Supporting Information

Multiple Data Imputation Methods Advances Risk Analysis and Treatability of Co-occurring Inorganic Chemicals in Groundwater

Akhlak U. Mahmood[2§], Minhazul Islam[1§], Alexey V. Gulyuk[2§], Emily Briese[1], Carmen A. Velasco[1], Mohit Malu[3], Naushita Sharma[1], Andreas Spanias[3], Yaroslava G. Yingling[2*], and Paul Westerhoff[1*]

§ A. U. M., M.I. and A. V. G. contributed equally to this paper

[1]School of Sustainable Engineering & the Built Environment, Arizona State University, Tempe, Arizona 85287, United States
[2]Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States
[3]School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, Arizona 85287, United States

*Corresponding email address: p.westerhoff@asu.edu ; Phone: 480-965-2885
*Corresponding email author: yara_yingling@ncsu.edu

Summary of Contents: 28 pages total, 5 tables, 12 figures.

Supporting Text

List of Figures

List of Tables

Supporting Text

1. List of 248 water quality parameters (sorted alphabetically) in raw downloaded dataset

Acid Neutralization Potential As %$CaCO_3$; Acidity ; Acidity, (H+) ; Acidity, mineral methyl orange (as $CaCO_3$) ; Acidity, total, phenolphthalein (as $CaCO_3$) ; Albuminoid nitrogen ; Alkalinity ; Alkalinity, Hydroxide ; Alkalinity, Phenolphthalein (total hydroxide+1/2 carbonate) ; Alkalinity, bicarbonate ; Alkalinity, carbonate ; Alkalinity, total ; Aluminum ; Ammonia ; Ammonia and ammonium ; Ammonia-nitrogen ; Ammonium ; Antimony ; Apparent color ; Argon ; Arsenate ; Arsenic ; Arsenic ion (3+) ; Arsenic ion (5+) ; Arsenite ; Barium ; Barometric pressure ; Beryllium ; Bicarbonate ; Biochemical oxygen demand, non-standard conditions ; Biochemical oxygen demand, standard conditions ; Biologically reactive iron ; Bismuth ; Boron ; Bromide ; Bromine ; Cadmium ; Calcium ; Carbon ; Carbon dioxide ; Carbon monoxide ; Carbonate ; Cations-Anions ; Cerium ; Cesium ; Chemical oxygen demand ; Chemical oxygen demand, (high level) ; Chemical oxygen demand, (low level) ; Chlorate ; Chloride ; Chlorine ; Chromium ; Chromium(III) ; Chromium(VI) ; Cloud cover (percent) ; Cobalt ; Color ; Colored dissolved organic matter (CDOM) ; Conductivity ; Copper ; Corrosion & scaling control, Langelier Saturation Index ; Cyanide ; Density of water at 20 deg C ; Depth ; Depth of pond or reservoir in feet ; Depth to water level below land surface ; Depth, Secchi disk depth ; Depth, from ground surface to well water level ; Detergent, severity (choice list) ; Deuterium/Hydrogen ratio ; Dissolved oxygen (DO) ; Dissolved oxygen saturation ; Dysprosium ; Elevation, groundwater surface, MSL ; Elevation, water surface, MSL ; Erbium ; Europium ; Evaporation ; Extractable fuel hydrocarbons (C13-C22 DRO) ; Ferric ion ; Ferrous ion ; Fish Kill, Severity (choice list) ; Fixed dissolved solids ; Fixed suspended solids ; Floating Garbage Severity (choice List) ; Floating algae mat - severity (choice list) ; Floating debris - severity (choice list) ; Floating sludge - severity (choice list) ; Flow ; Flow

rate, instantaneous ; Fluoride ; Fluorine ; Gadolinium ; Gallium ; Gas bubble severity (choice list) ; General pathology (text) ; Germanium ; Gold ; Hafnium ; Hardness, Ca, Mg ; Hardness, carbonate ; Hardness, non-carbonate ; Height, gage ; Helium ; Holmium ; Hydrogen ; Hydrogen sulfide ; Hydrolyzable phosphorus ; Hydroxide ; Ice cover, floating or solid - severity (choice list) ; Indium ; Inorganic carbon ; Inorganic nitrogen (nitrate and nitrite) ; Iodide ; Iodine ; Iron ; Kjeldahl nitrogen ; Krypton ; Lanthanum ; Lead ; Light attenuation at measurement depth ; Light attenuation, depth at 99% ; Lime (chemical), dolomitic ; Lithium ; Lutetium ; Magnesium ; Manganese ; Mercury ; Moisture content ; Molybdenum ; Neodymium ; Neon ; Nickel ; Niobium ; Nitrate ; Nitrate + Nitrite ; Nitrite ; Nitrogen ; Nitrogen, mixed forms (NH_3), (NH_4), organic, (NO_2) and (NO_3) ; Nitrogenous biochemical oxygen demand ; Nutrient-nitrogen ; Odor threshold number ; Odor, atmospheric ; Oil and Grease ; Organic Nitrogen ; Organic phosphorus ; Orthophosphate ; Osmium ; Osmotic pressure ; Oxidation reduction potential (ORP) ; Oxygen ; Palladium ; Partial pressure of dissolved gases ; Perchlorate ; Phosphate-phosphorus ; Phosphorus ; Platinum ; Potassium ; Praseodymium ; Precipitation ; Pressure ; Purge Volume ; RBP High water mark ; RBP Stream width ; Radium ; Relative humidity ; Reservoir storage ; Rhenium ; Ronnel ; Rubidium ; Ruthenium ; Salinity ; Samarium ; Scandium ; Selenate ; Selenite ; Selenium ; Settleable solids ; Silica ; Silicon ; Silver ; Sodium ; Sodium adsorption ratio [(Na)/(sq root of 1/2 Ca + Mg)] ; Sodium carbonate ; Sodium plus potassium ; Sodium, percent total cations ; Solar irradiation, local ; Soluble Reactive Phosphorus (SRP) ; Specific conductance ; Specific gravity ; Stream flow, instantaneous ; Stream flow, mean. daily ; Stream stage ; Strontium ; Sulfate ; Sulfide ; Sulfite ; Sulfur ; Sulfur hexafluoride ; Sum of anions ; Sum of cations ; Surface area ; Tantalum ; Tellurium ; Temperature, air ; Temperature, air, deg C ; Temperature, air, deg F ; Temperature, water ; Temperature, water, deg F ; Temperature, wet bulb ; Terbium ; Thallium ; Thiocyanate ;

Thulium ; Tide stage ; Tin ; Titanium ; Total Kjeldahl nitrogen ; Total Kjeldahl nitrogen (Organic N & NH_3) ; Total Nitrogen, mixed forms ; Total Phosphorus, mixed forms ; Total carbon ; Total dissolved solids ; Total fixed solids ; Total hardness ; Total solids ; Total suspended solids ; Total volatile solids ; True color ; Tungsten ; Turbidity ; Turbidity Field ; Turbidity severity (choice list) ; UV 254 ; Vanadium ; Volatile dissolved solids ; Volatile suspended solids ; Water level (probe) ; Water level in well, MSL ; Water level in well, depth from a reference point ; Water level reference point elevation ; Wind velocity ; Xenon ; Ytterbium ; Yttrium ; Zinc ; Zirconium ; pH

2. List of elements removed in data cleaning operation

Bismuth ; Fixed dissolved solids ; Temperature, air, deg C ; Fish Kill, Severity (choice list) ; Ferric ion ; Sodium, percent total cations ; Density of water at 20 deg C ; Stream flow, mean. daily ; Arsenic ion (5+) ; Odor threshold number ; Depth of pond or reservoir in feet ; Hafnium ; Pressure ; Relative humidity ; Ytterbium ; Osmotic pressure ; Extractable fuel hydrocarbons (C13-C22 DRO) ; Flow ; Erbium ; Total Nitrogen, mixed forms ; Temperature, air ; Temperature, wet bulb ; Er ; Ronnel ; Depth ; Radium ; RBP High water mark ; Temperature, water, deg F ; Corrosion & scaling control, Langelier Saturation Index ; Barometric pressure ; Tantalum ; Rhenium ; RBP Stream width ; Temperature, air, deg F ; Cloud cover (percent) ; Apparent color ; Sum of anions ; Germanium ; True color ; Ferrous ion ; Dysprosium ; Turbidity Field ; Alkalinity, bicarbonate ; Fixed suspended solids ; Indium ; Sulfite ; Water level in well, MSL ; Flow rate, instantaneous ; Biologically reactive iron ; Water level in well, depth from a reference point ; Selenite ; Depth, from ground surface to well water level ; Thiocyanate ; Lutetium ; Temperature, water ; Water level reference point elevation ; Stream stage ; Depth_Unit ; Light attenuation at measurement depth ; Palladium ; Settleable solids ; Evaporation ; Deuterium/Hydrogen ratio ; Depth, Secchi

disk depth ; Cesium ; Light attenuation, depth at 99% ; Water level (probe) ; General pathology (text) ; Color ; Purge volume ; Colored dissolved organic matter (CDOM) ; Chlorate ; Tellurium ; Total carbon ; Europium ; Neodymium ; Xenon ; Precipitation ; Praseodymium ; Height, gage ; Elevation, water surface, MSL ; Gadolinium ; Bromine ; Solar irradiation, local ; Reservoir storage ; Ruthenium ; Stream flow, instantaneous ; Selenate ; Niobium ; Floating Garbage Severity (choice List) ; Moisture content ; Acidity, total, phenolphthalein (as $CaCO_3$) ; Salinity ; Helium ; Holmium ; Gold ; Depth to water level below land surface ; Wind velocity ; Lime (chemical), dolomitic ; Arsenic ion (3+) ; Alkalinity, Phenolphthalein (total hydroxide+1/2 carbonate) ; Sodium plus potassium ; Thulium ; Yttrium ; Samarium ; Terbium ; Platinum ; Acidity ; Elevation, groundwater surface, MSL ; Krypton ; Surface area

3. Selected groundwater parameters (sorted alphabetically) from North Carolina

Acidity, (H+) ; Alkalinity ; Aluminum ; Ammonia and ammonium ; Antimony ; Arsenic ; Barium ; Beryllium ; Bicarbonate ; Biochemical oxygen demand, standard conditions ; Boron ; Bromide ; Cadmium ; Calcium ; Carbonate ; Chemical oxygen demand, (high level) ; Chloride ; Chromium ; Cobalt ; Copper ; Fluoride ; Hardness, Ca, Mg ; Hardness, non-carbonate ; Iron ; Kjeldahl nitrogen ; Lead ; Lithium ; Magnesium ; Manganese ; Mercury ; Molybdenum ; Nickel ; Nitrate ; Nitrite ; Nitrogen, mixed forms (NH_3), (NH_4), organic, (NO_2) and (NO_3) ; Organic Nitrogen ; Orthophosphate ; Oxygen ; Phosphate-phosphorus ; Phosphorus ; Potassium ; Selenium ; Silica ; Silver ; Sodium ; Specific conductance ; Strontium ; Sulfate ; Thallium ; Total dissolved solids ; Total solids ; Vanadium ; Zinc ; pH

4. Selected groundwater parameters (sorted alphabetically) from Arizona

Acidity, (H+) ; Alkalinity ; Alkalinity, Hydroxide ; Alkalinity, Phenolphthalein (total hydroxide+1/2 carbonate) ; Alkalinity, bicarbonate ; Alkalinity, carbonate ; Alkalinity, total ; Aluminum ; Ammonia and ammonium ; Ammonia-nitrogen ; Antimony ; Arsenic ; Barium ; Beryllium ; Bicarbonate ; Biochemical oxygen demand, standard conditions ; Boron ; Bromide ; Cadmium ; Calcium ; Carbonate ; Chemical oxygen demand ; Chemical oxygen demand, (high level) ; Chloride ; Chromium ; Chromium(III) ; Chromium(VI) ; Cobalt ; Copper ; Dissolved oxygen (DO) ; Fluoride ; Hardness, Ca, Mg ; Hardness, carbonate ; Hardness, non-carbonate ; Hydroxide ; Inorganic carbon ; Iron ; Kjeldahl nitrogen ; Lead ; Lithium ; Magnesium ; Manganese ; Mercury ; Molybdenum ; Nickel ; Nitrate ; Nitrite ; Nitrogen, mixed forms (NH_3), (NH_4), organic, (NO_2) and (NO_3) ; Organic Nitrogen ; Orthophosphate ; Oxidation reduction potential (ORP) ; Oxygen ; Perchlorate ; Phosphorus ; Potassium ; Selenium ; Silica ; Silver ; Sodium ; Specific conductance ; Strontium ; Sulfate ; Sulfide ; Thallium ; Total dissolved solids ; Total fixed solids ; Total hardness ; Total suspended solids ; Turbidity ; Vanadium ; Zinc ; pH

5. Data limitations

- Multiple data sources
    - Private wells
    - Unknown use: domestic, industrial, irrigation
- Data generation
    - States
    - Tribes
    - Local agencies
    - Research grants

- Analytical methods and detection limits
    - Change with time and technology.
- Inconsistent units

6. Data Curation

The following steps were implemented in the data curation portion of the study:

a. Over 20 million data points of groundwater quality parameters were downloaded from the Water Quality Portal database.
b. First, parameters irrelevant to groundwater quality or having very low percent of data availability were removed from the dataset, including air temperature, depth, and stream velocity. We selected the data points that were groundwater samples and collected from water media.
c. All (~500) categorical values were removed from the dataset. These categorical values were recorded for calcium concentration, detergent severity, floating algae mat severity, gas bubble severity, odor and grease, turbidity severity etc.
d. Then the dataset was classified into several states; NC and AZ were selected for the study.
e. After completing the above fixation in the dataset, we combined the concentration files with the sampling location files using the "MonitoringLocationIdentifier" field. This field connects the sampling location file and concentration data file by acting as a database connector.
f. The water quality parameters existing in the NC and AZ datasets were sub-divided into six groups: Physical, Major Non-Metal, Major Metals, Minor Metals, Nutrients, Minor Non-Metals.

g. Also, six measurement types of water quality parameters were found in AZ: Not Detected, Detected Not Quantified, Systematic Contamination, Present Below Quantification Limit, Present Above Quantification Limit, Not Reported. In the NC dataset, only Not Detected, Detected Not Quantified, and Present Above Quantification Limit were found. The Water Quality Portal defines "Not Detected" as data was looked for but was not observed/detected within defined laboratory reporting limits or method detection limits, and "Present Below Quantification Limit" as data was found less than defined laboratory reporting limits or method detection limits. We assigned half of the detection limit to data points indicated as both "Not Detected" and "Present Below Quantification Limit" in both the NC and AZ datasets. This approach allowed us to maximize the information retrieved from the sparse groundwater data matrix of the Water Quality Portal.

Detection limit status of water quality parameters in North Carolina

Detection Limit Status	**#Data points**
Not Detected	28,238
Detected Not Quantified	1,691
Present Above Quantification Limit	37

Detection limit status of water quality parameters in Arizona

Detection Limit Status	**#Data points**
Not Detected	87,117
Detected Not Quantified	4,066
Systematic Contamination	452
Present Below Quantification Limit	267

Present Above Quantification Limit	28
Not Reported	27

h. Units for each water parameter were fixed using a unit conversion dictionary. The goal was to make all the units consistent for each component. For this step, a unit conversion dictionary was prepared for all the components.

i. Datasets included some outliers that had no physical significance for the analysis. In this process, some negative concentration values were removed from the dataset, as were some pH values that exceeded the range of 0-14.

j. Each row in this datafile represented a fingerprint sample. To identify these fingerprint samples, we combined "ActivityStartDate" and "MonitoringLocationID" and created another field in the file called "SamplelD". This process identified rows in the datafile as fingerprints by converting the groundwater component names from the column values to column names.

k. To exclude old datasets with questionable analytical methods or reporting limits in the training set of our imputation model, we excluded data points older than the year 1974 in accordance with the Clean Water Act (CWA) effective year 1972.

l. In the data cleaning process, we also removed values beyond the range of 5* standard deviation to avoid biases in the machine learning model results. Analysis is shown below in **Figure S1** and **Figure S2**.

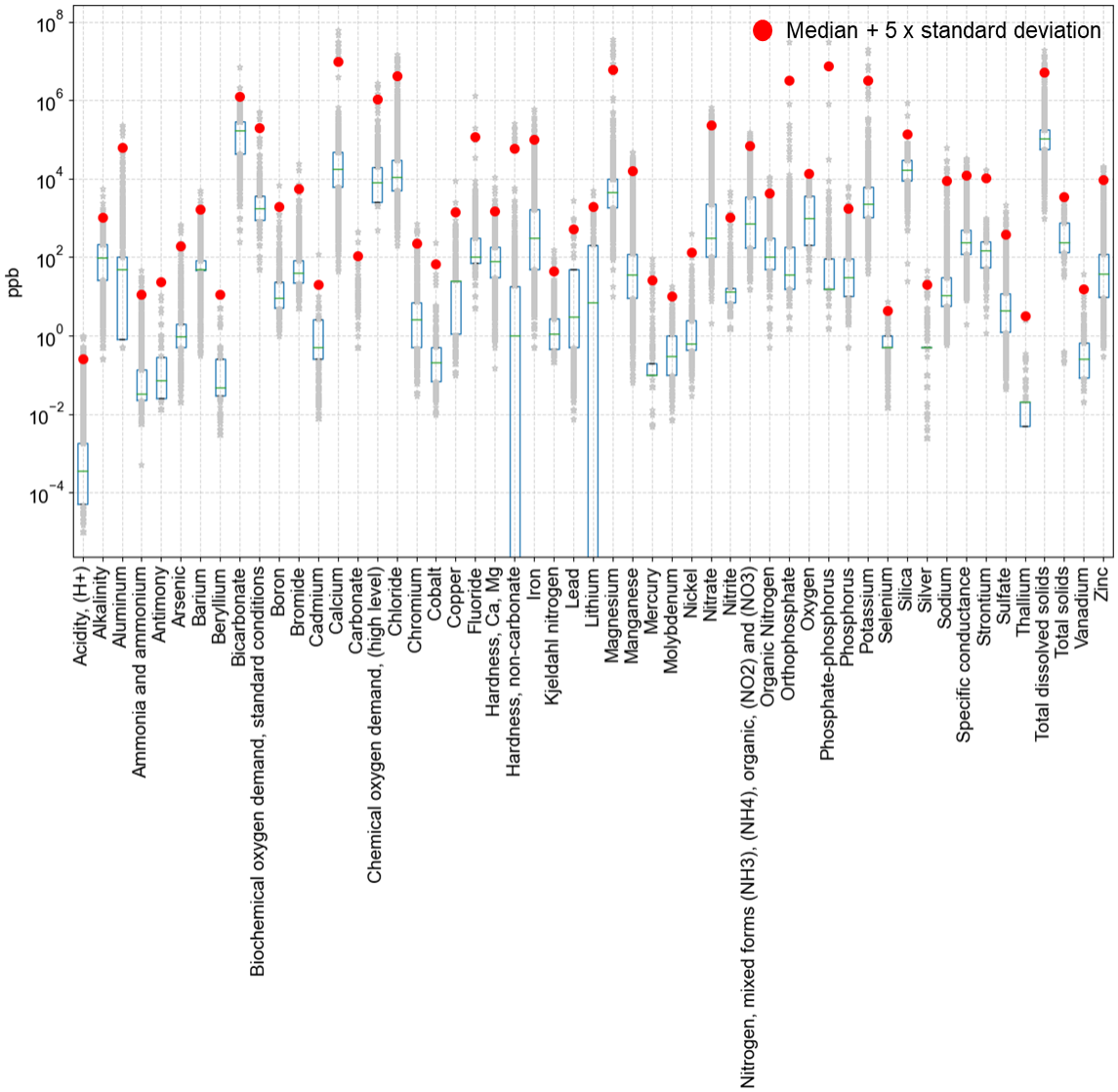


Figure S1: Outliers beyond 5 standard deviation values were filtered from the NC dataset.

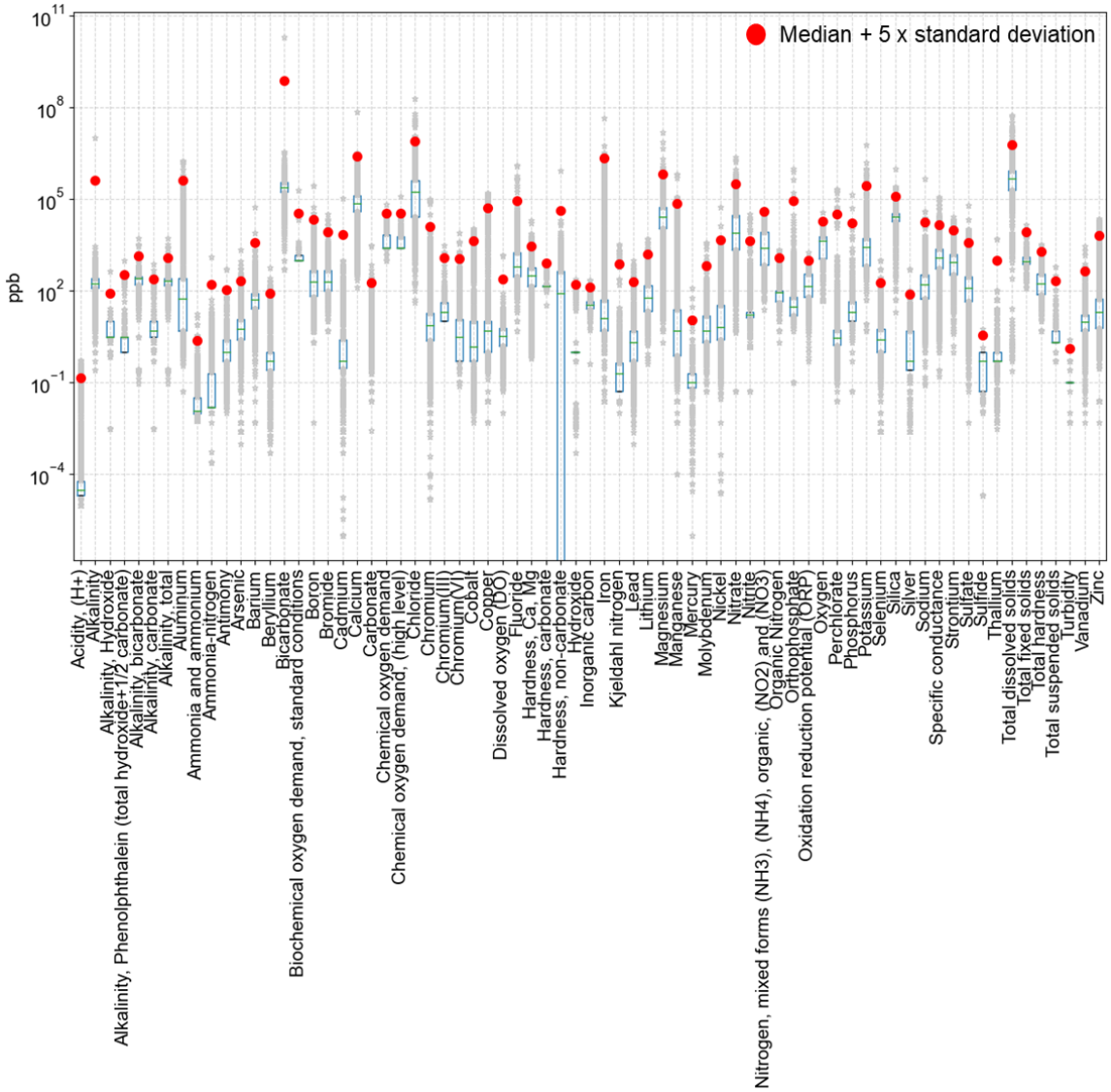


Figure S2: Outliers beyond 5 standard deviation values were filtered from the AZ dataset.

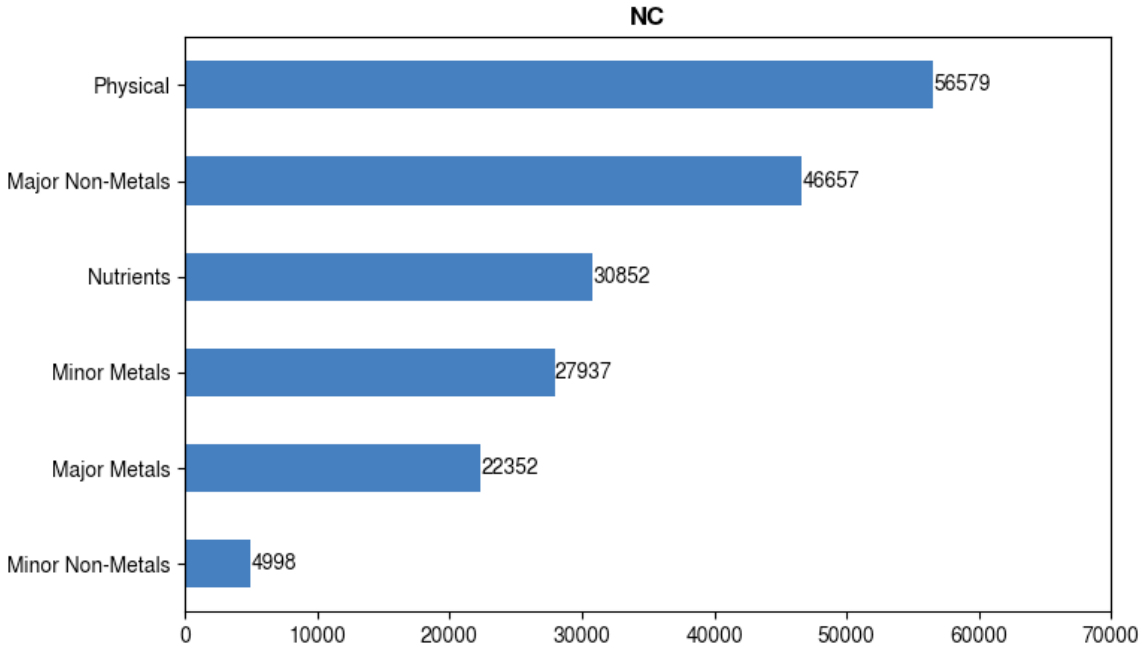


Figure S3: Number of data entries (x-axis) for characteristic groups in NC dataset.

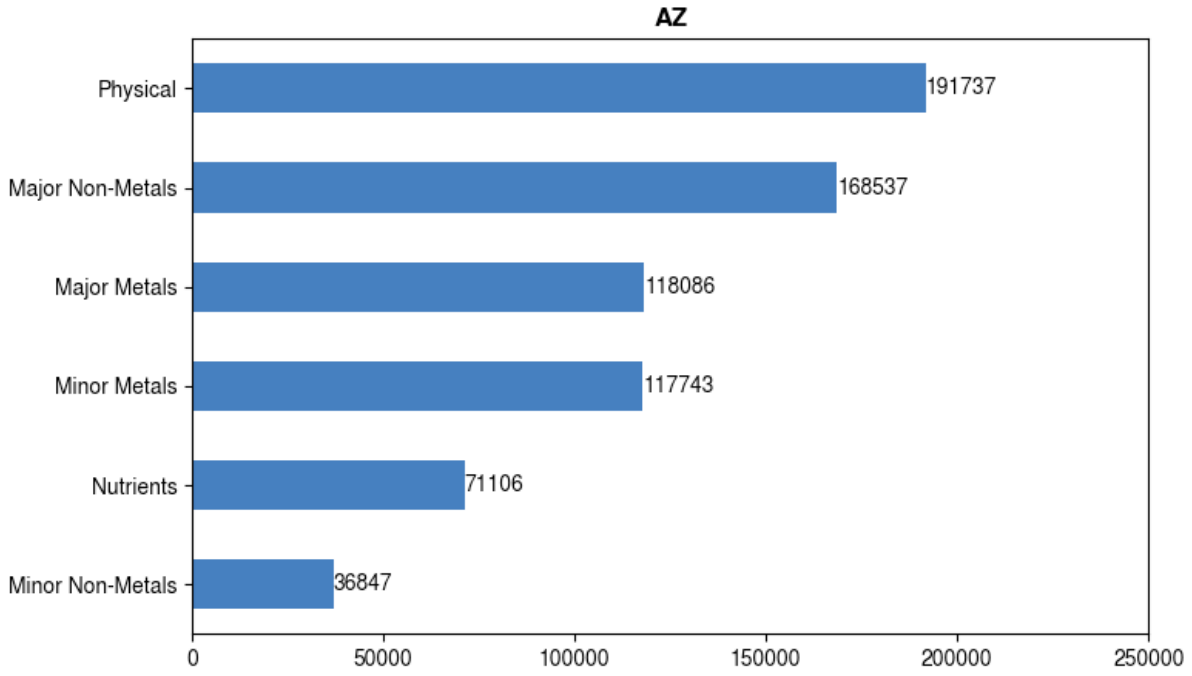


Figure S4: Number of data entries (x-axis) for characteristic groups in AZ dataset.

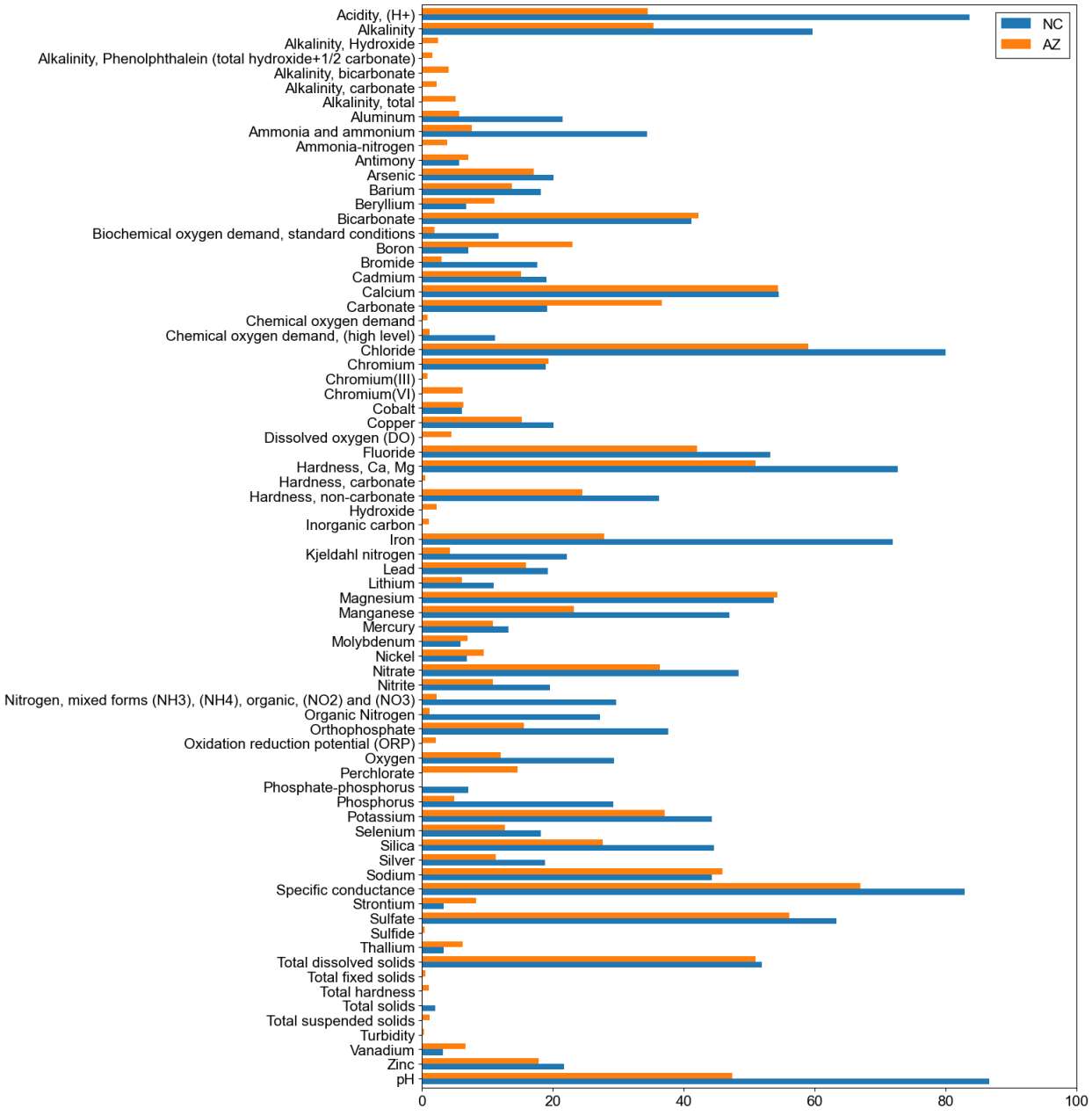


Figure S5: Percent occurrence of groundwater parameters for databases from North Carolina and Arizona.

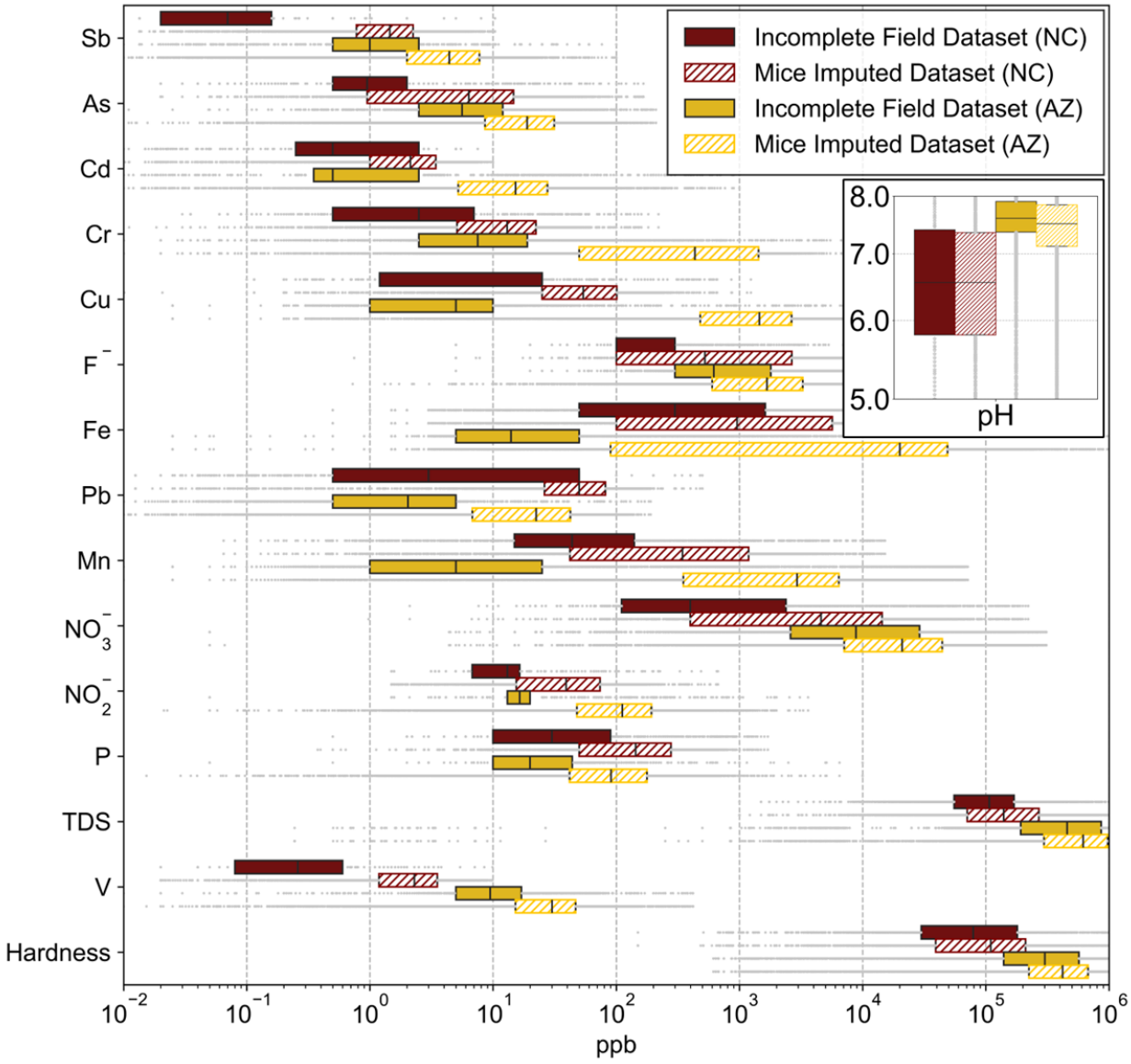


Figure S6: Distribution and variability between incomplete field data and MICE imputed dataset. Solid bars represent measurements of field collected samples, while hashed-filled bars represent imputed datasets. The bar and whisker plot show median values with a vertical line within the bar, ends of the bar represent 25th and 75th percentiles, and grey datapoints are outside those percentiles. Companion plot for AMELIA is provided as Figure 2.

Table S1: Total number of groundwater samples analyzed for each chemical of health concern in Arizona and North Carolina.

Water Quality Parameters	AZ		NC	
	Field	**Imputed**	**Field**	**Imputed**
Antimony (Sb)	2,918	26,784	339	3,523
Arsenic (As)	6,922	26,784	1,194	3,523
Cadmium (Cd)	6,185	26,784	1,130	3,523
Copper (Cu)	6,144	26,784	1,188	3,523
Chromium (Cr)	7,848	26,784	1,125	3,523
Lead (Pb)	6,469	26,784	1,138	3,523
Manganese (Mn)	9,004	26,784	1,949	3,523
Vanadium (V)	2,730	26,784	193	3,523
Fluoride (F^-)	12,017	26,784	1,822	3,523
Nitrate (NO_3^-)	6,576	26,784	1,793	3,523
Nitrite (NO_2^-)	4,311	26,784	1,167	3,523
Hardness	8,300	26,784	2,014	3,523
Phosphorus (P)	1,955	26,784	1,732	3,523
Total Dissolved Solids (TDS)	9,426	26,784	1,969	3,523
Iron (Fe)	9,955	26,784	2,171	3,523
Total =	**91,765**	**401,760**	**20,924**	**52,845**

Table S2: Imputation ± error of median values to evaluate performance. Negative errors indicate imputation is underpredicting and positive errors indicate overprediction.

GW Elements	AMELIA Imputation		MICE Imputation	
	AZ	**NC**	**AZ**	**NC**
Antimony	0.120	1.919	3.42	1.38
Arsenic	- 0.238	0.0036	13.3	5.40
Cadmium	0.0041	0.0132	14.8	1.63
Chromium	- 2.51	2.28	428	10.5
Copper	0.0337	- 0.345	144	29.
Fluoride	231	6.12	1054	425
Iron	5.43	0	20093	656
Lead	1.30	42.0	20.3	47
Manganese	0	- 6.84	2940.	300
Nitrate	- 2398	- 1.43	12142	4206
Nitrite	1.02	34.8	95.3	26.1
Phosphorus	- 4.99	1.02	70.8	113.7
TDS	96936	10333	160832	33166
Vanadium	1.02	1.05	20.6	2.034
pH	0.0242	0.100	- 0.0966	0
± Error	**6318**	**694**	**1327**	**259**
Absolute minimum error	**0**	**0**	**0.0966**	**0**
Absolute maximum error	**96936**	**10333**	**160832**	**33166**

7. Correlation matrix analysis to identify highly correlated features in the dataset

Figure S7 and **Figure S8** show 1:1 correlation analysis between chemical parameters. Some variables had stronger correlations in one state than the other. For instance, antimony (Sb) and nitrite (NO_2^-) had a strong positive correlation (r=0.74) in NC but not in AZ (r=0.47). In AZ, arsenic and vanadium were positively correlated (r=0.74) while arsenic and vanadium had a strong positive correlation in AZ. Other variables had weaker correlations in both locations, indicating that they were less influenced by local factors. Fluoride, for example, had very low correlations with all other variables in both NC and AZ. Additionally, some variables had negative correlations in both locations, indicating that they varied in opposite directions. Nitrate and pH, for instance, had a weak correlation in AZ and a weak negative correlation in NC. Differences in the correlations may reflect the differences in the water sources, environmental conditions, and human activities in the two locations. While a few chemicals showed modest correlations (r>0.5), the lack of exact 1:1 chemical correlation support the need for the fingerprinting machine learning based upon implicit to both AMELIA and MICE.

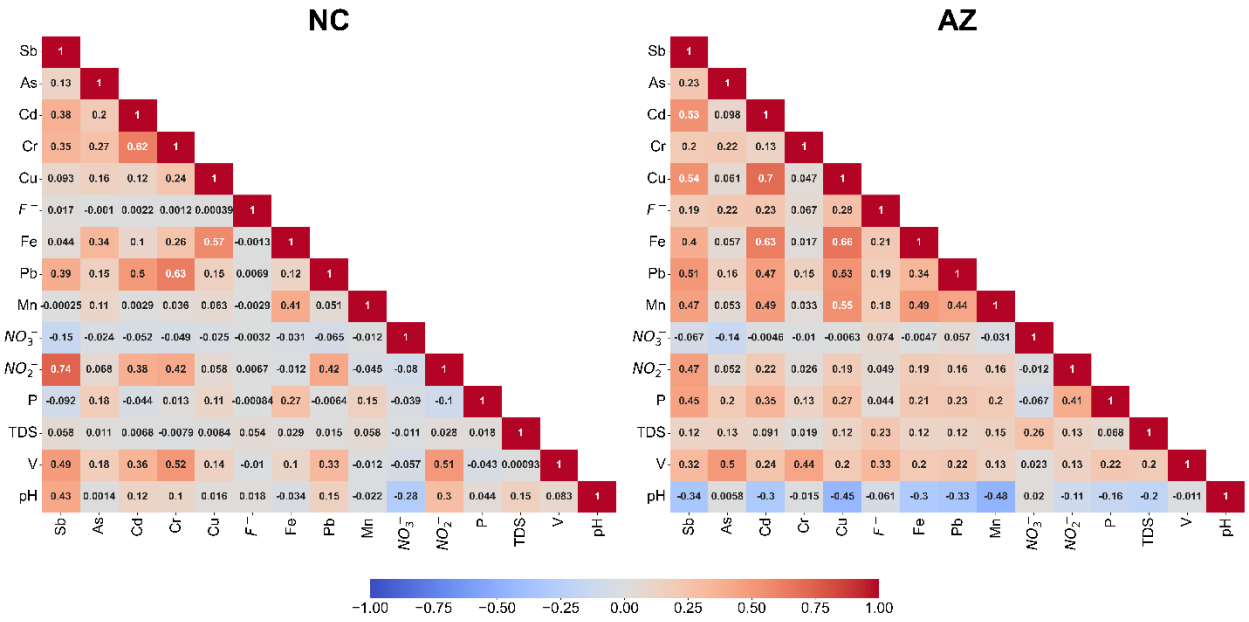


Figure S7: Co-occurrence among the AMELIA imputation predicted groundwater metals.

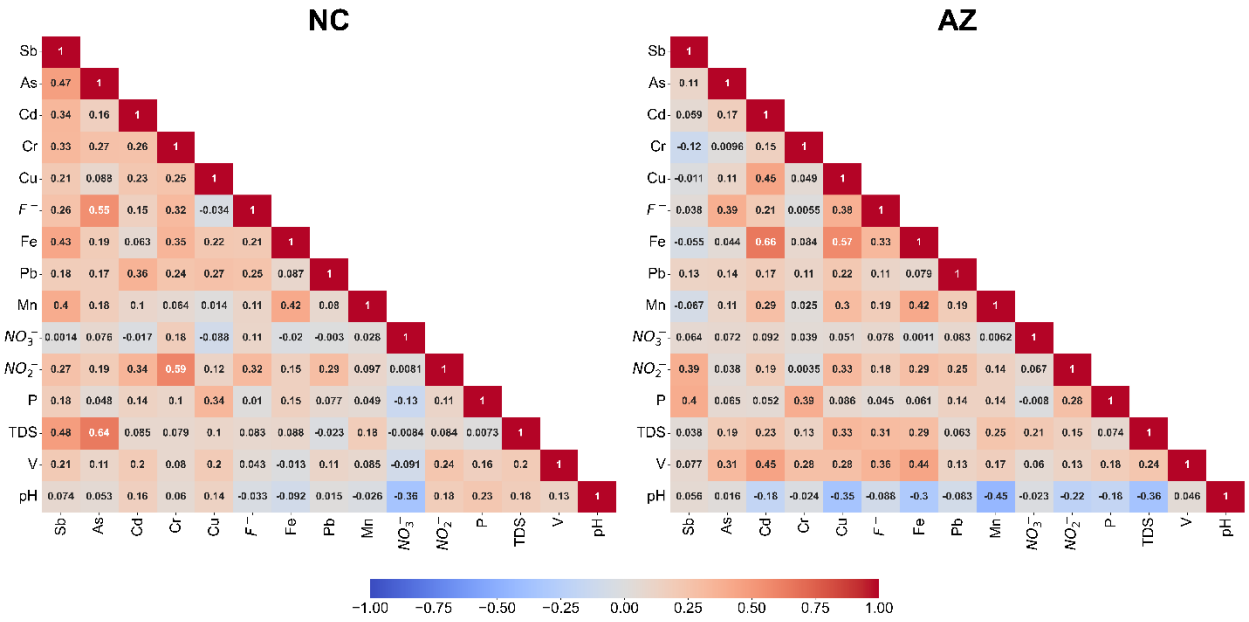


Figure S8: Co-occurrence among the MICE imputation predicted groundwater metals.

Table S3: Water treatability statistics associated with Figure 5 and Table 2. Values under each column show the number of groundwater samples in each water treatment method. Only field and AMELIA imputed data are shown below.

Co-occurrence scenario	Data Source	I	II	III	IV	TOTAL # samples
All Data	Field AZ	3,991	1,280	68	66	5,405
	Imputed AZ	15,286	10,281	929	288	26,784
	Field NC	1,007	34	122	4	1,167
	Imputed NC	3,339	35	145	4	3,523
Si > 20 ppm	Imputed AZ	13,676	9,883	875	279	24,713
	Imputed NC	1,154	25	108	4	1,291
P > As	Imputed AZ	15,018	8,677	904	217	24,816
	Imputed NC	3,331	30	131	3	3,495
V > As	Imputed AZ	13,890	6,311	593	105	20,899
	Imputed NC	1,921	15	76	0	2,012
P > As, V > As	Imputed AZ	13,707	5,677	587	99	20,070
	Imputed NC	1,918	15	76	0	2,009

Table S4: Summary of hardness from field and **AMELIA imputed data. *Units of hardness reported as mg$CaCO_3$/L

Location	Data Source(s)	Soft	Slightly Hard	Moderately Hard	Hard	Very Hard	Total # samples
		<20	**20-60**	**60-120**	**120-180**	**>180**	
Arizona	**Field samples**	569	519	1,137	1,309	4,766	8,300
	Imputed data	591	538	1,375	2,019	22,261	26,784
North Carolina	**Field samples**	401	595	403	231	384	2,014
	Imputed data	457	1,110	1,017	418	521	3,523

Table S5: Summary of total dissolved solids (TDS) from the field and **AMELIA imputed data.

State	Data Source	# values associated with different customer perceptions of water quality and TDS ranges (mg/L)					Total # samples
		Excellent	Good	Fair	Poor	Aesthetically Unpleasant	
		< 150 mg/L	150 to 250 mg/L	250 to 500 mg/L	500 to 1000 mg/L	> 1000 mg/L	
Arizona	Field samples	2,118	1,910	1,917	1,993	1,488	9,426
	Imputed data	2,523	2,788	6,066	12,224	3,183	26,784
North Carolina	Field samples	1,463	288	126	54	38	1,969
	Imputed data	2,703	521	179	63	57	3,523

8. Amelia imputation diagnostic test for Sb and V

The performance of data imputation methods can vary based on the structure and characteristics of the initial dataset. To address concerns about over-imputation and bias, we tested two algorithms, MICE and AMELIA, and evaluated them separately. Each method has its strengths and limitations, and their accuracy and reliability can differ depending on the dataset's nature and structure. To further evaluate some of the highly missing groundwater quality parameters (e.g., Sb, V) in the original dataset, we considered using a relatively small portion of data, which has the most complete data for the water quality parameters that were missing at a comparatively high rate (i.e., Sb, V). We ran cross-validation for the imputation methods. For this validation, we tested the Amelia method, which worked better among the two models.

Figure S9 presents representative diagnostic visualizations generated using the Amelia imputation method. The entire dataset was intentionally masked to simulate a scenario of high

completeness, particularly regarding the availability of Sb data for both Arizona (AZ) and North Carolina (NC). The upper panels of Figure S9 compare density plots between the original and imputed datasets. It was observed that the Amelia-imputed Sb data for both AZ and NC displayed a slight tendency to overestimate values. Moreover, the imputed data demonstrates a normal distribution, while the field monitoring data reveals a complex bimodal distribution (**Figure S9**). This suggests that the imputation model tends to simplify the inherent complexity of the field data. In this context, the AMELIA imputation method utilizes a multivariate normal distribution algorithm to fill in the missing values, aligning with the nature of the distribution. The bottom panel of Figure S9 illustrates the results of a diagnostic evaluation of overimputation, where certain data points from the observed dataset were intentionally omitted to test the performance of the Amelia algorithm as the amount of missing data increases (indicated by colors ranging from green to red, with red indicating higher rates of missing data). The plot also assesses the accuracy of imputing the target variable, along with a 90% confidence interval. Data points that align closely with the diagonal line indicate more precise predictions. Our findings demonstrate that with fewer missing data points, the model was better at predicting lower Sb values for AZ than higher values. However, as the amount of missing data increased, the model's predictions for AZ became less confident. For NC, the model was more confident in its predictions but less accurate overall. In summary, the Amelia model performed better at predicting Sb data for AZ compared to NC, likely due to the higher amount of missing data and the lower volume of available Sb data for NC.

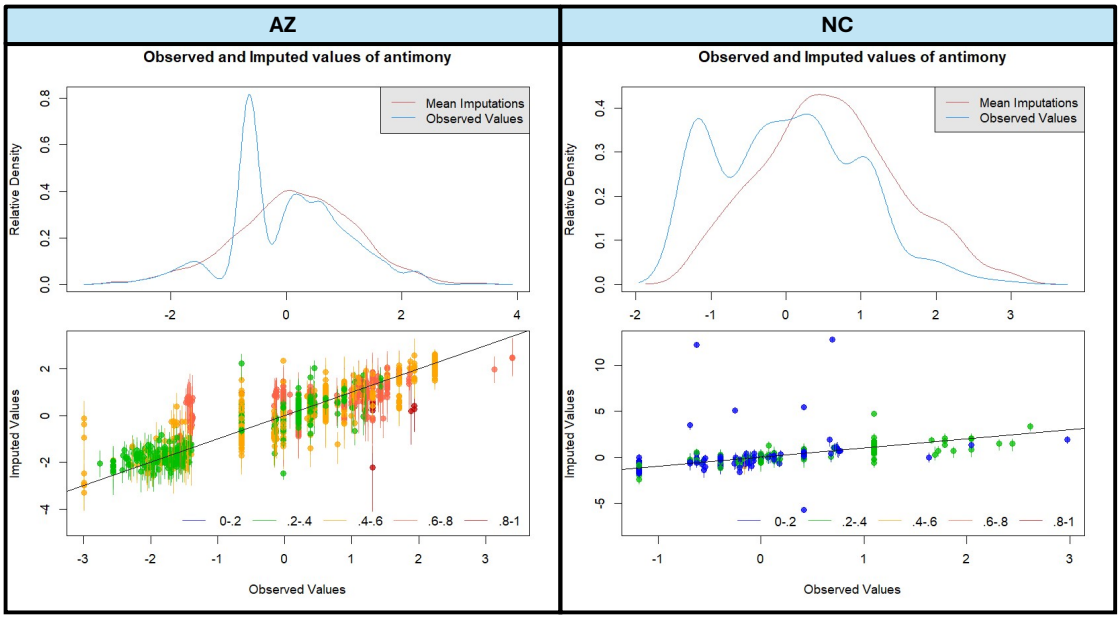


Figure S9: Amelia imputation diagnostic plots for Antimony (Sb) data.

Figure S10 shows a comparable analysis for vanadium as Figure S9 demonstrated for antimony. Results in Figure S10 indicate that the averages of the imputed data for both AZ and NC were somewhat higher than what was observed in the field. Further analysis, specifically diagnostics on overimputation, revealed that the Amelia model showed better performance with imputing the vanadium concentration in AZ when compared to those in NC.

Based upon findings discussed relative to Figures S9 and S10, ongoing research involves collaborating with State agencies to collect more field samples in regions where our models predicted elevated pollutant concentrations, but where very limited field data currently exists. This new data will be integrated into further model refinements to reduce uncertainty of predictions.

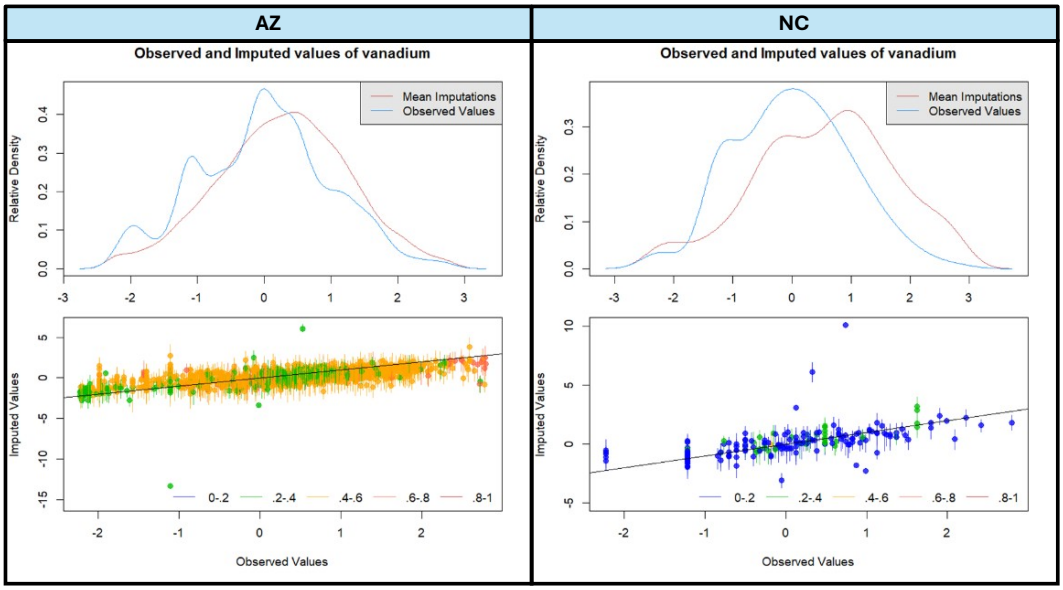
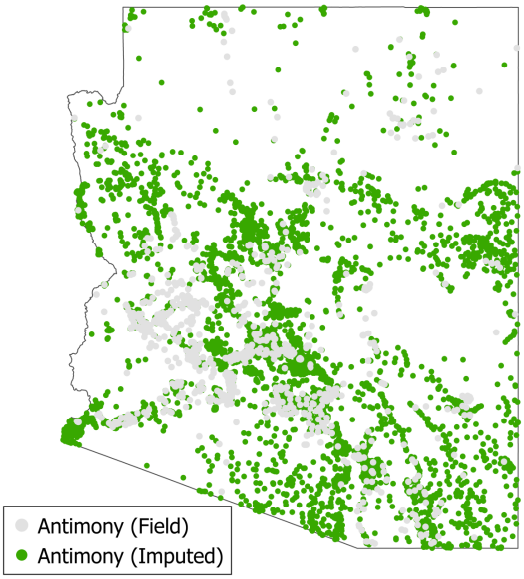
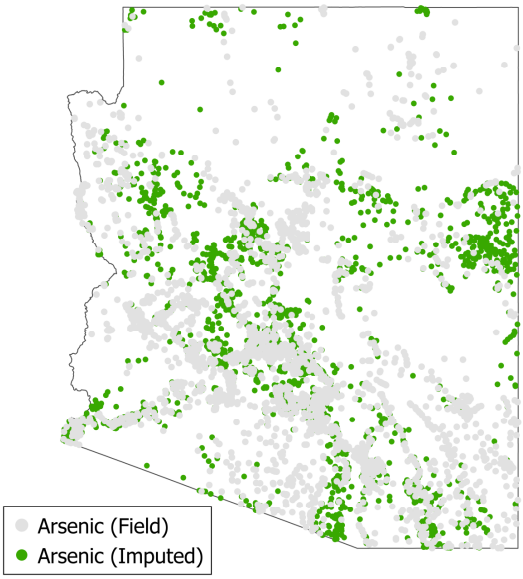
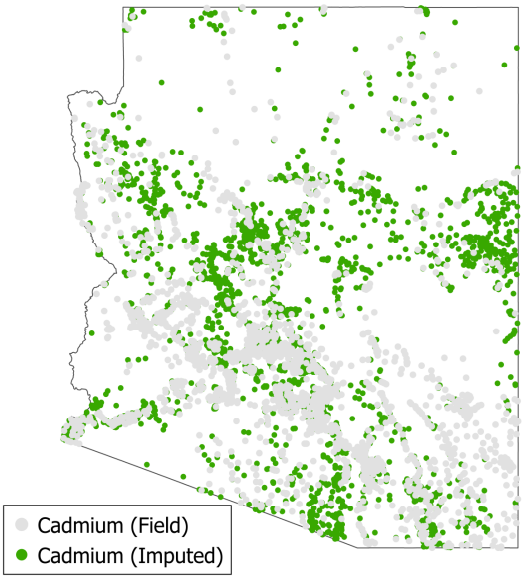
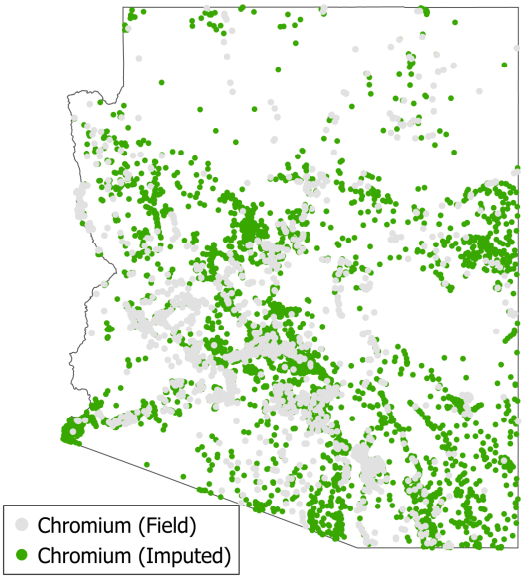


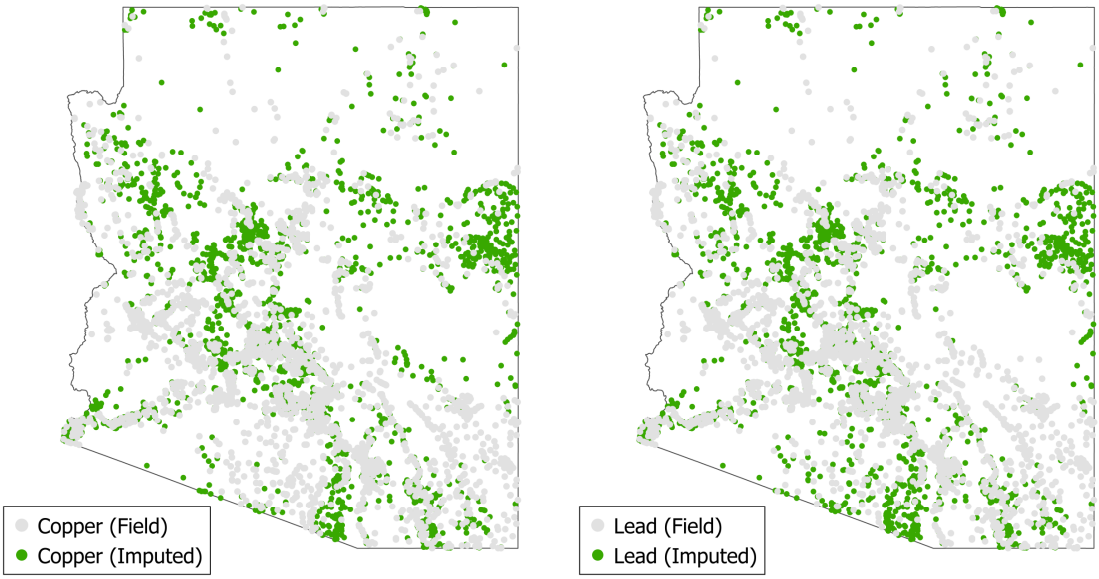
Figure S10: Amelia imputation diagnostic plots for Vanadium (V) data.

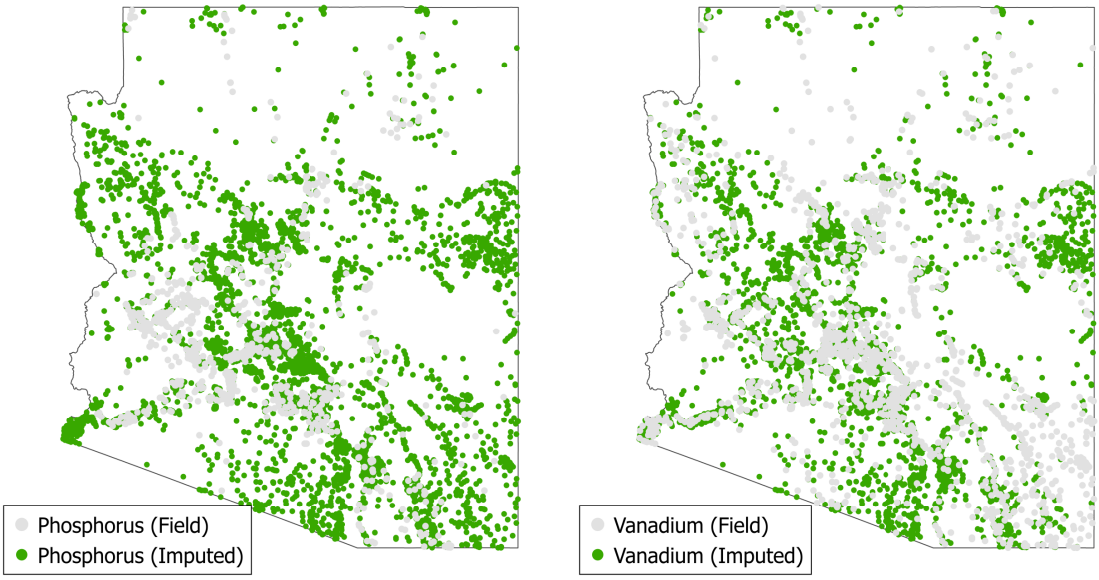

Antimony (Field)
Antimony (Imputed)


Arsenic (Field)
Arsenic (Imputed)


Cadmium (Field)
Cadmium (Imputed)


Chromium (Field)
Chromium (Imputed)


Copper (Field)
Copper (Imputed)
Lead (Field)
Lead (Imputed)


Phosphorus (Field)
Phosphorus (Imputed)
Vanadium (Field)
Vanadium (Imputed)

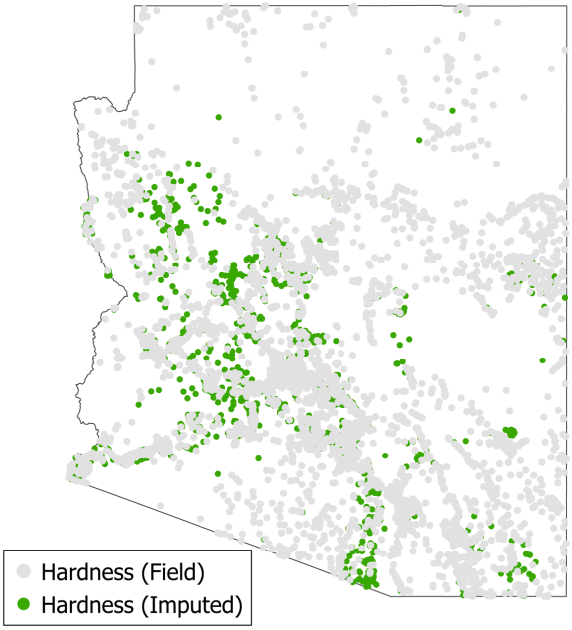


Figure S11: Geospatial locations for field (grey symbols) and AMELIA imputed (green symbols) data of Arizona (antimony, arsenic, cadmium, chromium, copper, lead, phosphorus, vanadium, and hardness).

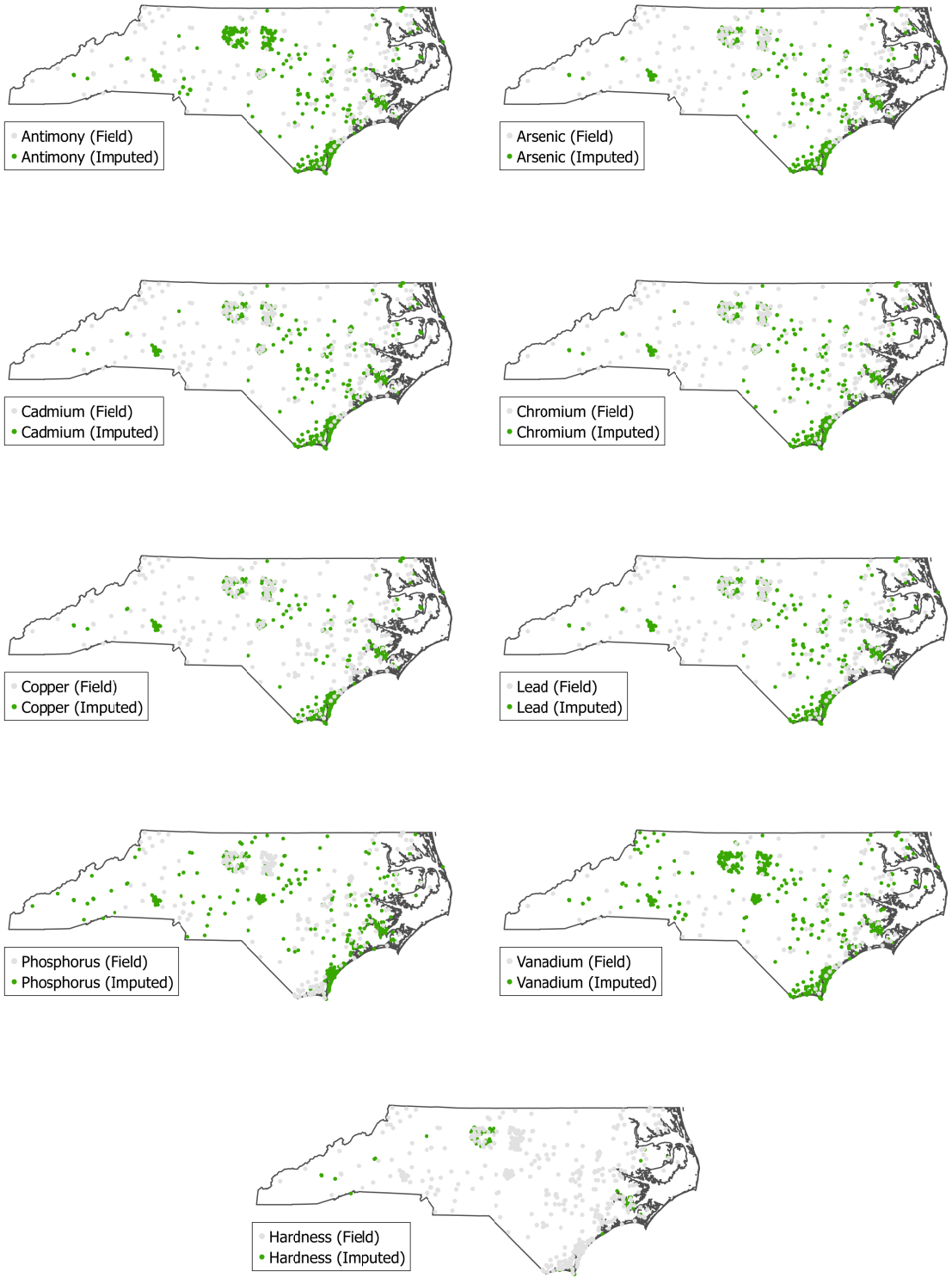


Figure S12: Geospatial locations for field (dark red symbols) and AMELIA imputed (light red symbols) data of North Carolina (antimony, arsenic, cadmium, chromium, copper, lead, phosphorus, vanadium, and hardness).