Estimating the high-arsenic domestic-well population in the conterminous United States

By Joseph D. Ayotte, Laura Medalie, Sharon L. Qi, Lorraine C. Backer, and Bernard T. Nolan

Supporting Information

16 pages
3 figures
4 tables

Table SI_1. List of potential independent variables for LR model with data-source citations

Data set	Description	units or scale	Citation or URL
Geologic and geochemical variables			
Bedrock Geology: King and Beikman	Bedrock geology based on King and Beikman 1974 map	Categorical variables	Schruben, P.G., Arndt, R.E., Bawiec, W.J., 1997, Geology of the Conterminous United States at 1:2,500,000 Scale--A Digital Representation of the 1974 P.B. King and H.M. Beikman Map: U.S. Geological Survey Digital Data Series DDS-11 release 2, 26 p.
Bedrock Geology: Reed	Geology for North America (Reed)	Categorical variables	Reed, J.C. Jr., and Bush, C.A., 2005, Generalized geologic map of the Conterminous United States, ed 1.2: U.S. Geological Survey Map.
Bedrock Geology: State Maps	mosaic of state geology maps	Categorical variables	Schweitzer, Peter N. , 2011, Combined geologic map data for the conterminous US derived from the USGS state geologic map compilation.
Stream Sediment Geochemistry	Geochemical data across the U.S. based primarily on stream sediments analyzed using a consistent set of methods.	Number, in mg/kg	U.S. Geological Survey, 2004, The National Geochemical Survey - database and documentation, Edition 1.5: U.S. Geological Survey Open-File Report 2004-1001
Soil Geochmestry	Geochemical data across the U.S. based on multi horizon sediments analyzed using a consistent set of methods.	Number, in mg/kg	Smith, D.B., Cannon, W.F., Woodruff, L.G., Solano, Federico, and Ellefsen, K.J., 2014, Geochemical and mineralogical maps for soils of the conterminous United States: U.S. Geological Survey Open-File Report 2014–1082, 386 p., https://dx.doi.org/10.3133/ofr20141082.
Surficial Geology	20 categories describing the nature and origin of surficial materials, with emphasis on carbonate versus non-carbonate materials.	Categorical variables	Cress, Jill, Soller, David, Sayre, Roger, Comer, Patrick, and Warner, Harumi, 2010, Terrestrial ecosystems—Surficial lithology of the conterminous United States: U.S. Geological Survey Scientific Investigations Map 3126, scale 1:5,000,000, 1 sheet.
Hydrologic variables			
Annual Evapotranspiration (ET11)	Actual ET represents the part of irrigation water that is evaporated and/or transpired and is not available for immediate reuse.	Integer, in mm	Savoca, M.E., Senay, G.B., Maupin, M.A., Kenny, J.F., and Perry, C.A., 2013, Actual evapotranspiration modeling using the operational Simplified Surface Energy Balance (SSEBop) approach: U.S. Geological Survey Scientific Investigations Report 2013-5126, 16 p.
Base Flow Index	The ratio of annual baseflow to the total annual runoff at 1-km grid spacing.	Integer, unit	Wolock, D.M., 2003, Base-flow index grid for the conterminous United States: U.S. Geological Survey Open-File Report 03–263, digital data set
PET	Potential Evapotranspiration	Number, in inches	James Falcone, USGS, written communication, 2015

Precipitation	30-year normal annual precipitation for 1981 through 2010	Number, in inches	http://prism.oregonstate.edu
Precipitation Minus PET	Precipitation minus potential evapotranspiration	Integer, in inches/year	https://www.usgs.gov/media/images/map-gridded-values-1971-2000-avg-precipitation-minus-avg-pet
Recharge to Groundwater	Mean annual natural groundwater recharge created by multiplying a grid of BFI by a grid of mean annual runoff values.	Integer, in mm/year	Wolock, D.M., 2003, Estimated mean annual natural ground-water recharge in the conterminous United States: U.S. Geological Survey Open-File Report 03-311, raster digital data.
Process variables			
Closed Basins	Binary data to indicate whether basin is closed or open. A closed drainage basin allows no outflow to external bodies of water.	0 or 1	Coordinated effort between the USDA-NRCS, USGS, and the EPA. The Watershed Boundary Dataset (WBD) was created from a variety of sources from each state and aggregated into a standard national layer for use in strategic planning and accountability. Watershed Boundary Dataset from https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/water/watersheds/dataset/
Distance to stream	Hydrography	Number, in meters	U.S. Geological Survey, 2016, USGS Small-scale Dataset - Streams and Waterbodies of the United States 200512 Shapefile: U.S. Geological Survey
Evaporites	Location of evaporites in subsurface	Categorical variables	Weary, D.J., and Doctor, D.H., 2014, Karst in the United States: A digital map compilation and database: U.S. Geological Survey Open-File Report 2014–1156, 23 p.
Flow Distance Downstream	Based on NHD plus database, distance of point downstream from watershed boundary	Integer, in meters	Richard Moore, U.S. Geological Survey, 2014, written commun.
Flow Distance Percent	Based on NHD plus database, percent of point downstream compared to stream length	Percent x 100, unit	Richard Moore, U.S. Geological Survey, 2014, written commun.
Flow Distance Upstream	Based on NHD plus database, distance of point upstream from stream outlet	Integer, in meters	Richard Moore, U.S. Geological Survey, 2014, written commun.
Percent Irrigated Land	Estimated percentage of agricultural land subject to a combination of irrigation sources at 1 km grid spacing	Percent x 100, unit	Wieczorek, M., 2005. This data set represents the estimated percentage of the 1-km grid cell that is covered by or subject to the agricultural conservation practice (CPIS05), Combination of Irrigation Sources (CIS) on agricultural land by county (nri_is05): U.S. Geological Survey Raster Digital Data.

State Soil Geographic (STATSGO) Data Base	STATSGO soil characteristics for the conterminous United States	Various numeric variables	https://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml
Topographic Wetness Index (TWI)	A steady state wetness index used to quantify topographic control on hydrological processes: a function of slope and the upstream contributing area per unit width orthogonal to the flow direction.	Integer	Wolock, D.M., 2003, Saturation overland flow estimated by TOPMODEL for the conterminous United States: U.S. Geological Survey Open-File Report 03-264, raster digital data.
		Other features	
Bouguer Gravity	Gravity anomalies produced by density variations within the rocks of the Earth's crust and upper mantle.	Number, in milligal	Kucks, Robert P., 1999, Bouguer gravity anomaly data grid for the conterminous US: U.S. Geological Survey Digital Data Series DDS-9.
Ecoregions	Ecoregions denote areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources.	Categorical variables	US Environmental Protection Agency, 2013, Level III Ecoregions of the Conterminous United States, U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL).
Elevation	Basic elevation data derived from DEMs	Number, in meters	Gesch, D.B., 2007, The National Elevation Dataset, in Maune, D., ed., Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd Edition: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing, p. 99-118.
Groundwater regions	Classification system of the occurrence and availability of groundwater.	Categorical variables	Heath, R.C., 1984, Ground-water regions of the United States: U.S. Geological Survey Water-Supply Paper 2242, 78 p.
Hydrologic Landscape Regions and Variables	U.S. watersheds grouped according to their similarity in landscape and climate characteristics and their associated hydrologic factors.	Categorical variables	Wolock, D.M., 2003, Hydrologic landscape regions of the United States: U.S. Geological Survey Open-File Report 03-145, raster digital data.
Isogravity	Gravitational potential	Number, in milligal	Kucks, Robert P., 1999, Isostatic residual gravity anomaly data grid for the conterminous US
Landcover	Multi-Resolution Land Characteristics Consortium (MRLC) National Land Cover Database (NLCD) at 1 km grid spacing	Categorical variables	Nakagaki, N., Price, C.V., Falcone, J.A., Hitt, K.J., and Ruddy, B.C.,Enhanced National Land Cover Data 1992 (NLCDe 92), http://water.usgs.gov/lookup/getspatial?nlcde92

Percent Tile Drains	County-based data	Percent x 100, unit	http://www.wri.org/publication/assessing-us-farm-drainage
Stream Density	Density of streams	Number, per square mile	U.S. Geological Survey, 2016, USGS Small-scale Dataset - Streams and Waterbodies of the United States 200512 Shapefile: U.S. Geological Survey
Volcano Distance	Distance of well from the nearest volcano	Number, in meters	Smithsonian Institution, Global Volcanism Program, National Atlas of the United States, and the United States Geological Survey, 2004, U.S. National Atlas Volcanoes, ESRI® Data & Maps, vector digital data.

SI_2 Stacked aquifers analysis

Many areas of the United States have horizontally layered aquifers where domestic wells may be drilled in an upper layer, such as unconsolidated sand and gravel of glacial or alluvial origin, but can also be drilled in a deeper layer, such as porous bedrock. Although not completely understood, the complex interrelationship between various geochemical and physical factors that control arsenic concentrations in groundwater (Welch and others, 2000) means that water withdrawn from distinct layered aquifers may be characterized by different arsenic concentrations, as was found in northern Pennsylvania (Low and Galeone, 2007). Because all of the potential independent variables examined in the models for this study were based on 2-dimensional representations in space, in areas of multiple layered aquifers, there was considerable potential for the arsenic signal from a relatively high-As aquifer to get diluted by mixing results with the signal from a relatively low-As aquifer.

We addressed this concern by adding a general aquifer field to the dataset for each well. Information for many of the wells that was retrieved from the National Water Information System (NWIS) was entered by USGS hydrologists and included some designation of either aquifer name or aquifer code. Information from five NWIS fields related to aquifers, at times variously or inconsistently populated, was consolidated into the single general aquifer field. About 14% of the wells in our study did not have any information with which to assign a general aquifer and were given the aquifer designation of 'unknown'.

Frequency distributions of wells with As > 10 ug/L in each state were compared by general aquifer (table S2). For each state, the general aquifer with the largest percentage of domestic wells with As > 10 ug/L was flagged as potentially dominant in areas of stacked aquifers. For example, in Arizona, 63% of the wells in the sand and gravel aquifer had As > 10, whereas 33% of wells in the carbonate rock aquifer, 26% of wells in unknown aquifers, and 10% in the bedrock aquifer had As > 10; thus, the sand and gravel aquifer was potentially dominant. In order to designate a general aquifer as potentially dominant for a state, that aquifer needed a minimum of 25 wells and it needed to be a designated (not unknown) aquifer. A dominant general aquifer was not considered for Ohio, even though the distribution of high As concentrations appeared to be greater in the sandstone aquifer than other aquifers (33% in the sandstone compared to 20% of wells in the sand and gravel aquifer), because the six wells in the sandstone aquifer were too few to generate confidence in the sample.

Map layers of bedrock geology and well locations for states with potentially dominant general aquifers were examined in detail in ArcMap to decide whether to take action by removing wells for the regression analysis. For the states in question, at least two maps of well locations overlain onto geology were scrutinized: one map showing wells from the potentially dominant layered aquifer and one or more maps showing wells in each of the subordinate aquifers. The visual snapshots of contrasting well locations from different types of aquifers relative to the underlying geology were used to justify an action for dealing with the potentially layered aquifer situation. The idea was that removal of some well data from input to the regression models, if certain conditions were met, could be justified in order to improve the predictive power of the regression models. Predictive power could be improved because all of the regression variables are based on a 2-dimensional grid (using x and y map locations); the presence of two different populations of the dependent variable (arsenic concentrations GT1 or GT10) from different aquifer layers, where one population has a higher concentration than the other, could result in dilution of the signal in the regression estimation.

Necessary conditions for omitting well records were that wells in the respective state tapped at least 2 different aquifers, wells from at least 2 different aquifers were interspersed throughout some part of the state (i.e. were not in completely distinct geographic areas because then they wouldn't be stacked), and the arsenic concentrations of wells in the different aquifers appeared to be different. Most states had situations with wells that resulted in "no action" or no removal of well records (table S2, right-most column). For example, figure SI_2_1 shows the ArcMap plots for the 2 potential layered aquifers in Pennsylvania. The sandstone aquifer (fig. SI_2_1a), which has 6 percent of wells with arsenic concentration > 10 ug/L is potentially dominant over the carbonate and sandstone aquifer (fig. SI_2_1b, 3 percent of wells with arsenic concentration > 10 ug/L). However, because wells for these two types of aquifers generally do not overlap, layering is not likely to dilute the regression models, and no action is appropriate. There were 6 types of justification for no action: (1) cases of potentially dominant layered aquifers if the wells in different aquifers were in distinct parts of the state (illustrated with data from Pennsylvania; fig. SI_2_1); (2) there was an insufficient number of wells in one or more of the aquifers to make a difference (seen with data from Connecticut); (3) the distribution of high arsenic concentrations (> 10 ug/L) in the layered aquifers was not different enough to distinguish one aquifer from another (seen with data from Arkansas); (4) the modeled geologic units had similar percentages of wells with > 10 ug/L to eliminate the potential dilution effect of the layered aquifer (seen with data from Kansas); (5) wells in unknown aquifers had higher concentrations of arsenic than wells in named aquifers (seen with data from Washington); or (6) there was only one type of generic aquifer identified for wells in the State (seen with data from South Dakota).

Results from this analysis suggested removal of 208 records for the regression analysis for wells in Idaho, Indiana, Missouri, Nebraska, and Oregon, as described in the "Action and justification" column of table S2. If locations of wells from the potentially dominant aquifer and at least one other aquifer were interspersed within some area of the State, then we assumed the presence of a layered aquifer system in that area. Furthermore, if data from wells in the different layers suggested distinct arsenic populations (that is different percents of arsenic > 10 ug/L), then omission of well records from the subordinate aquifer(s) was justified. For example, using data from Nebraska, wells in the sand and gravel aquifer (fig. SI_2_2a) are interspersed in the x-y plane with wells in other aquifers (fig. SI_2_2b). Additionally, wells in the sand and gravel aquifer showed a potentially different population of arsenic concentrations > 10 ug/L than wells in other aquifers (table S2; 10 percent for sand and gravel versus 0 percent each for sandstone and unknown aquifers). These conditions justified including the 15 Nebraska well records that were not in the sand and gravel aquifer with the set of wells removed from the original dataset.

A comparison of regression results between the full dataset and the test dataset with these 208 wells removed showed that there are small differences in logistic regression model results when some wells are removed from the dataset via the subordinate aquifer analysis. Differences in classification table results for predictions of As > 1, if present, were all less than 0.3 percent. Differences in classification table results for predictions of As > 10 were slightly more substantial; improvements resulting from removing 208 wells were up to 1 percent for sensitivity, false positives and false negatives at a couple of cutpoints. However, the percent correct did not change by more than 0.1 for any cutpoint, most classification table metrics at most cutpoints shows no improvements, and the c value and the H&L statistic were slightly worse in the model with wells removed. Because these improvements to the logistic regression models are negligible, probably because the adjustment only affected 1 percent of the data, the full dataset was used in all subsequent analyses.

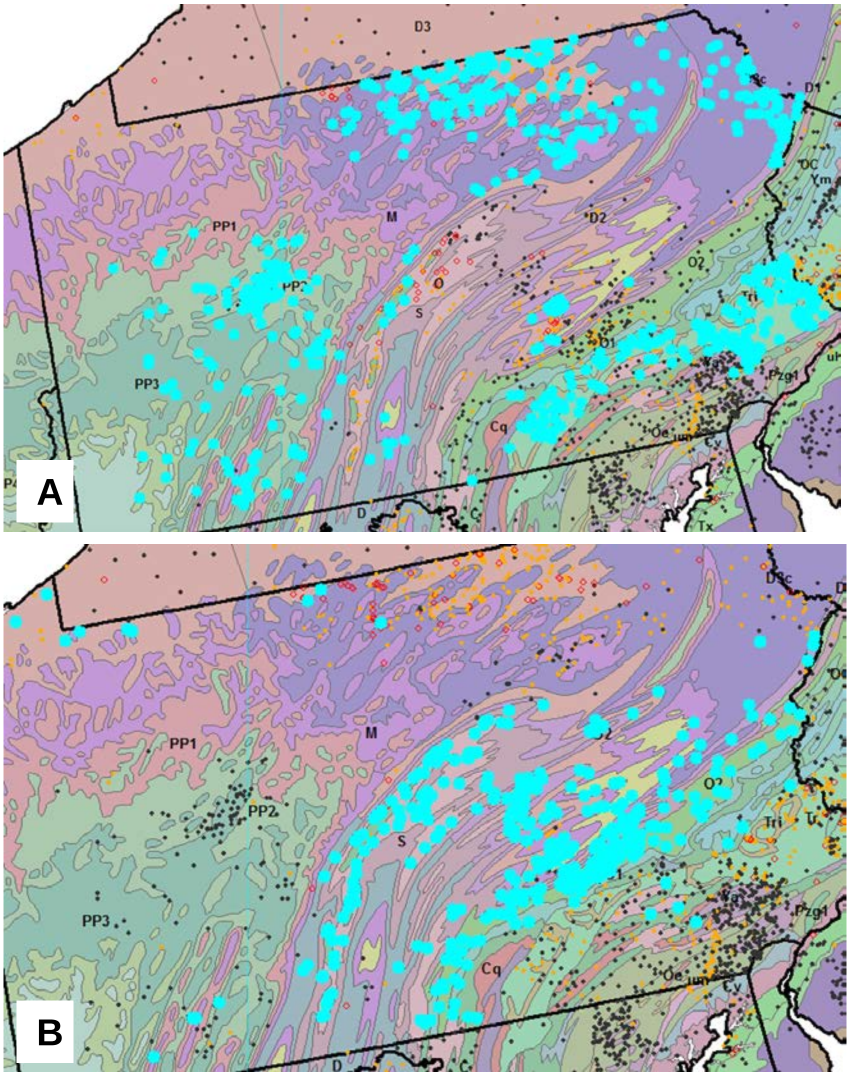


Figure SI_2_1. Maps of Pennsylvania with underlying geology showing wells selected from *A*, the sandstone aquifer and *B*, carbonate and sandstone aquifer.

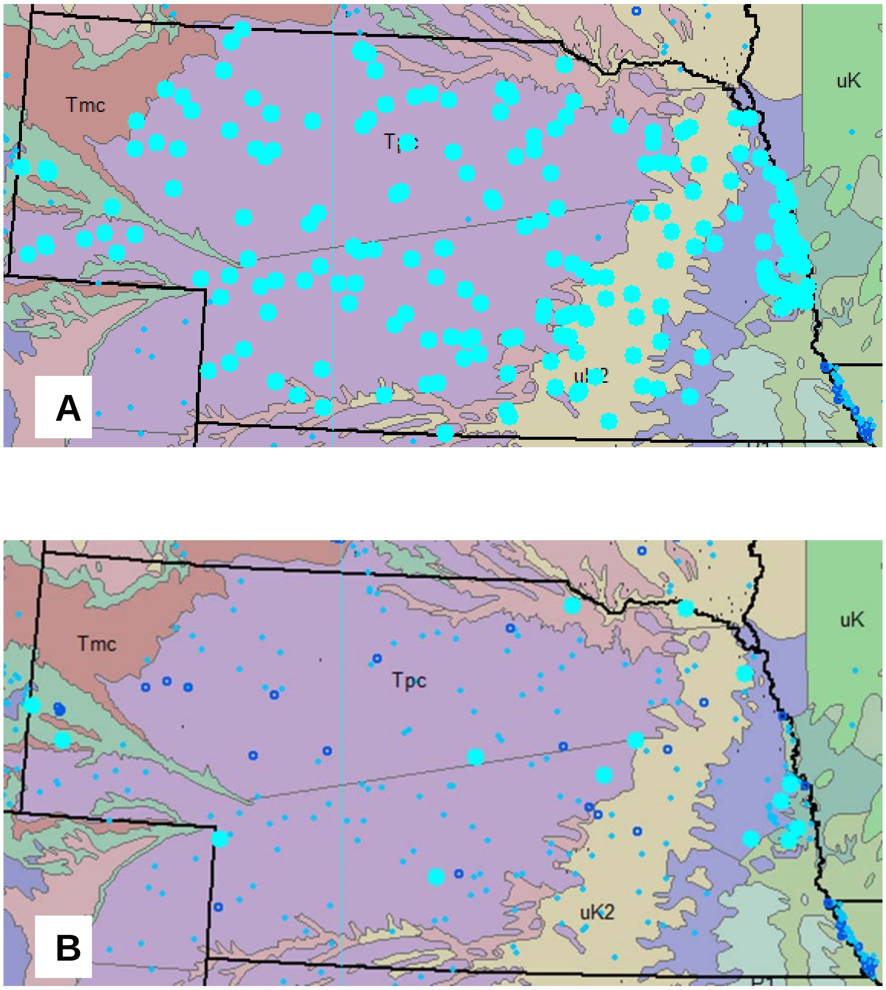


Figure SI_2_2. Maps of Nebraska with underlying geology showing wells selected from *A*, the sand and gravel aquifer and *B*, all other aquifers.

Table SI_2. Number and percent of wells with high arsenic concentrations by type of aquifer, potentially dominant layered aquifer, and action for layered aquifer situations, by State.

[BDRK, bedrock; CARB & SDST, carbonate and sandstone; CARB, carbonate rock; CRYS, crystalline rock; SD & G, sand and gravel; SDST, sandstone; SemiS, semiconsolidated sand; UNK, unknown, VOLC, volcanic rock; Action and justification column explanations– No action: geology, No action because modeled geologic units eliminate potential dilution effect of layered aquifer; No Action: similar distribution, No action because the arsenic distribution in layered aquifers is similar; No Action: small effect, No action because effect of removing wells would be small; No Action necessary, Not a layered aquifer situation; No Action: unknown, No action because wells in unknown aquifers have higher concentrations than wells in named aquifers; No Action: distinct areas, Not a layered aquifer situation because wells are in different parts of the state.]

State	BDRK	CARB & SDST	CARB	CRYS	SD & G	SDST	SemiS	UNK	VOLC	Potentially dominant layered aquifer	Action and justification for area with potentially layered aquifers
	Number of wells and percent having As > 10										
Alaska	3				42			296		SD & G	?? Don't have geologic map of Alaska
	0				29			24			
Alabama		2	2		27	3		5		None	No Action: small effect and similar distribution
		0	0		0	0		0			
Arkansas			41		28	146		1		None	No Action: similar distribution
			0		0	0		0			
Arizona	10		3		245			303		SD & G	No action: geology (Q and Tpc)
	10		33		63			26			
California					1,237		24	162		SD & G	No Action: geology (Q, Kg)
					10		0	6			
Colorado[1]			30		284	261	51	148		SDST, SD & G	No Action: similar distribution
			0		3	4	2	0			
Connecticut				17	10	4		1		SDST	No Action: small effect
				6	0	25		0			
Delaware			12		13					None	No Action: similar distribution
			0		0						
Florida		1	94		43			16		CARB	No Action: distinct areas
		0	3		0			0			
Georgia		1	80		5	2		3		SD & G	No Action: small effect
		0	1		20	0		0			
Iowa					72	2	1			SD & G	No Action: small effect
					3	0	0				
Idaho			7		525		9	1,090	78	SD & G	Remove wells for analysis that are VOLC (geologic unit Qv); this is the only area where wells in different aquifers are interspersed within a geologic unit
			14		17		11	20	3		
Illinois					65			17		SD & G	No Action: similar distribution
					22			24			
Indiana	23		17		147	6		5		SD & G	Remove wells for analysis whose aquifer is not designated as SD & G
	0		0		6	0		20			
Kansas			4		125	80	5	37		SD & G	No Action: geology (Tpc and Q) and distinct areas
			0		2	1	0	11			

Kentucky						30 0		14 0	SDST	No Action: similar distribution
Louisiana					123 1				Single aquifer	No Action necessary
Massachusetts				78 14	8 0	3 33		483 16	CRYS	No Action: unknown
Maryland		18 11	108 0		149 12	42 2		1 0	SD & G, CARB & SDST	No Action: geology (D, Tm, Qp)
Maine								67 96	Single aquifer	No Action necessary
Michigan					40 10	12 0		31 87	SD & G	No Action: distinct areas
Minnesota					8,038 13				Single aquifer	No Action necessary
Missouri			115 1		245 9	97 0		5 0	SD & G	Remove wells for analysis in geologic unit PP3 that are not SD & G; this is the only area where wells in different aquifers are interspersed within a geologic unit
Mississippi					37 5				Single aquifer	No Action necessary
Montana					405 21				Single aquifer	No Action necessary
North Carolina			117 5		1 0			9 20	CARB	No Action: small effect
North Dakota					10 20	13 8		1 0	SD & G	No Action: small effect
Nebraska					209 10	10 0		5 0	SD & G	Remove wells for analysis whose aquifer is not designated as sand and gravel
New Hampshire				455 18					Single aquifer	No Action necessary
New Jersey						133 8			Single aquifer	No Action necessary
New Mexico			1 0		64 14				Single aquifer	No Action necessary
Nevada					68 62			178 61	None	No Action: similar distribution
New York	1 0		46 0	1 0	128 7	149 5	1 0	16 0	SD & G, SDST	No Action: geology (Ym) and equal distribution
Ohio		9 0	132 17		82 20	6 33		10 0	None	No Action: distinct areas
Oklahomoa						508			Single aquifer	No Action necessary

				3					
Oregon[2]			225 8		27 4	60 17	10 0	SD & G	Remove wells for analysis that are in the SemiS or VOLC aquifers because they are interspersed with wells in the SD & G aquifer
Pennsylvania	401 3			656 6				SDST	No Action: distinct areas
South Carolina		48 0	1 0					Single aquifer	No Action necessary
South Dakota			262 4					Single aquifer	No Action necessary
Tennessee	23 0	25 0	19 0						No Action: similar distribution
Texas			1,768 13					Single aquifer	No Action necessary
Utah			96 7	18 0	30 27	40 8		SemiS	No Action: small effect
Washington			80 4			603 9	1 0	UNK	No Action: unknown
Wisconsin		52 2	137 2	62 0				None	No Action: similar distribution
West Virginia	38 5		26 23	108 3		17 0		SD & G	No Action: geology (PP4)
Wyoming	6 0	8 13	95 0	59 2	12 0	14 7		None	No Action: distinct areas

[1]Wells in carbonate rock and of unknown aquifer overlap each other more than other aquifers

[2]Wells in sand and gravel and unknown aquifers are interspersed in Tmv group only

SI_3. Logistic regression model coefficients and coefficient diagnostics for probability of arsenic > 10 µg/L

[All parameters have 1 degree of freedom; GeoChem, stream sediment geochemistry; SurfGeo, surficial geology; KB, Bedrock geology: King and Beikman; HydrLand, hydrologic landscape regions and variables]

Model variable	Dataset (from SI_1)	Description of variable	Model coefficient	Standard error	Wald chi-square[1]	Pr > ChiSq[2]	Exp(Est)[3]	Standardized coefficient
Intercept		--	-0.4071	0.3217	1.6009	0.2058	0.666	--
as_idw_c2	GeoChem	Average arsenic concentrations in the c2 horizon	0.0281	0.00649	18.8291	<.0001	1.029	0.113
be_idw_c	GeoChem	Average beryllium concentrations in the c horizon	0.227	0.0575	15.5837	<.0001	1.255	0.0797
BFI	Base flow index	lower is less base flow	-0.0161	0.00187	73.5277	<.0001	0.984	-0.1643
bi_idw_c	GeoChem	Average bismuth concentrations in the c horizon	-1.7461	0.4267	16.7437	<.0001	0.174	-0.1179
Cv	KB	Cambrian volcanic rocks	1.5296	0.4782	10.2327	0.0014	4.616	0.0495
D	KB	Devonian	2.3586	0.4593	26.3739	<.0001	10.576	0.0608
D3	KB	Upper Devonian	1.0811	0.2155	25.1791	<.0001	2.948	0.0751
De_geo	KB	Devonian, eugeosynclinal	2.4582	0.22	124.8841	<.0001	11.684	0.1331
DSe	KB	Devonian and Silurian, eugeosynclinal	2.2263	0.1563	202.9346	<.0001	9.266	0.1803
HGA	STATSGO	Hydrologic soil group A	-0.0087	0.002	18.9205	<.0001	0.991	-0.008
IVR	SurfGeo	Alkaline intrusive volcanic rock	-2.3869	0.6968	11.7359	0.0006	0.092	-0.0556
lc11	Landcover	Open water	0.7503	0.1659	20.4643	<.0001	2.118	0.0557
lc82	Landcover	Row crops	0.3544	0.0794	19.915	<.0001	1.425	0.0709
M	KB	Mississippian	1.6014	0.4375	13.4009	0.0003	4.96	0.0575
mo_idw_a	GeoChem	Average molybdenum concentrations in the c horizon	-0.1724	0.0376	20.9854	<.0001	0.842	-0.1117
Oe	KB	Ordovician, eugeosynclinal	2.1915	0.2359	86.2648	<.0001	8.948	0.1241
Percent_ti	Percent Tile Drains	--	0.0338	0.00328	106.3286	<.0001	1.034	0.1442
PPT81_10	Precipitation	--	-0.0887	0.00445	396.6356	<.0001	0.915	-0.7063
Pzg1	KB	Lower Paleozoic granitic rocks	2.2046	0.2319	90.4046	<.0001	9.067	0.1221
Pzg2	KB	Middle Paleozoic granitic rocks	2.2257	0.1799	153.0722	<.0001	9.26	0.1557
Pzmi	KB	Phanerozoic mafic intrusives	2.5313	0.3564	50.4421	<.0001	12.57	0.0763
Q	KB	Quaternary	0.5623	0.0927	36.7672	<.0001	1.755	0.1095
Qp	KB	Pleistocene	1.2272	0.1985	38.2307	<.0001	3.412	0.1013
recharge	Recharge to groundwater	--	0.00375	0.000369	103.4729	<.0001	1.004	0.2909

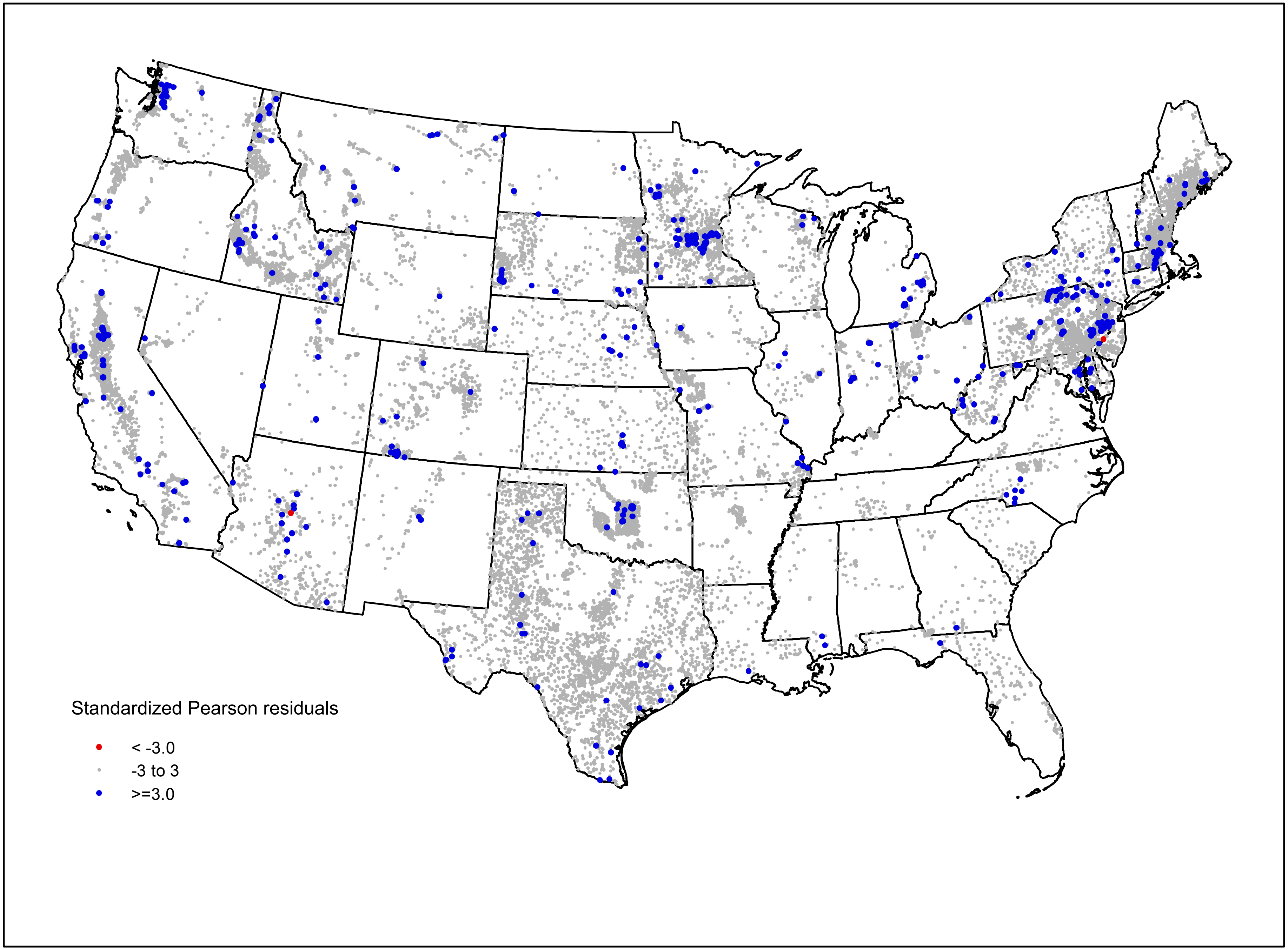
RELIEF	HydrLand	maximum elevation minus minimum elevation in the watershed	-0.00067	0.000085	61.9952	<.0001	0.999	-0.1543
ROCKDEPAVE	STATSGO	Average depth to rock	0.0201	0.0037	29.4895	<.0001	1.02	0.1234
S3	KB	Upper Silurian (Cayugan)	-3.5907	1.0241	12.2928	0.0005	0.028	-0.1264
SAND	HydrLand	Average percent of sand in soil	0.00872	0.00221	15.6126	<.0001	1.009	0.0761
sb_idw_c	GeoChem	Average antimony concentrations in the c horizon	0.3636	0.0391	86.6457	<.0001	1.438	
Se	KB	Silurian, eugeosynclinal	2.5402	0.1526	277.2204	<.0001	12.682	0.191
SLOPEAVE	STATSGO	Average slope	0.0338	0.0045	56.3191	<.0001	1.034	0.01557
SLS	SurfGeo	Saline lake sediment	1.2504	0.1843	46.0366	<.0001	3.492	0.0783
strmden	Stream density	--	1.7426	0.4669	13.9332	0.0002	5.712	0.0576
Tm	KB	Miocene	1.0529	0.2435	18.6984	<.0001	2.866	0.0752
Tmc	KB	Miocene, continental	0.9262	0.1839	25.3703	<.0001	2.525	0.06
Tp	KB	Pliocene	1.6539	0.23	51.7139	<.0001	5.228	0.0765
Tpc	KB	Pliocene, continental	1.0682	0.1006	112.7915	<.0001	2.91	0.1493
Tpv	KB	Pliocene, volcanic	1.8952	0.1973	92.2267	<.0001	6.654	0.0948
Tr	KB	Triassic	0.8601	0.2242	14.7241	0.0001	2.364	0.068
uK3	KB	Taylor Group	-2.2969	0.5831	15.5185	<.0001	0.101	-0.1991
WEG	STATSGO	Average values for wind erodibility group	-0.1314	0.0201	42.8712	<.0001	0.877	-0.1194
WTDEPAVE	STATSGO	Average depth to water	-0.2133	0.0252	71.3999	<.0001	0.808	-0.1697

[1]The test statistic testing the null hypothesis that a predictor's regression coefficient is zero, given the other predictor variables are in the model. It is the squared ratio of the estimate to the standard error of the respective predictor.

[2]The p-value of the Wald chi-square test statistic.

[3]The odds ratio determined by exponentiating the estimate. This is interpreted as: for a one unit change in the predictor variable, the odds ratio for a potitive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

SI_4. Standardized Pearson residuals of the predicted probabilities of arsenic exceeding 10 micrograms per liter in groundwater for the logistic regression model.



SI 5. Influence diagnostics for individual observations in logistic model to predict As > 10 µg/L.

The table below shows some of the influence diagnostic statistics and other information for these potential outliers. Although some observed As concentrations are less than and others are greater than 10, predicted values are all less than 10 µg/L. In general, small predicted values correspond to large observation values and large predicted values correspond to small observation values. That this model outcome is opposite from the expected outcome illustrates why these points might be influential. The three table sections show model results with (A) all values; (B) the 2 most extreme influential cases deleted; and (C) 18 additional influential cases deleted. There is almost no difference in model results, as indicated by information in rows below that begin with 'AIC', between these scenarios.

Results of logistic regression models with 0, 2, and 20 most influential points removed

Case	Unique identifier	Binary variable[1]	Arsenic Concentration (µg/L)		Influence Diagnostic Statistic			State
			Observed	Predicted	Standard Pearson Residual	Leverage	DBETA	
		1. Full model						
		AIC: G_M = 9,603; p = < 0.0001; R^2 = 0.13; R^2 (max re-scaled)= 0.26; H&L = 0.0035; c = 0.807						
14220	430014075141901	1	17.4	0.0002	65.0878	0.0003	-0.03	NY
17475	461738112441701	0	0.2	0.1675	-0.4485	0.1818	0.01	MT
		2. 2 Most Extreme Influential Cases Deleted						
		AIC: G_M = 9,584 p = < 0.0001; R2 = 0.13; R2 (max re-scaled)= 0.26; H&L = 0.0061; c = 0.808						
3704	333408111490301	1	11	0.0148	8.1568	0.0080	-0.01	AZ
3821	334832111385401	1	12.24	0.0173	7.5400	0.0062	0.00	AZ
3953	342048111572501	0	2	0.8903	-2.8495	0.0087	0.02	AZ
9682	395608075042601	0	1	0.8670	-2.5528	0.0388	0.05	NJ
11902	411631098124901	1	10.9	0.0044	15.0170	0.0015	-0.05	NE
16399	444239111055301	1	10.1	0.1035	2.9427	0.0390	0.00	NJ
16441	444533095310301	0	<1	0.2967	-0.6496	0.0027	0.00	MN
16442	444534111104601	1	12.8	0.0451	4.5998	0.0187	0.00	MT
16537	445139096553301	0	<0.5	0.0064	-0.0804	0.0022	0.00	SD
16538	445142096394501	1	13	0.0050	14.0837	0.0017	0.00	SD
16794	451138096391401	0	0.5	0.0605	-0.2539	0.0004	0.00	SD
16795	451138096531301	1	12	0.0050	14.1664	0.0017	-0.01	SD
17445	461509112484701	0	0.8	0.1345	-0.3942	0.1510	0.01	MT
17490	461841114110701	0	<1	0.1710	-0.4542	0.0043	0.00	MT
17492	461848112470601	0	1.2	0.1317	-0.3895	0.1496	0.01	MT
17533	462324116282101	0	<1	0.0550	-0.2412	0.0007	0.00	ID
17535	462345112501101	0	2.2	0.0915	-0.3173	0.1410	0.01	MT
18487	480933120040801	0	0.5	0.7564	-1.7622	0.0766	0.05	WA
		3. 18 Additional Influential Cases Deleted						
		AIC: G_M = 9,499; p = < 0.0001; R2 = 0.13; R2 (max re-scaled)= 0.27; H&L = 0.0136; c = 0.81						

1 The binary variable for logistic regression is 0 or 1 depending on whether the arsenic concentration is less (0) than or greater (1) than 10 µg/L