1	Supporting	Inform	ation
-	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~		

# **3** A Statistical Approach for Identifying Private Wells Susceptible to Perfluoroalkyl

## 4 Substances (PFAS) Contamination

5 Xindi C. Hu<sup>a,b,c,†</sup>, Beverly Ge<sup>a</sup>, Bridger J. Ruyle<sup>a</sup>, Jennifer Sun<sup>a</sup>, Elsie M. Sunderland<sup>a, c</sup>

# 6 Author Affiliations

7 <sup>a</sup>Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University,

## 8 Cambridge, MA, USA

- 9 <sup>b</sup>Mathematica, Inc., Oakland, CA, USA
- 10 <sup>c</sup>Department of Environmental Health, Harvard T.H Chan School of Public Health, Boston, MA,
- 11 USA
- 12
- 13 *Correspondence to*: Xindi C. Hu, Mathematica, Inc., 505 14th street, 8th floor, Oakland, CA
- 14 94612 USA. Telephone: 5102854675. Email: chu@mathematica-mpr.com

16	Table	of	Contents

17	Figure S1. Observed PFOA and PFOS concentration in New Hampshire private wells (n =
18	2366). Data are from NHDES domestic well sampling campaign, 2014 - 2017. <sup>1</sup>
19	S1. Data processing steps for PFAS concentration data from New Hampshire (NH) domestic
20	wells5
21	Table S1. Detection limit (DL) for PFAS measured in domestic well waters in NH5
22	Table S2. NAICS codes for identifying PFAS sources in EPA Facility Registry Service
23	Figure S2. Sensitivity analysis of different atmospheric buffer distances on industrial impact
24	scores for the plastics and rubber industry and textile manufacturing
25	S2. Detailed description of environmental predictors
26	Table S3. List of independent model variables (point sources and environmental factors)
27	Figure S3. Confusion matrix for categorical models (logistic regression and classification
28	random forest)11
29	S3. Regression random forest model
30	Table S4. Model performance for regression random forest    13
31	Table S5. Standardized <sup>a</sup> logistic regression model coefficients (± standard error)14
32	Table S6. Tuning of hyperparameters in the random forest model and area under the Receiver
33	Operating Characteristics curve (AUROC)16
34	Figure S4. Locations of potential PFAS sources in NH. White dots represent the location of wells
35	sampled in this study. The number of wells sampled that were influenced by each source are as
36	follows: Airports ( $n = 36$ ); Military Bases ( $n = 51$ ); Other Industries (potential sources) ( $n =$

37	1471); Plastics ( $n = 1320$ ); Textiles ( $n = 573$ ); Wastewater treatment plants (WWTP) ( $n = 203$ ).
38	
39	S4. Review of previous literature on machine learning models for drinking water contaminants.
40	
41	Table S7. Previously published predictive models for toxicants in private wells
42	References
43	



46 Figure S1. Observed PFOA and PFOS concentration in New Hampshire private wells (n =

2366). Data are from NHDES domestic well sampling campaign, 2014 - 2017.<sup>1</sup> 

# 49 S1. Data processing steps for PFAS concentration data from New Hampshire (NH) 50 domestic wells

51	One common limitation of secondary data is that samples measured at different times
52	may have different detection limits, creating a multiple censoring problem (Table S1). We chose
53	the median detection limit (DL) as a uniform DL and treated samples with DLs at or below this
54	value as non-detects. The uniform DLs for PFPeA, PFHxA, PFHpA, PFOA, and PFOS were 5.0,
55	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
56	does not significantly skew the sample distribution. We removed 254 (1.6%) samples with DLs
57	that are more than five times of the uniform DL due to data quality concerns. 790 wells were
58	sampled multiple times (2 to 48 times), often due to those wells having concentrations
59	approaching but not exceeding the standard. For these wells, we use the average PFAS
60	concentration detected. The mean coefficient of variation across multiple samples for the same
61	well was below 25% for all PFAS, suggesting relatively low temporal variability.

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

		DL (ng/L) †						
Compound Name	Acronym	min	median	75th	98th	max		
				percentile	percentile			
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		

63 <sup>†</sup> 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 acrossbatches.

NAICS and	Description of Industry	Number of unique sites in NH
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

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

67 Notes: We used the North American Industrial Classification System (NAICS) codes and the US EPA

68 Facility Registry Service (FRS) codes that correspond to industries that are known to use and release

69 PFAS We identified the locations of these potential PFAS sources using for any time before October

70 2017, the year when the latest samples were collected.



73 Figure S2. Sensitivity analysis of different atmospheric buffer distances on industrial



### 76 S2. Detailed description of environmental predictors

77 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 78 79 density and sand/silt/clay content, (3) hydrologic variables such as precipitation and groundwater 80 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,<sup>2</sup> 81 but was not consistently collected in the NHDES sample campaign. We therefore used the 82 83 annual minimum depth to water table in gSSURGO as a proxy.<sup>3</sup> No statewide data were available for the groundwater table. Thus, we used elevation as a proxy. Datasets for NH were 84 85 accessed through the US Geologic Survey and NHDES websites in the format of raster files or 86 spatial shapefiles (Table S2). To assign independent variables to each well, we overlaid the raster 87 or shapefile that contained data for NH with the well location in a Geographic Information 88 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 89 90 recharge were available at the 1 km x1 km scale, and precipitation was available at 4 km x 4 km 91 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 92 the variable across all wells with available information. We chose this imputation method 93 because it preserves the mean distribution of variables.<sup>4</sup> 94

# 96 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	e Impacts				
Impact: Plastics and Rubber Products Manufacturing	0	0	1.11× 10 <sup>-3</sup>	6.22× 10 <sup>-2</sup>	3.42× 10 <sup>-1</sup>	9.23× 10 <sup>-1</sup>	km <sup>-1 [1]</sup>	15
Impact: Textiles Manufacturing and Related Activities	0	0	0	0	6.33× 10 <sup>-1</sup>	9.15× 10 <sup>-1</sup>	km <sup>-1</sup>	15
Impact: Airports	0	0	0	0	0	6.36× 10 <sup>-1</sup>	km <sup>-1</sup>	15
Impact: military sites	0	0	0	0	4.50× 10 <sup>-2</sup>	9.63× 10 <sup>-1</sup>	km <sup>-1</sup>	15
Impact: wastewater treatment plants	0	0	0	0	1.25× 10 <sup>-1</sup>	9.51× 10 <sup>-1</sup>	km <sup>-1</sup>	15
Impact: Potential sources	0	0	8.27× 10 <sup>-3</sup>	7.57× 10 <sup>-1</sup>	4.19	7.80	km <sup>-1</sup>	15
			Geologic	Variables				
Bedrock Type: Metasedimentary and Metavolcanic Rocks of the Merrimack Trough			Unitless <sup>[2]</sup>	16				
Depth to bedrock	6.00× 10 <sup>-5</sup>	30.8	30.8	30.8	41.0	76.0	meter <sup>[3]</sup>	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 <sup>[4]</sup>	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 <sup>-5</sup>	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 <sup>-5</sup>	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							
			Soil Geoc	hemistry					
Silt: Total - Mineral particles 0.002 mm - 0.05 mm in equivalent diameter as a weight percentage of <2 mm fraction	2.00× 10 <sup>-1</sup>	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 <sup>-2</sup>	9.50× 10 <sup>-1</sup>	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 <sup>-1</sup>	64.0	79.5	81.2	85.9	95.0	%	17	
Bulk Density at a water tension of 1/3 bar	5.50× 10 <sup>-2</sup>	1.14	1.31	1.47	1.63	1.79	g/mL	17	
Available Water Capacity <sup>[6]</sup>	4.02× 10 <sup>-3</sup>	8.80× 10 <sup>-2</sup>	1.19× 10 <sup>-1</sup>	1.32× 10 <sup>-1</sup>	3.17× 10 <sup>-1</sup>	5.30× 10 <sup>-1</sup>	vol. water/vol. soil	17	
Cation Exchange Capacity at pH 7.0, as estimated by the ammonium acetate method	5.01× 10 <sup>-3</sup>	9.00× 10 <sup>-2</sup>	7.65× 10 <sup>-1</sup>	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 x 10 <sup>4</sup>	1.23 x 10 <sup>4</sup>	1.31 x 10 <sup>4</sup>	2.50 x 10 <sup>4</sup>	1.59 x 10 <sup>5</sup>	g Carbon	17	
Soil thickness	7.20× 10 <sup>-1</sup>	14	20.4	45.6	105	149	cm	17	

<sup>[3]</sup> 10-meter resolution grid dataset.
<sup>[4]</sup> 4-kilometer resolution grid dataset.
<sup>[5]</sup> 1-kilometer resolution grid dataset
<sup>[6]</sup> The quantity of water that the soil is capable of storing 105

Notes:

<sup>&</sup>lt;sup>[1]</sup> 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.

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

108 Figure S3. Confusion matrix for categorical models (logistic regression and classification

109 random forest).

### 111 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.<sup>5</sup>

122 Performance of the continuous model (regression random forest) was moderate to poor across the five PFAS with pseudo-R<sup>2</sup> values ranging from 0.024 for PFOS to 0.52 for PFPeA 123 124 (Table S6). This performance is similar to modeling studies for other toxicants in groundwater 125 with a similar sample size (see SI Section S4 for more details). The lowest performance likely 126 reflects the limited detectable concentration data available for PFOS (n = 465). The sample size 127 for the regression random forest is much smaller than the categorical models due to the exclusion 128 of samples below detection. The intended purpose of this type of statistical model is as a 129 screening tool to prioritize field sampling. Thus, we conclude based on these results that 130 classification models that can use all available data are preferable. Classification random forest 131 models may be preferred over continuous models because they can use all data collected in 132 monitoring programs, avoiding poor performance for PFAS like PFOS in this study that had a 133 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 <sup>2</sup>	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. 136

Industry <sup>¶</sup> Plastics and rubber $1.2 \pm 0.13^{***}$ $0.57 \pm 0.12^{***}$ $0.88 \pm 0.11^{***}$ $0.18 \pm 0.09^*$ $0.20 \pm 0.11$ $0.33 \pm 0.10^{***}$ Textile manufacturing $-0.60 \pm 0.18^{**}$ $-0.52 \pm 0.13^{***}$ $0.42 \pm 0.10^{**}$ $-0.29 \pm 0.13^*$ $0.28 \pm 0.10^{**}$ Military sites $-0.53 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^*$ $0.40 \pm 0.11^{***}$ WWTP <sup>b</sup> $0.53 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^*$ $0.40 \pm 0.11^{***}$ Potential sources <sup>c</sup> $0.31 \pm 0.11^{***}$ $0.24 \pm 0.10^*$ $0.43 \pm 0.10^{***}$ $0.40 \pm 0.11^{***}$ Bedrock type <sup>‡</sup> $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $0.33 \pm 0.20^{*}$ $0.32 \pm 0.11^{**}$ $0.32 \pm 0.11^{**}$ Monthly precipitation potential $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20^{*}$ $0.44 \pm 0.19^{**}$ $0.34 \pm 0.10^{**}$ Market rable depth Groundwater recharge $0.26 \pm 0.11^{*}$ $0.26 \pm 0.09^{**}$ $0.19 \pm 0.11^{*}$ $0.0.$		PFPeA	PFHxA	PFHpA	PFOA	PFOS	PFAS
Plastics and rubber Textile manufacturing Military sites $1.2 \pm 0.13^{***}$ $0.57 \pm 0.12^{***}$ $0.88 \pm 0.11^{***}$ $0.18 \pm 0.09^*$ $0.20 \pm 0.11$ $0.33 \pm 0.10^{***}$ Military sites $-0.52 \pm 0.13^{***}$ $0.42 \pm 0.10^{**}$ $-0.29 \pm 0.13^*$ $0.28 \pm 0.09^*$ WWTP <sup>b</sup> $0.53 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^*$ $0.40 \pm 0.11^{***}$ Potential sources <sup>c</sup> $0.31 \pm 0.11^{**}$ $0.28 \pm 0.09^{**}$ $0.43 \pm 0.10^{***}$ $0.40 \pm 0.11^{***}$ Bedrock type <sup>‡</sup> $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $-0.37 \pm 0.10^{***}$ $0.32 \pm 0.11^{***}$ $0.32 \pm 0.11^{***}$ Monthly precipitation potential $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{***}$ Mater table depth Groundwater recharge $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{***}$ $-0.72 \pm 0.12^{***}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent silt $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$				Industry <sup>¶</sup>			
Textile manufacturing Military sites $-0.60 \pm 0.18^{**}$ $-0.52 \pm 0.13^{***}$ $0.42 \pm 0.10^{**}$ $-0.29 \pm 0.13^{*}$ $0.28 \pm 0.10^{**}$ WWTP <sup>b</sup> $0.53 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^{*}$ $0.40 \pm 0.11^{***}$ Potential sources <sup>6</sup> $0.31 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^{**}$ $0.43 \pm 0.10^{***}$ Bedrock type <sup>‡</sup> $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $0.32 \pm 0.12^{**}$ $0.32 \pm 0.11^{***}$ $0.32 \pm 0.11^{***}$ Monthly precipitation Low runoff potential $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{**}$ $0.34 \pm 0.10^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ Soil $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$	Plastics and rubber	$1.2 \pm 0.13^{***}$	$0.57 \pm 0.12^{\ast\ast\ast}$	$0.88 \pm 0.11^{\ast\ast\ast}$	$0.18\pm0.09^{\ast}$	$0.20\pm0.11$	$0.33 \pm 0.10^{\ast\ast\ast}$
manufacturing $-0.50 \pm 0.18^{**}$ $-0.52 \pm 0.13^{***}$ $0.42 \pm 0.10^{**}$ $-0.22 \pm 0.13^{***}$ $0.28 \pm 0.10^{***}$ Military sites $-0.22 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^{*}$ $0.40 \pm 0.11^{***}$ WWTP <sup>b</sup> $0.53 \pm 0.13^{***}$ $0.44 \pm 0.11^{***}$ $0.28 \pm 0.09^{**}$ $0.19 \pm 0.09^{*}$ $0.40 \pm 0.11^{***}$ Potential sources <sup>6</sup> $0.53 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Monthly       precipitation $-0.29 \pm 0.12^{*}$ $0.33 \pm 0.20^{*}$ $0.32 \pm 0.11^{**}$ $0.34 \pm 0.10^{**}$ Monthly $-0.29 \pm 0.12^{*}$ $0.33 \pm 0.20^{*}$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20^{*}$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Stope gradient $0.21 \pm 0.13^{*}$ $0.19 \pm 0.12^{*}$ $-0.19 \pm 0.10^{*}$ $-0.17 \pm 0.10^{**}$ Percent slit $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11^{**}$ $-0.39 \pm 0.12^{**$	Textile	0. (0. ). 0. 10**	0.50 . 0.10***		0.40 . 0.10**	0.00 + 0.12*	0.00 1.0**
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	Military sites	$-0.60 \pm 0.18$	$-0.52 \pm 0.13$		$0.42 \pm 0.10^{-4}$	$-0.29 \pm 0.13^{\circ}$	$0.28 \pm 0.10^{\circ}$
WW11 $0.53 \pm 0.13^{**}$ $0.44 \pm 0.11^{**}$ $0.28 \pm 0.09^{*}$ $0.19 \pm 0.09^{*}$ $0.40 \pm 0.11^{**}$ Potential sources <sup>E</sup> $0.31 \pm 0.11^{**}$ $0.24 \pm 0.10^{*}$ $0.43 \pm 0.10^{***}$ Geo       Geo       1.1 \pm 0.15^{***} $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $-0.32 \pm 0.12^{**}$ $-0.37 \pm 0.10^{***}$ $0.38 \pm 0.11^{***}$ Monthly       precipitation $-0.29 \pm 0.12^{*}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Groundwater $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{**}$ $-0.72 \pm 0.12^{**}$ $-0.17 \pm 0.10^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10^{**}$ Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11^{***}$ $-0.39 \pm 0.12^{***}$ Percent sand $-0.25 \pm 0.15^{*}$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$	W/W/TDb	0.52 . 0.12***	0 44 + 0 11***	0.00 + 0.00**	$-0.22 \pm 0.13$		0 40 1 0 1 1 ***
Potential sourcest $0.31 \pm 0.11^{**}$ $0.24 \pm 0.10^{**}$ $0.43 \pm 0.10^{***}$ Bedrock type <sup>‡</sup> $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Monthly       precipitation $-0.39 \pm 0.12^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{**}$ $0.34 \pm 0.10^{***}$ Mow runoff $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Groundwater $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{***}$ $0.17 \pm 0.10$ $0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Solpe gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ Bulk density $0.55 \pm 0.25^{*}$ $0.58 \pm 0.28^{*}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$	W W II Dotontial cources€	$0.53 \pm 0.13$	$0.44 \pm 0.11$	$0.28 \pm 0.09^{\circ}$	$0.19 \pm 0.09^{\circ}$		$0.40 \pm 0.11$
GeoBedrock type‡ $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $-0.32 \pm 0.12^{**}$ $-0.37 \pm 0.10^{***}$ $-0.38 \pm 0.11^{***}$ Monthly precipitation Low runoff potential $-0.29 \pm 0.12^{*}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth Groundwater recharge $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ SoilPercent clay Percent silt $-0.43 \pm 0.16^{**}$ Percent sand Bulk density $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$	Fotential sources		$0.31 \pm 0.11^{++}$	$0.24 \pm 0.10^{\circ}$		$0.43 \pm 0.10^{+++}$	
Bedrock type* $0.64 \pm 0.16^{***}$ $0.69 \pm 0.17^{***}$ $1.1 \pm 0.15^{***}$ $1.3 \pm 0.12^{***}$ $0.92 \pm 0.11^{***}$ Depth to bedrock $0.19 \pm 0.12^{***}$ $-0.32 \pm 0.12^{**}$ $-0.37 \pm 0.10^{***}$ $-0.38 \pm 0.11^{***}$ Monthly       precipitation $-0.29 \pm 0.12^{*}$ $0.32 \pm 0.10^{***}$ $0.32 \pm 0.11^{***}$ $0.32 \pm 0.11^{***}$ Low runoff $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{***}$ $0.34 \pm 0.10^{**}$ Groundwater $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$	D 1 1 . *			Geo			
Depth to bedrock $0.19 \pm 0.12^{***}$ $-0.32 \pm 0.12^{**}$ $-0.37 \pm 0.10^{***}$ $-0.38 \pm 0.11^{***}$ Monthly       Precipitation $-0.29 \pm 0.12^{*}$ $0.32 \pm 0.11^{**}$ $0.32 \pm 0.11^{**}$ Monthly $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Groundwater $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{***}$ $0.17 \pm 0.10^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.58 \pm 0.28^{*}$ $0.58 \pm 0.28^{*}$ $0.58 \pm 0.28^{*}$	Bedrock type <sup>4</sup>	$0.64 \pm 0.16^{***}$	$0.69 \pm 0.17^{***}$	$1.1 \pm 0.15^{***}$	$1.3 \pm 0.12^{***}$		$0.92 \pm 0.11^{***}$
HydroMonthly precipitation $-0.29 \pm 0.12^*$ $0.32 \pm 0.11^{**}$ Low runoff potential $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^*$ $0.34 \pm 0.10^{**}$ Water table depth Groundwater recharge $0.26 \pm 0.11^*$ $-0.72 \pm 0.12^{***}$ $0.34 \pm 0.10^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ $-0.17 \pm 0.10$ SoilPercent clayPercent sald $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^*$ $0.58 \pm 0.28^*$ $0.32 \pm 0.10^{***}$	Depth to bedrock	$0.19 \pm 0.12^{***}$		$-0.32 \pm 0.12^{**}$	$-0.37 \pm 0.10^{***}$		$-0.38 \pm 0.11^{***}$
Monthly precipitation $-0.29 \pm 0.12^*$ $0.32 \pm 0.11^{**}$ Low runoff potential $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^*$ $0.34 \pm 0.10^{**}$ Water table depth Groundwater recharge $0.26 \pm 0.11^*$ $-0.72 \pm 0.12^{***}$ $0.26 \pm 0.09^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ $-0.17 \pm 0.10$ Soil       Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$	NC 41			Hydro			
Low function $0.53 \pm 0.16^{**}$ $0.57 \pm 0.18^{**}$ $0.33 \pm 0.20$ $0.45 \pm 0.19^{*}$ $0.34 \pm 0.10^{**}$ Water table depth $0.26 \pm 0.11^{*}$ $-0.72 \pm 0.12^{***}$ $0.26 \pm 0.09^{**}$ Groundwater $0.26 \pm 0.09^{**}$ $0.26 \pm 0.09^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^{*}$ $-0.17 \pm 0.10$ Percent clay $-0.43 \pm 0.16^{**}$ $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent silt $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.34 \pm 0.10^{***}$	Monthly precipitation		$\textbf{-}0.29\pm0.12^{*}$			$0.32 \pm 0.11^{**}$	
Water table depth $0.26 \pm 0.11^*$ $-0.72 \pm 0.12^{***}$ Groundwater $0.26 \pm 0.09^{**}$ $0.26 \pm 0.09^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ Soil       Soil         Percent clay $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ Percent silt $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$	potential	$0.53 \pm 0.16^{**}$	$0.57 \pm 0.18^{**}$	$0.33 \pm 0.20$		$0.45 \pm 0.19^{*}$	$0.34 \pm 0.10^{**}$
Groundwater $0.26 \pm 0.09^{**}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ $-0.17 \pm 0.10$ Soil         Percent clay         Percent silt $-0.43 \pm 0.16^{**}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^*$ $0.25 \pm 0.15^{***}$ $0.58 \pm 0.28^*$	Water table depth			$0.26 \pm 0.11^{*}$		$-0.72 \pm 0.12^{***}$	
recharge $0.26 \pm 0.09^{14}$ Slope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ $-0.17 \pm 0.10$ Soil       Soil $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent slit $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^*$ $0.34 \pm 0.10^{***}$	Groundwater				0.00**		
Stope gradient $0.21 \pm 0.13$ $0.19 \pm 0.12$ $-0.19 \pm 0.10$ $-0.20 \pm 0.13^*$ $-0.17 \pm 0.10$ Soil       Soil $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent silt $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.34 \pm 0.10^{***}$	recharge				$0.26 \pm 0.09^{+1}$		
Soil         Percent clay $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent silt $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$	Slope gradient	$0.21 \pm 0.13$	$0.19 \pm 0.12$	~ •	$-0.19 \pm 0.10$	$-0.20 \pm 0.13^*$	$-0.17 \pm 0.10$
Percent clay $-0.56 \pm 0.12^{***}$ $0.19 \pm 0.11$ $-0.39 \pm 0.12^{***}$ Percent silt $-0.43 \pm 0.16^{**}$ $0.34 \pm 0.10^{***}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.58 \pm 0.28^{*}$				Soil			
Percent silt $-0.43 \pm 0.16^{**}$ Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$	Percent clay				$-0.56 \pm 0.12^{***}$	$0.19 \pm 0.11$	$-0.39 \pm 0.12^{***}$
Percent sand $-0.25 \pm 0.15$ $-0.69 \pm 0.30$ $0.34 \pm 0.10^{***}$ Bulk density $0.58 \pm 0.28^{*}$ $0.34 \pm 0.10^{***}$	Percent silt		$-0.43 \pm 0.16^{**}$				
Bulk density $0.58 \pm 0.28^*$	Percent sand	$-0.25 \pm 0.15$	$\textbf{-0.69} \pm 0.30$		$0.34 \pm 0.10^{\ast\ast\ast}$		
0.38 ± 0.28	Bulk density		$0.58\pm0.28^{\ast}$				
Available water capacity $0.37 + 0.13^{**}$ $0.34 + 0.12^{**}$ $-0.32 + 0.12^{**}$	Available water		$0.37 \pm 0.13^{**}$	$0.34 \pm 0.12^{**}$		$-0.32 \pm 0.12^{**}$	
Organic carbon $0.57 \pm 0.15$ $0.54 \pm 0.12$	Organic carbon		$0.57 \pm 0.15$	$0.54 \pm 0.12$		$-0.52 \pm 0.12$	
content $0.40 \pm 0.13^{**}$	content	$0.40 \pm 0.13^{**}$					
Soil thickness $-0.19 \pm 0.13$ $-0.15 \pm 0.09$ $-0.13 \pm 0.09$	Soil thickness	$\textbf{-0.19} \pm 0.13$			$\textbf{-0.15} \pm 0.09$		$\textbf{-0.13} \pm 0.09$
Saturated	Saturated						
conductivity $0.57 \pm 0.20^{***}$ $-0.82 \pm 0.20^{***}$	conductivity			$0.57 \pm 0.20^{***}$		$-0.82 \pm 0.20^{***}$	
Cation exchange	Cation exchange						
$\frac{\text{capacity}}{\text{C} \text{Statistics}^{\dagger}}$	C Statistics <sup>†</sup>		<b>A</b>		A	<b>•</b>	$0.20 \pm 0.12$
C-Stausucs* 0.70 0.69 0.74 0.68 0.65 0.66	AUROC	0.70	0.69	0.74	0.68	0.65	0.66
$(95\% \text{ CI})^{\text{f}} \qquad 0.68 \ (0.65, \ 0.72) \qquad 0.67 \ (0.65, \ 0.70) \qquad 0.72 \ (0.70, \ 0.75) \qquad 0.66 \ (0.64, \ 0.68) \qquad 0.63 \ (0.61, \ 0.64) \qquad 0.64 \ (0.62, \ 0.66)$	(95% CI) <sup>£</sup>	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	n	1617	1725	2253	2373	2376	2379

## 139 Table S5. Standardized<sup>a</sup> logistic regression model coefficients (± standard error).

<sup>a</sup> Standardized coefficients are unitless, normalized values so that the variances of dependent and independent

141 variables are equal to 1 and can be compared because they reflect how many standard deviations PFAS

142 concentrations will change per standard deviation in the predictor variable.

<sup>b</sup> WWTP = wastewater treatment plants.

144 Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.01

145 Variables not selected by the logistic regression models are pH calculated by the 1:1 soil-water ratio, and available

146 water storage from the surface to reported depth of soil profile.

147 ¶ 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.

<sup>‡</sup>Bedrock type = Metasedimentary and Metavolcanic Rocks of the Merrimack Trough

- 150 <sup>†</sup> 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.
- 151 152 <sup>e</sup> Potential sources: sources considered include semiconductor, printing, metal plating, textile mills, petroleum and
- 153 coal products manufacturing, chemical manufacturing.
- 154 <sup>‡</sup>Concordance (C) statistics are used to assess model discrimination, which means how well the model can separate 155 the wells with detectable concentrations from those with non-detect. C statistics ranges from 0.5 to 1, and values
- 156 around 0.7 generally indicate a good model.
- 157 <sup>f</sup>Area under the Receiver Operating Characteristics curve (AUROC) is used to evaluate the classification model's
- 158 performance. The mean AUROC and its 95% confidence interval (CI) is calculated by 10-fold cross validation.
- 159

#### Table S6. Tuning of hyperparameters in the random forest model and area under the

#### **Receiver Operating Characteristics curve (AUROC)**

### 

Classification Random Forest <sup>1</sup>							
	Wors	ormance	Best performance				
	$mtry^2$	ns <sup>3</sup>	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:

<sup>1</sup>The number of trees used was 1000 given that the out-of-bag error converged by then across all compounds.

<sup>2</sup>Number of features randomly sampled at each node out of an original 26 features. <sup>3</sup>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).

# 176 S4. Review of previous literature on machine learning models for drinking water177 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 183 184 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 185 average an accuracy rate of 79% (range: 37% - 92%), which is similar to the results shown here. 186 187 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 188 189 classification threshold is set to maximize accuracy. In practice, this can be adjusted so that some 190 true negative rate is sacrificed in order to reduce false negatives.

191 Performance of the regression random forest models in this study was comparable to those for predicting groundwater nitrate contamination from prior work.<sup>2, 6, 7</sup> Similar to the 192 193 classification random forest model, groundwater recharge and monthly precipitation were 194 consistently among the most important predictors for all five PFAS modeled, in addition to 195 impacts from industrial sources such as plastics manufacturing, printing and textile 196 manufacturing. A model for groundwater nitrate concentrations in Germany with a comparable sample size to ours (1890 wells) had an R<sup>2</sup> of 0.54.<sup>6</sup> In Iowa and North Carolina, continuous 197 models similarly had low predictive performance ( $R^2 < 0.33$ ) for predicting groundwater nitrate 198

- 199 concentrations in 22,000 private wells sampled.<sup>7</sup> Common challenges for such modeling include
- 200 dealing with a low fraction of samples with detectable concentrations, and limited data on some
- 201 important spatial predictors, particularly those relating to local groundwater flow conditions that
- are not always available at statewide or larger spatial scales.<sup>8</sup>

size     technique       1     Anning     2012     Arizona, California, Colorado, Nevada, New     Not     Random     Correct 48.6%, 92.5.8%, Underpredicted, 25.6%     9       2     Anning     2012     Arizona, California, Colorado, Nevada, New     NR     Random     Correct 36.7%, 9     9       2     Anning     2012     Arizona, Colorado, Nevada, New     NR     Random     Correct 36.7%, 9     9       3     Nolan     2014     Central     Nitrate     314     Logistic     Predict       3     Nolan     2014     Central     Nitrate     318     Random     Predict       4     Nolan     2014     Central     Nitrate     318     Random     Predict       5     Nolan     2014     Central     Nitrate     318     Random     Predict       5     Nolan     2014     Central     Nitrate     318     Random     Predict       6     Nolan     2014     Central     Nitrate     928     Logistic     Predict       7     Nolan     2014     Central     Nitrate     937     Random     Predict       7     Nolan     2014     Central     Nitrate     937     Random     Predict	No.	Author	Year	Location	Compound	Sample	ML	Performance	Ref
1       Anning 2012       Arizona, Nitrate Not Random Correct 86.0%, 0       California, reported forest 0verpredicted 25.8%, 0         2       Anning 2012       Arizona, Arsenic NR Random Correct 36.7%, 0       Underpredicted, 25.6%         2       Anning 2012       Arizona, Arsenic NR Random Correct 36.7%, 0       3.5%, 0         2       Anning 2012       Arizona, Arsenic NR Random Correct 36.7%, 0       3.5%, 0         2       Anning 2012       Arizona, Arsenic NR Random Correct 36.7%, 0       3.3.5%, 0         3       Nolan 2014       Central Nitrate 314       Logistic Predict 2         4       Nolan 2014       Central Nitrate 318       Random regression nitrate>4mg/L; accuracy 69.7%, sensitivity 60.0%, sensitivity 82.1%, specificity 77.3%         5       Nolan 2014       Central Nitrate 318       Random Predict 2         6       Nolan 2014       Central Nitrate 928       Logistic Predict 2         7       Nolan 2014       Central Nitrate 927       Random Predict 2         7       Nolan 2014       Central Nitrate 927       Random Predict 2         7       Nolan 2014       Central Nitrate 927       Random Predict 2         8       Nolan 201						size	technique		
	1	Anning	2012	Arizona,	Nitrate	Not	Random	Correct 48.6%,	9
Colorado, Nevada, New Mexico, and Utah       (NR)       25.8%, Underpredicted, 25.6%         2       Anning       2012       Arizona, California, Colorado, Nevada, New Mexico, and Utah       NR       Random       Correct 36.7%, 9       9         3       Nolan       2014       Central Valley, CA       Nitrate       314       Logistic regression       Predict       2         4       Nolan       2014       Central Valley, CA       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central Valley, CA       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central Valley, CA       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central Valley, CA       Nitrate       928       Logistic regression       mitrate>49%, sensitivity 84.2%, sensitivity				California,		reported	forest	Overpredicted	
Nevada, New       Unda       25.6%         Mexico, and       25.6%         2       Anning       2012       Arizona, Arsenic       NR       Random       Correct 36.7%, 9         2       Anning       2012       Arizona, Arsenic       NR       Random       Correct 36.7%, 9         2       Anning       2012       Arizona, Arsenic       NR       Random       Correct 36.7%, 9         2       Anning       2012       Arizona, Arsenic       NR       Random       Overpredicted         2       Colorado, New       Utah       29.8%       Underpredicted       29.8%         3       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict				Colorado,		(NR)		25.8%,	
Mexico, and Utah     25.6%       2     Anning     2012     Arizona, Arsenic Colorado, Nevada, New Mexico, and     NR     Random forest     Correct 36.7%, 0.83.5%, Underpredicted     9       3     Nolan     2014     Central Valley, CA     Nitrate     314     Logistic regression     Predict     2       4     Nolan     2014     Central Valley, CA     (shallow well)     regression     nitrate>40.0%, sensitivity 60.0%, specificity 70.4%     2       5     Nolan     2014     Central Valley, CA     Nitrate     318     Random forest     regression     accuracy 63.7%, sensitivity 60.0%, specificity 70.4%       5     Nolan     2014     Central Valley, CA     Nitrate     318     Random forest     regression     accuracy 71.7%, sensitivity 82.5%       6     Nolan     2014     Central Valley, CA     Nitrate     928     Logistic regression     regression       7     Nolan     2014     Central Valley, CA     Nitrate     937     Random Predict     2       7     Nolan     2014     Central Valley, CA     Nitrate     937     Random Predict     2       8     Nolan     2014     Central Valley, CA     Nitrate     937     Random Predict     2       8     Nolan     2014     Central Val				Nevada, New				Underpredicted,	
Utah       Varizona, Arsenic       NR       Random       Correct 36.7%, 9         2       Anning       2012       Arizona, Arsenic       NR       Random       Correct 36.7%, 9         3       Nolan       2014       California, Colorado, New       Underpredicted       33.5%, 10         3       Nolan       2014       Central       Nitrate       314       Logistic       Predict       2         3       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nit				Mexico, and				25.6%	
2       Anning       2012       Arizona, Arsenic       NR       Random       Correct 36.7%, 9         California, Colorado, Nevada, New       Forest       Overpredicted       33.5%, 9         Mexico, and       29.8%       Utah       29.8%         3       Nolan       2014       Central       Nitrate       314       Logistic       Predict       2         3       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predi				Utah					_
California, forest Overpredicted Colorado, 33.5%, Nevada, New Mexico, and 29.8% Underpredicted 28.8% 3 Nolan 2014 Central Nitrate 314 Logistic Predict 2 Valley, CA (shallow regression nitrate>4mg/L; accuracy 69.7%, specificity 70.4% specificity 70.4% Valley, CA (shallow regression nitrate>4mg/L; accuracy 69.7%, sensitivity 60.0%, specificity 70.4% Specificity 71.7% Sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 62.1%, sensitivity 84.2%, specificity 73.8% 66 Nolan 2014 Central Nitrate 928 Logistic Predict 2 Valley, CA (deep well) regression accuracy 68.9%, sensitivity 7 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 29.1%, specificity 73.8% 8 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 29.1%, specificity 93.1%, sensitivity 29.1%, specificity 93.1%, sensitivity 29.1%, specificity 93.1%, sensitivity 29.1%, specificity 93.1%, sensitivity 29.1%, specificity 93.3%, sensitivity 51.3%, sensitivity 51.3%, se	2	Anning	2012	Arizona,	Arsenic	NR	Random	Correct 36.7%,	9
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$				California,			forest	Overpredicted	
Nevada, New Mexico, and Utah 3 Nolan 2014 Central Nitrate 314 Logistic Predict $^2$ Sensitivity 69,0%, sensitivity 61,1%, specificity 70,4% Valley, CA (shallow forest nitrate-4mg/L; sensitivity 81,2%, sensitivity 82,1%, sensitivity 84,2%, sensitivity 84,				Colorado,				33.5%,	
Mexico, and Utah29.8%3Nolan2014CentralNitrate314LogisticPredict23Nolan2014CentralNitrate314LogisticPredict24Nolan2014CentralNitrate318RandomPredict25Nolan2014CentralNitrate318RandomPredict25Nolan2014CentralNitrate318RandomPredict25Nolan2014CentralNitrate318RandomPredict26Nolan2014CentralNitrate318RandomPredict26Nolan2014CentralNitrate928LogisticPredict27Nolan2014CentralNitrate928LogisticPredict27Nolan2014CentralNitrate937RandomPredict27Nolan2014CentralNitrate937RandomPredict28Nolan2014CentralNitrate937RandomPredict29Rodriguez-2014Granada city,Nitrate175RandomPredict29Rodriguez-2014Granada city,Nitrate175LogisticPredict nitrate1010Rodriguez-2014Granada city,Nitrate175Logistic<				Nevada, New				Underpredicted	
3       Nolan       2014       Central       Nitrate       314       Logistic       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2				Mexico, and				29.8%	
5       Notan       2014       Central       Nutrate       514       Logishic       intrate>4mg/L; accuracy 69.7%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 69.0%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.5%         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central <td>2</td> <td>Nolan</td> <td>2014</td> <td>Central</td> <td>Nitrata</td> <td>214</td> <td>Logistic</td> <td>Dradiat</td> <td>2</td>	2	Nolan	2014	Central	Nitrata	214	Logistic	Dradiat	2
4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2	3	Inolali	2014	Velley CA	(shallow	514	Logistic	ritroto /mg/L:	
<ul> <li>Weil)</li> <li>accuracy 05.7%, specificity 60.0%, specificity 70.4%</li> <li>Valley, CA (shallow well)</li> <li>classification accuracy 71.7%, sensitivity 65.1%, specificity 77.3%</li> <li>Nolan 2014</li> <li>Central Nitrate 318</li> <li>Random Predict 2</li> <li>Sensitivity 65.1%, specificity 77.3%</li> <li>Solan 2014</li> <li>Central Nitrate 318</li> <li>Random Predict 2</li> <li>Sensitivity 65.1%, specificity 77.3%</li> <li>Solan 2014</li> <li>Central Nitrate 318</li> <li>Random Predict 2</li> <li>Valley, CA (shallow well)</li> <li>regression accuracy 68.9%, sensitivity 84.2%, specificity 55.8%</li> <li>Solan 2014</li> <li>Central Nitrate 928</li> <li>Logistic Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression intrate&gt;4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 91.3%, specificity 92.1%, specificity 96.3%</li> <li>Nolan 2014</li> <li>Central Nitrate 937</li> <li>Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; accuracy 81.5%, sensitivity 25.1%, specificity 96.3%</li> <li>specificity 96.3%&lt;</li></ul>				valley, CA	(shahow		regression	$\frac{1111 \text{ ate}}{4111 \text{ g/L}}$	
4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         6       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2					wen)			accuracy 09.770,	
4       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         5       Nolan       2014       Central       Nitrate       318       Random       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2								sensitivity 09.076,	
<ul> <li>Valley, CA (shallow vell)</li> <li>Valley, CA (shallow vell)</li> <li>Classification intrate-Amg/L; accuracy 71.7%, sensitivity 65.1%, specificity 77.3%</li> <li>Nolan 2014 Central Nitrate 318 Random Predict 2</li> <li>Valley, CA (shallow vell)</li> <li>regression accuracy 68.9%, sensitivity 84.2%, specificity 55.8%</li> <li>Nolan 2014 Central Nitrate 928 Logistic Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression nitrate-Amg/L; accuracy 80.8%, sensitivity 29.1%, specificity 52.8%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression accuracy 81.2%, sensitivity 25.1%, sensitivity 21.1%, specificity 96.3%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate-4mg/L; accuracy 81.2%, sensitivity 51.3%, specificity 95.3%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate-4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 99.3%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 1</li> <li>Galiano Spain Forest 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 1</li> <li>Galiano Spain Forest 75 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10<td>1</td><td>Nolan</td><td>2014</td><td>Central</td><td>Nitrata</td><td>218</td><td>Pandom</td><td>Specificity 70.476</td><td>2</td></li></ul>	1	Nolan	2014	Central	Nitrata	218	Pandom	Specificity 70.476	2
<ul> <li>Valley, CA (shallow well)</li> <li>classification classification</li> <li>sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, sensitivity 65.1%, specificity 77.3%</li> <li>Valley, CA (shallow well)</li> <li>regression accuracy 68.9%, sensitivity 84.2%, specificity 55.8%</li> <li>Nolan 2014 Central Nitrate 928 Logistic Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression nitrate&gt;4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 92.9%, sensitivity 29.1%, specificity 94.9%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression nitrate&gt;4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 96.3%</li> <li>Sensitivity 25.1%, specificity 96.3%</li> <li>Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression accuracy 81.2%, sensitivity 51.3%, specificity 96.3%</li> <li>Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 96.3%</li> <li>Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 96.3%</li> <li>Somg/L; 80.46%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Lo</li></ul>	4	Inolali	2014	Velley CA	(shallow	510	forest	ritroto /mg/L:	
				valley, CA	(shanow		alassification	$\frac{1111 \text{ ate}}{4111 \text{ g/L}}$	
5 Nolan 2014 Central Nitrate 318 Random Predict <sup>2</sup> Valley, CA (shallow forest nitrate>4mg/L; accuracy 68.9%, sensitivity 84.2%, specificity 55.8% 6 Nolan 2014 Central Nitrate 928 Logistic Predict <sup>2</sup> Valley, CA (deep well) regression nitrate>4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9% 7 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 89.7% 9 Rodriguez- Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup> Galiano Spain regression >50mg/L; 80.46% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = <sup>11</sup>					wen)		classification	accuracy / 1.7%	
5 Nolan 2014 Central Nitrate 318 Random Predict 2 Valley, CA (shallow well) regression accuracy 68.9%, sensitivity 84.2%, specificity 55.8% 6 Nolan 2014 Central Nitrate 928 Logistic Predict 2 Valley, CA (deep well) regression nitrate>4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9% 7 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.2%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 51.3%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.2%, sensitivity 25.1%, specificity 95.3% 8 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; accuracy 81.5%, sensitivity 51.3%, specificity 96.3% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 10 Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain regression >50mg/L; 80.46% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = 11								sensitivity 03.176,	
3       Notan       2014       Central       Nutate       318       Kandoni       Intrate>4mg/L;         4       well)       regression       accuracy 68.9%,       sensitivity 84.2%,       specificity 55.8%         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         6       Nolan       2014       Central       Nitrate       928       Logistic       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         9 <td>5</td> <td>Nolon</td> <td>2014</td> <td>Control</td> <td>Nitroto</td> <td>210</td> <td>Dandom</td> <td>Specificity 77.576</td> <td>2</td>	5	Nolon	2014	Control	Nitroto	210	Dandom	Specificity 77.576	2
<ul> <li>Valley, CA (shallow well)</li> <li>regression accuracy 68.9%, sensitivity 84.2%, specificity 55.8%</li> <li>6 Nolan 2014 Central Nitrate 928 Logistic Predict 2</li> <li>Valley, CA (deep well)</li> <li>regression nitrate&gt;4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9%</li> <li>7 Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; accuracy 80.8%, sensitivity 25.1%, specificity 96.3%</li> <li>8 Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3%</li> <li>8 Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; classification accuracy 81.5%, sensitivity 51.3%, specificity 96.3%</li> <li>8 Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; sensitivity 51.3%, specificity 96.3%</li> <li>9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10</li> <li>Galiano Spain forest &gt;50mg/L; 80.46%</li> <li>10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Galiano Spain 73.56%</li> <li>11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R<sup>2</sup> = 11</li> </ul>	3	INOIAII	2014	Velley CA	(shallow	518	forest	pitroto/mg/L:	-
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$				valley, CA	(shanow		rogragion	mulate=4mg/L,	
6 Nolan 2014 Central Nitrate 928 Logistic Predict 2 Valley, CA (deep well) regression nitrate>4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9% 7 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 10 Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain 75 Logistic Predict nitrate 10 Galiano Spain 73.56% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = 11					wen)		regression	accuracy 08.976,	
6 Nolan 2014 Central Nitrate 928 Logistic Predict 2 Valley, CA (deep well) regression nitrate>4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9% 7 Nolan 2014 Central Nitrate 937 Random Predict 2 Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 25.1%, sensitivity 55.3% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict 10 Galiano Spain forest sitrate>4mg/L; regression accuracy 81.5%, sensitivity 51.3%, specificity 89.7% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10 Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain regression >50mg/L; accuracy is 73.56% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = 11								specificity 55.8%	
<ul> <li>Notan 2014 Central Nitrate 925 Ebgistic Tredict Valley, CA (deep well)</li> <li>regression nitrate&gt;4mg/L; accuracy 80.8%, sensitivity 29.1%, specificity 94.9%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; classification accuracy 81.2%, sensitivity 51.3%, specificity 96.3%</li> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; regression accuracy 81.5%, sensitivity 51.3%, specificity 89.7%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10</li> <li>Galiano Spain</li> <li>forest &gt;50mg/L; 80.46%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Galiano Spain</li> <li>regression &gt;50mg/L; accuracy is 73.56%</li> <li>Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R<sup>2</sup> = 11</li> </ul>	6	Nolan	2014	Central	Nitrata	028	Logistic	Predict	2
<ul> <li>Valley, CA (deep well)</li> <li>Regression infrate-4ng/L; accuracy 80.8%, sensitivity</li> <li>29.1%, specificity</li> <li>94.9%</li> <li>7 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup></li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3%</li> <li>8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup></li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; regression accuracy 81.5%, sensitivity 51.3%, specificity 96.3%</li> <li>8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup></li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L; regression accuracy 81.5%, sensitivity 51.3%, specificity 80.7%</li> <li>9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate <sup>10</sup></li> <li>Galiano Spain forest &gt;50mg/L; 80.46%</li> <li>10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Galiano Spain regression &gt;50mg/L; accuracy is 73.56%</li> <li>11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R<sup>2</sup> = <sup>11</sup></li> </ul>	0	INOIAII	2014	Valley CA	(deen well)	920	regression	nitrate>/mg/L:	
7 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate <sup>10</sup> Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup> Galiano Spain regression >50mg/L; 80.46% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = <sup>11</sup>				valley, CA	(deep wen)		regression	$\frac{11110}{2} + \frac{112}{12}$	
7 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; classification accuracy 81.2%, sensitivity 25.1%, specificity 96.3% 8 Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup> Valley, CA (deep well) forest nitrate>4mg/L; regression accuracy 81.5%, sensitivity 51.3%, specificity 89.7% 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate <sup>10</sup> Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup> Galiano Spain 73.56% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = <sup>11</sup>								sensitivity	
7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         9       Rodriguez-       2014       Granada city,       Nitrate       175       Random       Predict nitrate       10         6aliano       Spain       forest       regression       >50mg/L; 80.46%       10         10       Rodriguez-       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         Galiano       Spain       regression       >50mg/L;       accuracy is       73.56%         11								20.1% specificity	
7       Nolan       2014       Central       Nitrate       937       Random       Predict       2         Valley, CA       (deep well)       forest       nitrate>4mg/L;       accuracy 81.2%,       sensitivity 25.1%,         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         9       Rodriguez-       2014       Granada city,       Nitrate       175       Random       Predict nitrate       10         9       Rodriguez-       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez-       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         Galiano       Spain       regression       >50mg/L;       accuracy is       73.56% </td <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>27.170, specificity 94.9%</td> <td></td>								27.170, specificity 94.9%	
<ul> <li>Valley, CA (deep well)</li> <li>Valley, CA (deep</li></ul>	7	Nolan	2014	Central	Nitrate	937	Random	Predict	2
<ul> <li>Nolan 2014 Central Nitrate 937 Random Predict <sup>2</sup></li> <li>Valley, CA (deep well)</li> <li>Valley, CA (deep well)</li> <li>Valley, CA (deep well)</li> <li>forest nitrate&gt;4mg/L;</li> <li>regression accuracy 81.5%,</li> <li>sensitivity 51.3%,</li> <li>specificity 89.7%</li> <li>Random Predict nitrate <sup>10</sup></li> <li>Galiano Spain</li> <li>Galiano Spain</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Spain Forest -50mg/L; 80.46%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Spain Forest -50mg/L; 80.46%</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate <sup>10</sup></li> </ul>	/	ivoluli	2014	Valley CA	(deen well)	))	forest	nitrate> $4mg/I$ ·	
<ul> <li>Nolan 2014 Central Nitrate 937 Random Predict 2</li> <li>Valley, CA (deep well)</li> <li>Forest nitrate&gt;4mg/L;</li> <li>regression accuracy 81.5%,</li> <li>sensitivity 51.3%,</li> <li>specificity 89.7%</li> <li>Random Predict 10</li> <li>Galiano Spain</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10</li> <li>Galiano Spain</li> <li>Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10</li> <li>Galiano Spain</li> <li>Spain</li> <li>Sp</li></ul>				valley, en	(deep wen)		classification	accuracy 81.2%	
8       Nolan 2014       Central Nitrate       937       Random       Predict       2         Valley, CA       (deep well)       forest nitrate>4mg/L;       regression       accuracy 81.5%, sensitivity 51.3%, specificity 89.7%         9       Rodriguez- 2014       Granada city, Nitrate       175       Random       Predict nitrate       10         9       Rodriguez- 2014       Granada city, Nitrate       175       Random       Predict nitrate       10         10       Rodriguez- 2014       Granada city, Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- 2014       Granada city, Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- 2014       Granada city, Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- 2014       Granada city, Nitrate       175       Logistic       Predict nitrate       10         11       Singh 2014       Indo-       Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11							clussification	sensitivity 25.1%	
8       Nolan       2014       Central       Nitrate       937       Random       Predict       2         Valley, CA       (deep well)       forest       nitrate>4mg/L;       regression       accuracy 81.5%,       sensitivity 51.3%,       specificity 89.7%         9       Rodriguez-       2014       Granada city,       Nitrate       175       Random       Predict nitrate       10         9       Rodriguez-       2014       Granada city,       Nitrate       175       Random       Predict nitrate       10         10       Rodriguez-       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         Galiano       Spain       forest       >50mg/L; 80.46%       10       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         Galiano       Spain       regression       >50mg/L;       accuracy is       73.56%         11       Singh       2014       Indo-       Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11								specificity 96 3%	
Valley, CA (deep well) 9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10 Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain 73.56% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = 11 Gauge Computing State 11 Computing State 11 State 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R <sup>2</sup> = 11 State 11	8	Nolan	2014	Central	Nitrate	937	Random	Predict	2
<ul> <li>9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10 Galiano Spain forest &gt;50mg/L; 80.46%</li> <li>10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain regression &gt;50mg/L; 80.46%</li> <li>11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, R<sup>2</sup> = 11 Galiano Chemical 409 Decision tree Test dataset, R<sup>2</sup> = 11</li> </ul>	0	rtoluli	2011	Valley CA	(deen well)	201	forest	nitrate>4mg/L:	
9       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Random       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Random       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Logistic       Predict nitrate       10         11       Singh       2014       Indo- Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11				valley, err	(accp wen)		regression	accuracy 81 5%	
9 Rodriguez- 2014 Granada city, Nitrate 175 Random Predict nitrate 10 Galiano Spain forest >50mg/L; 80.46% 10 Rodriguez- 2014 Granada city, Nitrate 175 Logistic Predict nitrate 10 Galiano Spain regression >50mg/L; accuracy is 73.56% 11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, $R^2 = 11$							regression	sensitivity 51.3%	
9       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Random       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Granada city, Galiano       Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Logistic       Predict nitrate       10         10       Rodriguez- Galiano       Spain       regression       >50mg/L; accuracy is       3.56%         11       Singh       2014       Indo- Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11								specificity 89 7%	
Galiano       Spain       forest       >50mg/L; 80.46%         10       Rodriguez-       2014       Granada city,       Nitrate       175       Logistic       Predict nitrate       10         Galiano       Spain       regression       >50mg/L;       accuracy is       73.56%         11       Singh       2014       Indo-       Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11	9	Rodriguez-	2014	Granada city	Nitrate	175	Random	Predict nitrate	10
10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Logistic regression       Predict nitrate       10         10       Rodriguez- Galiano       2014       Granada city, Spain       Nitrate       175       Logistic regression       Predict nitrate       10         11       Singh       2014       Indo- Currentia       Chemical       409       Decision tree       Test dataset, R <sup>2</sup> =       11	-	Galiano	2011	Spain	1 (Intate	170	forest	>50 mg/L : 80.46%	
GalianoSpainregression>50mg/L; accuracy is $73.56\%$ 11Singh 2014Indo- Chemical409Decision tree built test dataset, $R^2 = 11$	10	Rodriguez-	2014	Granada city	Nitrate	175	Logistic	Predict nitrate	10
11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, $R^2 = 11$	10	Galiano	-01.	Spain	1 111 000	170	regression	>50 mg/L	
11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, $R^2 = 11$		Culturio		Spann			10010001011	accuracy is	
11 Singh 2014 Indo- Chemical 409 Decision tree Test dataset, $R^2 = 11$								73 56%	
	11	Singh	2014	Indo-	Chemical	409	Decision tree	Test dataset. $R^2 =$	11
Gangetic oxygen boost 0.918	-	~B.i		Gangetic	oxvgen		boost	0.918	

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

			plains of	demand				
			north India	(COD)				10
12	Nolan	2015	Central	Nitrate	318	Boosted	Hold-out data,	12
			Valley, CA	(shallow		regression	R <sup>2</sup> =0.26	
				well)		trees		
13	Nolan	2015	Central	Nitrate	318	Artificial	Hold-out data,	12
			Valley, CA	(shallow		neural	$R^2=0.12$	
				well)		networks		
14	Nolan	2015	Central	Nitrate	318	Bayesian	Hold-out data,	12
			Valley, CA	(shallow		networks	$R^2=0.18$	
			···· - ) ; -	well)				
15	Wheeler	2015	Iowa	Nitrate	34 084	Random	Test dataset $R^2 =$	13
10	W neerer	2010	10.00	1 (Itilato	51,001	forest	0.38	
16	Wheeler	2015	Iowa	Nitrate	34 084	Random	Dredict nitrate	13
10	Wheeler	2015	Iowa	Initiate	54,004	forest	$\sum \frac{5ma}{I}$	
						loiest	/ Jilig/L,	
							Accuracy is 0.92,	
							sensitivity is 0.75,	
							specificity is 0.96	0
17	Ayotte	2016	Central	Arsenic	1,180	Boosted	Predict arsenic	8
			Valley, CA			regression	>10 µg/L;	
						trees	Accuracy is 0.91,	
							sensitivity is 0.39,	
							specificity is 0.96	
18	Avotte	2016	Central	Arsenic	1.180	Logistic	Predict arsenic	8
	<b>j</b>		Valley, CA		,	regression	>10 µg/L:	
			, anoj, eri			1081000000	Accuracy is 0.90	
							sensitivity is 0.18	
							specificity is 0.98	
10	Avotto	2017	Contiguous	Arsonia	20.450	Logistic	Dredict arcenic	14
19	Ayone	2017	Contiguous	Alsenic	20,430	Logistic		
			05			regression	>10 µg/L;	
							Accuracy is 0.90,	
							sensitivity is 0.14,	
•			<b>a</b> 1				specificity is 0.99	15
20	Ransom	2017	Central	Nitrate	5,170	Boosted	Hold-out data,	15
			Valley, CA			regression	R <sup>2</sup> =0.434	
						trees		
21	Rosecrans	2017	Central	Dissolved	2,767	Boosted	Hold-out data,	16
			Valley, CA	oxygen		regression	predict DO<0.5	
			-			trees	mg/L, AUC is	
							0.87	
22	Rosecrans	2017	Central	Manganese	2,767	Boosted	Hold-out data.	16
			Valley CA	U	,	regression	predict Mn>50	
			, anoj, eri			trees	$\mu\sigma/L$ AUC is	
							0.87	
23	Tecoriero	2017	Northeastern	Nitrata	10.866	Dandom	Dredict nitrate	17
23	resortero	2017	Wissonsin	Initiate	10,800	forest	Sma/L tost data	
			w isconsin			location	> Sing/L, lest data,	
						classification	accuracy 75%,	
24	т :	2017		т	520	D 1	AJC 0.80	17
24	l'esoriero	2017	Northeastern	Iron	539	Random	Predict iron >0.1	17
			Wisconsin			forest	mg/L, on out of	
						classification	bag training data,	
							accuracy is 74%,	
							AUC is 0.79	
25	Tesoriero	2017	Northeastern	Arsenic	1,275	Random	Predict arsenic >5	17
			Wisconsin			forest	$\mu$ g/L, on out of	
						classification	bag training data	
							<u> </u>	

							accuracy is 74%,	
26	Friekson	2018	North control	Arconio	2 202	Poostad	AUC IS 0.79 Dradiat argania	18
20	ETICKSOII	2018		Alsenic	5,285	rogragion		
			USA			trees	hold out dataset	
						lices	noiu-out uataset,	
							$\frac{1}{2} \frac{1}{2} \frac{1}$	
27	Podgorski	2018	India	Fluorida	12 600	Pandom	ROC IS 0.72 Predict fluoride	19
21	Fougoiski	2018	mula	riuoliue	12,000	format		
						Torest	>1.5111g/L,	
							AUC = 0.84	
20	D . 1	2010	Course la site	NUM	110	Dendem	AUC IS 0.84	20
28	Rodriguez-	2018	Granada city,	Nitrate	110	Kandom	Predict nitrate	20
	Galiano		Spain			forest	>50 mg/L; AUC is	
20	<b>—</b> :	2010	Б	D (* * 1		<b>D</b> 1	0.92	21
29	Irajanov	2018	France	Pesticides	NK	Random	Recalls of 0.84	21
						forest	and 0.86 for the	
							risky and not-	
							risky class	
	~ .						respectively	22
30	Canion	2019	Florida	Nitrate	1554	Random	Predict nitrate	22
						forest	>0.35 mg/L; AUC	
						classification	is 0.89, accuracy	
							is 0.83, sensitivity	
							is 0.79, specificity	
							is 0.86	
31	Canion	2019	Florida	Nitrate	1554	Random	Predict nitrate	22
						forest	>1.2 mg/L; AUC	
						classification	is 0.84, accuracy	
							is 0.79, sensitivity	
							is 0.54, specificity	
							is 0.89	
32	Knoll	2019	Germany	Nitrate	1890	Random	Predict nitrate	6
						forest	concentration,	
						regression	R <sup>2</sup> =0.54	
33	Messier	2019	North	Nitrate	22000	Multiple	Predict nitrate < 1	7
			Carolina			random	mg/L, 1 – 5 mg/L,	
						forest	and $\geq 5$ mg/L,	
						classification	overall accuracy	
							is 0.79	

# 205 **References**

206

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

209 2. Nolan, B. T.; Gronberg, J. M.; Faunt, C. C.; Eberts, S. M.; Belitz, K., Modeling nitrate at
210 domestic and public-supply well depths in the Central Valley, California. *Environmental science*211 & *technology* 2014, 48, (10), 5643-5651.

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

215 <u>https://gdg.sc.egov.usda.gov/</u>, 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), 815829.
- Liaw, A.; Wiener, M. randomForest: Breiman and Cutler's Random Forests for
  Classification and Regression., 4.6-14; 2018.

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.

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.

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.

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,

**241** *531*, 902-911.

242 13. Wheeler, D. C.; Nolan, B. T.; Flory, A. R.; DellaValle, C. T.; Ward, M. H., Modeling
243 groundwater nitrate concentrations in private wells in Iowa. *Science of the Total Environment*244 2015, 536, 481-488.

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

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.

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.

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.

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.

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.

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

267 21. Trajanov, A.; Kuzmanovski, V.; Real, B.; Perreau, J. M.; Džeroski, S.; Debeljak, M.,
268 Modeling the risk of water pollution by pesticides from imbalanced data. *Environmental Science*269 *and Pollution Research* 2018, *25*, 18781-18792.

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