accuracy is 0.90, sensitivity is 0.14, specificity is 0.99	14
20	Ransom	2017	Central Valley, CA	Nitrate	5,170	Boosted regression trees	Hold-out data, R^2=0.434	15
21	Rosecrans	2017	Central Valley, CA	Dissolved oxygen	2,767	Boosted regression trees	Hold-out data, predict DO<0.5 mg/L, AUC is 0.87	16
22	Rosecrans	2017	Central Valley, CA	Manganese	2,767	Boosted regression trees	Hold-out data, predict Mn>50 µg/L, AUC is 0.87	16
23	Tesoriero	2017	Northeastern Wisconsin	Nitrate	10,866	Random forest classification	Predict nitrate >5mg/L, test data, accuracy 75%, AJC 0.80	17
24	Tesoriero	2017	Northeastern Wisconsin	Iron	539	Random forest classification	Predict iron >0.1 mg/L, on out of bag training data, accuracy is 74%, AUC is 0.79	17
25	Tesoriero	2017	Northeastern Wisconsin	Arsenic	1,275	Random forest classification	Predict arsenic >5 µg/L, on out of bag training data,	17

							accuracy is 74%, AUC is 0.79	
26	Erickson	2018	North-central USA	Arsenic	3,283	Boosted regression trees	Predict arsenic >10 µg/L. On hold-out dataset, accuracy is 67%, ROC is 0.72	[18]
27	Podgorski	2018	India	Fluoride	12,600	Random forest	Predict fluoride >1.5mg/L, accuracy is 0.78, AUC is 0.84	[19]
28	Rodriguez-Galiano	2018	Granada city, Spain	Nitrate	110	Random forest	Predict nitrate >50mg/L; AUC is 0.92	[20]
29	Trajanov	2018	France	Pesticides	NR	Random forest	Recalls of 0.84 and 0.86 for the risky and not-risky class respectively	[21]
30	Canion	2019	Florida	Nitrate	1554	Random forest classification	Predict nitrate >0.35 mg/L; AUC is 0.89, accuracy is 0.83, sensitivity is 0.79, specificity is 0.86	[22]
31	Canion	2019	Florida	Nitrate	1554	Random forest classification	Predict nitrate >1.2 mg/L; AUC is 0.84, accuracy is 0.79, sensitivity is 0.54, specificity is 0.89	[22]
32	Knoll	2019	Germany	Nitrate	1890	Random forest regression	Predict nitrate concentration, R^2=0.54	[6]
33	Messier	2019	North Carolina	Nitrate	22000	Multiple random forest classification	Predict nitrate < 1 mg/L, 1 – 5 mg/L, and ≥5 mg/L, overall accuracy is 0.79	[7]

References

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

2. 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. *Environmental science & technology* **2014,** *48*, (10), 5643-5651.

3. U.S. Department of Agriculture, The Gridded Soil Survey Geographic (gSSURGO) Database for West Virginia. United States Department of Agriculture, Natural Resources Conservation Service. . In November 16, 2015 ed.; Available online at https://gdg.sc.egov.usda.gov/, 2016.

4. Musil, C. M.; Warner, C. B.; Yobas, P. K.; Jones, S. L., A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research* **2002,** *24*, (7), 815-829.

5. Liaw, A.; Wiener, M. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression.*, 4.6-14; 2018.

6. Knoll, L.; Breuer, L.; Bach, M., Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Science of The Total Environment* **2019,** *668*, 1317-1327.

7. Messier, K. P.; Wheeler, D. C.; Flory, A. R.; Jones, R. R.; Patel, D.; Nolan, B. T.; Ward, M. H., Modeling groundwater nitrate exposure in private wells of North Carolina for the Agricultural Health Study. *Science of The Total Environment* **2019,** *655*, 512-519.

8. Ayotte, J. D.; Nolan, B. T.; Gronberg, J. A., Predicting arsenic in drinking water wells of the Central Valley, California. *Environmental science & technology* **2016,** *50*, (14), 7555-7563.

9. 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.

10. Rodriguez-Galiano, V.; Mendes, M. P.; Garcia-Soldado, M. J.; Chica-Olmo, M.; Ribeiro, L., Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment* **2014,** *476*, 189-206.

11. Singh, K. P.; Gupta, S.; Mohan, D., Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. *Journal of Hydrology* **2014,** *511*, 254-266.

12. Nolan, B. T.; Fienen, M. N.; Lorenz, D. L., A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *Journal of Hydrology* **2015,** *531*, 902-911.

13. Wheeler, D. C.; Nolan, B. T.; Flory, A. R.; DellaValle, C. T.; Ward, M. H., Modeling groundwater nitrate concentrations in private wells in Iowa. *Science of the Total Environment* **2015,** *536*, 481-488.

14. 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.

15. Ransom, K. M.; Nolan, B. T.; Traum, J. A.; Faunt, C. C.; Bell, A. M.; Gronberg, J. A. M.; Wheeler, D. C.; Rosecrans, C. Z.; Jurgens, B.; Schwarz, G. E., A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Science of The Total Environment* **2017,** *601*, 1160-1172.

16. Rosecrans, C. Z.; Nolan, B. T.; Gronberg, J. M., Prediction and visualization of redox conditions in the groundwater of Central Valley, California. *Journal of Hydrology* **2017,** *546*, 341-356.

17. Tesoriero, A. J.; Gronberg, J. A.; Juckem, P. F.; Miller, M. P.; Austin, B. P., Predicting redox‐sensitive contaminant concentrations in groundwater using random forest classification. *Water Resources Research* **2017,** *53*, (8), 7316-7331.

18. Erickson, M. L.; Elliott, S. M.; Christenson, C.; Krall, A. L., Predicting geogenic Arsenic in Drinking Water Wells in Glacial Aquifers, North‐Central USA: Accounting for Depth‐Dependent Features. *Water Resources Research* **2018,** *54*, (12), 10,172-10,187.

19. Podgorski, J. E.; Labhasetwar, P.; Saha, D.; Berg, M., Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environmental science & technology* **2018,** *52*, (17), 9889-9898.

20. Rodriguez-Galiano, V. F.; Luque-Espinar, J. A.; Chica-Olmo, M.; Mendes, M. P., Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Sci Total Environ* **2018,** *624*, 661-672.

21. Trajanov, A.; Kuzmanovski, V.; Real, B.; Perreau, J. M.; Džeroski, S.; Debeljak, M., Modeling the risk of water pollution by pesticides from imbalanced data. *Environmental Science and Pollution Research* **2018,** *25*, 18781-18792.

22. Canion, A.; McCloud, L.; Dobberfuhl, D., Predictive modeling of elevated groundwater nitrate in a karstic spring-contributing area using random forests and regression-kriging. *Environmental Earth Sciences* **2019,** *78*, (9), 271.