Repurposing Underutilized Monitoring Data from Contaminated Sites for Sustainable Groundwater Characterization

Landi Laura[a]*, Rotiroti Marco[b], Zanotti Chiara[b], Amorosi Alessandro[a], Dinelli Enrico[a], Filippini Maria[a]

[a]*Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Zamboni 67, 40126 Bologna, Italy*

[b]*Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, 20126 Milan, Italy*

** Corresponding author: laura.landi11@unibo.it*

SUPPLEMENTARY MATERIAL

S1. Additional details on data collection

The preliminary site selection (see Section 2.2 of the main text) identified 108 contaminated sites as potentially suitable for this study, including 46 different types of activities (Tab. S1). The most frequently investigated activities were petrol stations and landfills, accounting for 30% and 11% of the collected sites, respectively. The temporal coverage of the collected samples was highly variable, spanning 43 years from February 1981 to December 2023 (Fig. S1).

The information associated to each sample, carefully harvested from the suitable sites, was systematically stored and aggregated into a unified, homogeneous data matrix, which includes:

- three unique IDs for sample, piezometer, and site, respectively;
- sampling date;
- 18 chemical species and physico-chemical parameters (Arsenic (As), Iron (Fe), Manganese (Mn), Sulfate (SO_4), Ammonium (NH_4), Nitrate (NO_3), Nitrite (NO_2), pH, ORP, Electrical Conductivity (EC), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Oxidizability (Ox), Chlorine (Cl), Calcium (Ca), Magnesium (Mg), Potassium (K) and Sodium (Na));
- 6 site-specific "index contaminants" related to the type of anthropogenic activity causing the contamination (Total Hydrocarbons (THC), Lead (Pb), Aluminum (Al), Boron (B), Total Organohalogenated compounds (TOH), Vinyl Chloride Monomer (VCM));
- information on sampling procedures (i.e., field filtering);
- laboratory performing the analysis;
- counter-analysis results from private and ARPA laboratories;
- duplicate samples (filtered/unfiltered);
- piezometer filter depth (beginning of the filter);

- piezometer total depth;
- position relative to the contamination source.

Concentrations of index contaminants were collected for each site depending on the type of contamination (Tab. S1). The most frequently analyzed index contaminants were Total Hydrocarbons (THC) and lead (Pb), measured in 83% and 29% of the collected sites, respectively.

Table S1 *Collected sites, main activities and index contaminants (empty cells denote unavailable information)*

Site ID	Main Activity	Index contaminant(s)
1	Metal Coating	-
2	Landfill	-
3	Petrol Station	THC
4	Petrol Station	-
5	Petrol Station	THC
6	Petrol Station	THC
7	Sugar Factory	THC
8	Sugar Factory	THC
9	Petrol Station	THC
10	Mechanical Factory	THC, Al, Pb
11	Concrete Manufactory	THC, Pb
12	Locomotive Storage	THC
13	Petrol Station	THC
14	Landfill	THC
15	Landfill	THC
16	Petrol Station	THC, Pb
17	Landfill	THC, Al, Pb
18	Agri-Food Factory	THC
19	Wood Factory	THC
20	Gas Extraction Plant	THC
21	Petrol Station	THC
22	Petrol Station	THC, Pb
23	Oil Storage	THC, Pb
24	Gas Extraction Plant	-
25	Pottery Factory	THC
26	Sugar Factory	-
27	Landfill	THC, Pb
28	Petrol Station	THC
29	Petrol Station	THC, Pb
30	Industrial Area	THC, Pb
31	Pesticide Storage and Commercial activities	THC, Pb
32	Petrol Station	THC, Pb
33	Sugar Factory	THC, Pb
34	Distillery	-
35	Oil Storage	THC
36	Bitumen Factory	THC, Pb
37	Landfill	-
38	Mechanical Factory	THC
39	Electrical components commerce and Textile Factory	THC
40	Bricks Factory	THC
41	Sugar Factory	THC
42	Landfill	-
43	Petrol Station	THC
44	Landfill	-
45	Waste Dumping	THC
46	Petrol Station	THC

47	Waste Dumping	Pb
48	Oil&Gas Extraction Area	THC
49	Electrical Appliances Factory	THC, Al, Pb
50	Sugar Factory	THC
51	Gas Plant	-
52	Treatment Plant for Liquid Waste	THC
53	Petrol Station	THC
54	Petrol Station	THC
55	Metallurgical Factory	THC
56	Petrol Station	THC
57	Gas Plant	THC
58	Petrol Station	THC
59	Trucks Storage and Remediation	THC, Pb
60	Farmhouse	THC, Pb
61	Farmhouse	-
62	Petrol Station	THC
63	Waste Dumping	THC
64	Petrol Station	THC, Pb
65	Clay Quarry	THC, Pb
66	Petrol Station	THC
67	Petrol Station	THC, Pb
68	Landfill	-
69	Tractors Factory	THC, TOH
70	Metal Scraps Storage	THC, Pb
71	Engines Factory	THC, TOH
72	Ironware Commerce	THC
73	Oil Storage	THC
74	Waste Dumping	-
75	Petrol Station	THC
76	Waste Dumping	THC
77	Metal Waste Storage	THC, Pb
78	Landfill	-
79	Aggregated Workshops	THC
80	Area for Vehicle Refueling	THC
81	Petrol Station	THC
82	Petrol Station	-
83	Geothermical Plant	THC, B
84	Geothermical Plant	THC, B
85	Building Material Manufactory	THC
86	Incinerator	THC, Pb
87	Building Material Manufactory	THC, Pb
88	Disposed Train Station	THC, Pb
89	Metallurgical Workshop	THC, TOH
90	Petrol Station	THC
91	Aggregated Industrial Activities	THC, Pb, TOH, CVM
92	Aggregated Industrial Activities	CVM
93	Aggregated Industrial Activities	THC, TOH, CVM
94	Distillery	THC
95	Petrol Station	THC
96	Petrol Station	THC, Pb
97	Petrol Station	THC, Pb
98	Petrol Station	THC, Pb
99	Disposed Hospital Area	THC
100	Disposed Barrack	THC
101	Commercial Area	THC
102	Petrol Station	THC
103	Hippodrome	THC
104	Recycling Plant	THC
105	Historical hydrocarbons contamination	THC

106	Petrol Station	THC
107	Landfill	Pb
108	Landfill	THC, Pb

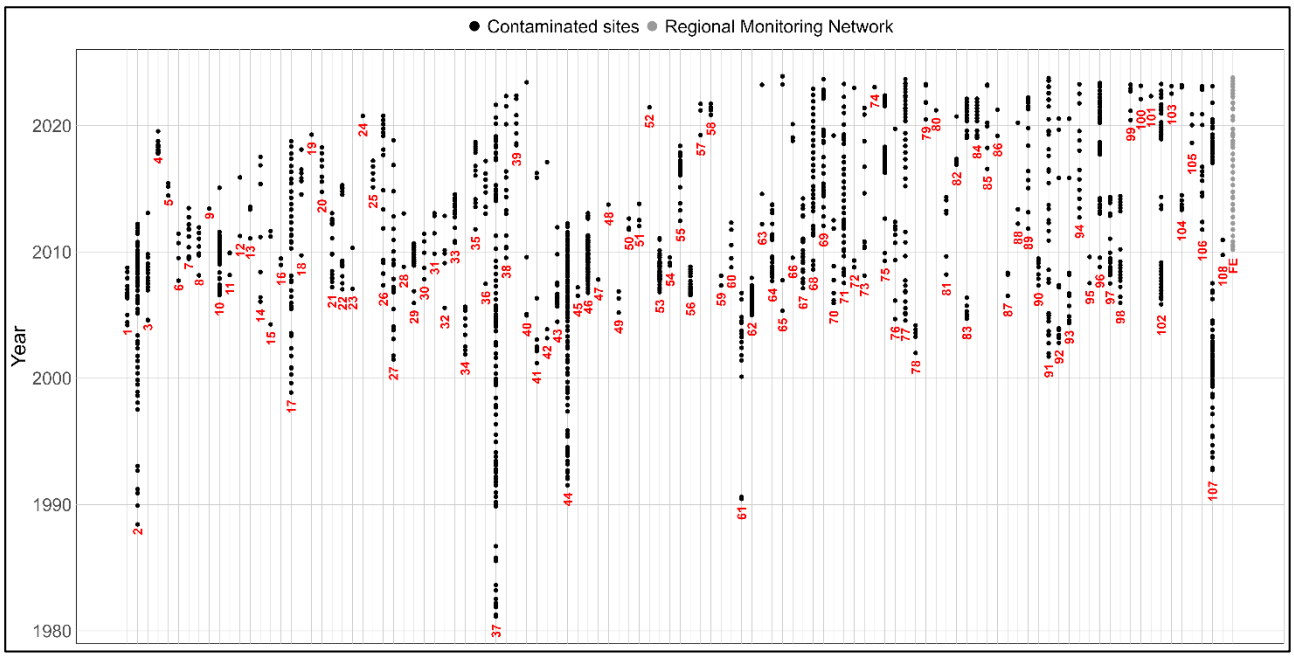

Fig. S1 *Temporal coverage of collected samples from contaminated sites and regional monitoring network*

S2. Additional details on the treatment workflow

S2.1 Censored data

In the sample matrix, censored data reported as below the limit of detection (LOD) were handled using the LOD/2 substitution method (see Section 2.3.1 of the main text). Due to the heterogeneity of data sources, detection limits varied for the same variable. To ensure consistency, we selected the minimum LOD with a significant presence (≥ 5% of all the censored data, Tab. S2) for each variable. Censored data were then replaced with half of this selected LOD.

Table S2 *Selected minimum significant LOD for management of censored data (empty cells denote the absence of censored data)*

Variable	min LOD
As (µg/L)	< 0.5
Fe (µg/L)	< 5
Mn (µg/L)	< 3
SO_4 (mg/L)	< 1
NH_4 (mg/L)	< 0.02
NO_3 (mg/L)	< 0.1
NO_2 (mg/L)	< 0.003
pH	-
ORP (mV)	-
EC (µS/cm)	-
COD (mg/L)	< 4
Ox (mg/L)	< 0.5

DO (mg/L)	< 0.01
Cl (mg/L)	-
Ca (mg/L)	-
Mg (mg/L)	-
K (mg/L)	< 1
Na (mg/L)	-

S2.2 Influence of field filtering on metal concentrations

One of the main challenges in analyzing an aggregated dataset is the variability of sampling protocols. In particular, field filtering has significant impacts on data reliability, as the concentration of certain species, especially metals, can be influenced by suspended colloidal particles. A comparison of duplicate samples (filtered vs. unfiltered) for As, Fe and Mn (Fig. S2), confirmed that unfiltered samples generally exhibit higher concentrations and greater variability. For this reason, we decided to rely exclusively on filtered samples.

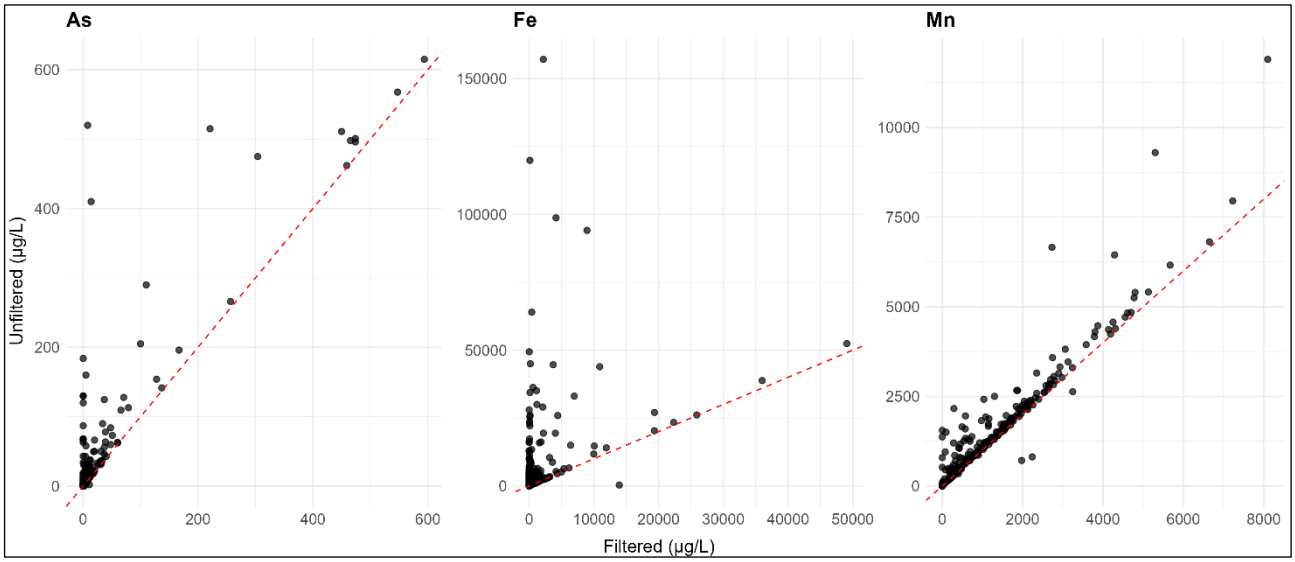

Fig. S2 *Comparison of filtered and unfiltered measurements for duplicate samples of As, Fe and Mn. 1:1 line is indicated in red*

S2.3 Data removal throughout treatment steps

Throughout the various steps of the treatment workflow, different values, samples and piezometers were progressively removed (Fig S3). From the initial 1'695 piezometers, 215 were excluded for tapping deeper aquifers, since this study focuses on the shallow aquifer. After evaluating the different sampling protocols, 7'037 out 11'974 samples labeled as unfiltered or unknown were removed (487 piezometers providing only unfiltered or unknown samples were entirely removed). After the analysis of the time series outliers 46 samples identified as anomalous and 398 "extreme" outliers were removed. The trend analysis identified and removed 111 piezometers associated with temporal trends and, therefore, potentially contaminated.

The pre-selection (PS) with site-specific criteria identified and removed 293 piezometers where the index contaminants exceeded at least once the PS threshold. Through the last step, site-scale spatial outliers were investigated and 101 piezometers considered anthropogenic anomalies were removed from the dataset.
After the two main steps of the workflow (see Section 2.3 in the main text), we obtained two data matrices: an "anthropized" and a "natural" dataset representing, respectively, the actual composition of shallow groundwater incorporating both natural and anthropogenic signals, and the groundwater natural composition. A comparison of the statistical summaries of chemical species between the two datasets illustrates the effect of removing the anthropogenic signal (Tab. S3).
Despite the natural dataset retaining only 49% of the piezometers constituting the anthropized dataset, the overall spatial distribution remains largely unaffected (Fig. S4), indicating that the anthropogenic signal was removed in a spatially homogeneous manner across the sites.

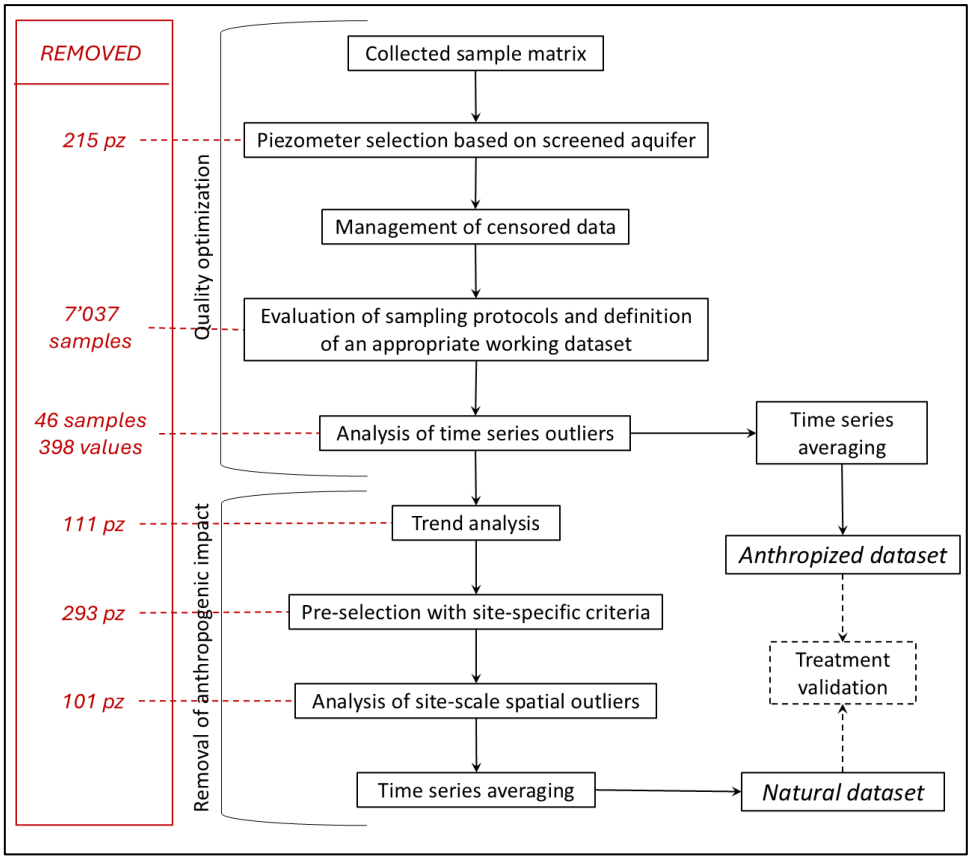

Fig. S3 *Workflow and summary of values/samples/piezometer removal for each step*

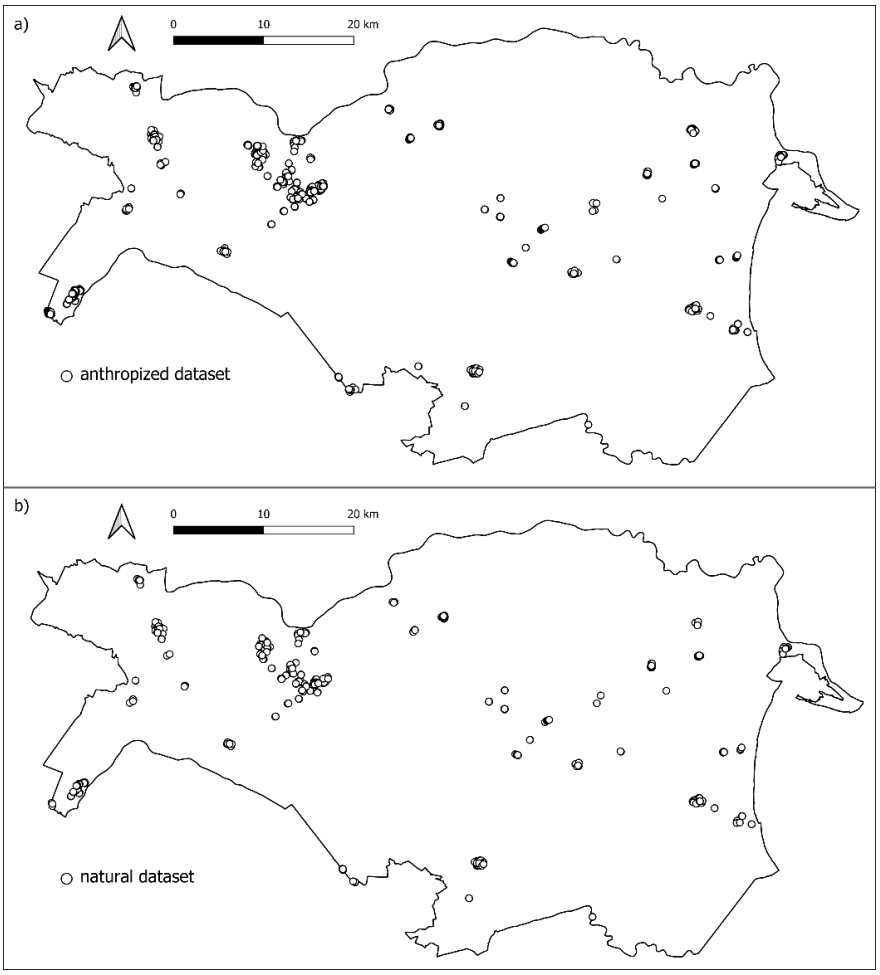

Fig. S4 *Spatial distribution of the anthropized (a) and natural (b) datasets*

Table S3 *Statistics of the main analites for the anthropized and natural datasets*

Variable	**Dataset**							
	Anthropized				**Natural**			
	min	*max*	*mean*	*sd*	*min*	*max*	*mean*	*sd*
Fe (µg/L)	2.50	205750.00	3739.25	10405.64	2.50	30734.50	2421.18	4297.94
Mn (µg/L)	1.50	17085.00	1249.87	1274.93	1.50	5649.50	1161.54	980.57
As (µg/L)	0.20	474.00	14.70	38.41	0.25	192.75	10.38	22.57
SO_4 (mg/L)	0.50	3240.50	396.38	609.97	0.50	2290.00	410.76	562.01
NH_4 (mg/L)	0.007	68.03	5.40	9.98	0.007	14.04	2.38	3.38
NO_3 (mg/L)	0.005	774.34	19.64	89.94	0.005	55.78	4.37	9.79
NO_2 (mg/L)	0.003	10.10	0.15	0.87	0.005	0.55	0.04	0.10
pH	5.77	11.99	7.32	0.60	6.48	11.99	7.25	0.60
ORP (mV)	-422.00	235.00	-59.97	97.10	-345.75	235.00	-51.69	100.55
EC (µS/cm)	190.00	39720.50	3105.98	5510.54	190.00	37502.50	3741.15	6263.64
COD (mg/L)	6.50	4290.00	166.33	543.89	6.50	81.00	26.54	17.91
Ox (mg/L)	0.25	27.30	6.38	5.85	0.25	18.40	4.26	4.10
DO (mg/L)	0.001	26.81	1.77	2.39	0.001	7.14	1.41	1.69
Cl (mg/L)	16.25	16400.00	1943.94	4400.31	16.75	15660.00	1557.57	4046.64
Ca (mg/L)	56.50	623.00	206.61	126.92	68.00	425.75	189.96	77.41
Mg (mg/L)	12.50	276.50	81.78	59.59	15.40	168.75	78.71	41.57
K (mg/L)	0.50	111.75	13.69	21.01	0.50	45.00	8.15	11.17
Na (mg/L)	21.00	710.00	255.86	196.69	21.00	710.00	205.20	171.14

S3. Additional details on multivariate analysis and geochemical zonation

Through multivariate analysis (see Section 2.4 of the main text) we identified six clusters (C1-C6; Tab. S4), each representing a hydrogeochemical facies associated with different redox zones and brackish groundwater (see Section 3.3 of the main text).

For each cluster we identified the most representative variables: C1-C4 are mainly represented by the Mn-Fe pair, C5 by As and C6 by EC. After the cluster identification, all the previously imputed missing values were removed, in order to consider only the real measured values. Moreover, all piezometers where the value of key variables for the pertaining cluster had been previously imputed were discarded (164 piezometers), minimizing the influence of missing data imputation on the analysis.

Cluster C1 (35 piezometers) is characterized by strongly positive scores of FN3, suggesting relatively higher ORP values and slightly reductive processes where Mn is mobilized in groundwater (FN1 negative scores) while Fe and As are stable in the solid phase (median ORP, Mn, Fe and As of 223.5 mV, 1,489 µg/L, 2.5 µg/L and 2.6 µg/L, respectively). Cluster C2 (133 piezometers) is characterized by slightly positive scores of FN3 and negative scores of FN1, suggesting a slight progression of reducing processes, with negative ORP and increasing Mn and Fe reduction, still dominated by Mn mobilization (median ORP, Mn, Fe and As of -2.9 mV, 844.8 µg/L, 227 µg/L and 2.5 µg/L, respectively). Cluster C3 (91 piezometers) is characterized by slightly negative scores of FN3, and suggests redox conditions comparable with C2, with an increased influence of brackish water (median Fe, Mn, As and EC of 607 µg/L, 867 µg/L, 2.8 µg/L and 1,701 µS/cm, respectively). Cluster C4 (21 piezometers) is primarily characterized by strongly positive scores of FN2, suggesting a transition to moderately reducing conditions associated with strong mobilization of Fe and Mn (median Fe, Mn and As of 13,350 µg/L, 1,960 µg/L and 16.3 µg/L respectively). Cluster C5 (21 piezometers) is predominantly characterized by strongly positive scores of FN1, indicating stronger reducing conditions associated with intense As and Fe mobilization (median As and Fe of 75.1 µg/L and 6,400 µg/L, respectively). Cluster C6 (15 piezometers) is characterized by strongly negative scores of FN3, suggesting brackish conditions in piezometers located near the coast (median EC of 20,865 µS/cm is one order of magnitude greater than that of the other clusters).

Table S4 *Centroids of Cluster Analysis applied to the FA scores of the natural dataset and key variable(s) representing each cluster. The higher value for each cluster is highlighted in bold*

Cluster	Factor			key variable(s)
	FN1	FN2	FN3	
C1	-0.26	-0.1	**2.19**	Fe-Mn
C2	-0.34	-0.18	**0.52**	Fe-Mn
C3	-0.09	-0.21	**-0.76**	Fe-Mn
C4	-0.24	**2.98**	0.69	Fe-Mn
C5	**3.47**	0.16	-0.62	As
C6	-0.06	-0.77	**-2.66**	EC

Additionally, cluster C5 is characterized by lower SO_4 concentration, likely indicating sulfate reduction processes occurring alongside As mobilization.

A scatterplot of EC vs SO_4 (Fig. S5) shows different conditions among the clusters.

Brackish groundwater of cluster C6 is clearly observed in the top-right corner of the plot, indicating the highest salinity due to the influence of seawater.

Piezometers from clusters C2, C3 and C4 tend to align with the 1:1 line, revealing higher salinity in cluster C3, which have more piezometers located near the cost.

In contrast, piezometers from cluster C5 (and some from cluster C4), which exhibit the most advanced reducing conditions, deviate from the 1:1 line suggesting a sulfate depletion likely attributable to sulfate reduction. C5 likely represents groundwater under strong reducing conditions where sulfate reduction processes occur alongside strong As mobilization (see Section 3.3.2 of the main text).

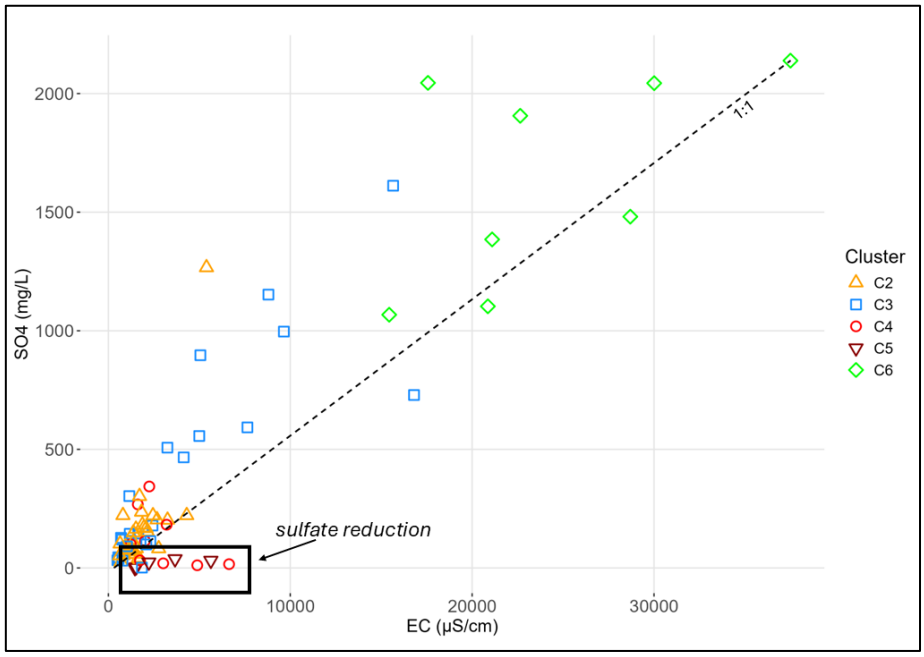

Fig. S5 *Scatterplot SO_4 vs. EC of clustered piezometers. Black frame indicates points where sulfate reduction is likely occurring. Cluster C1 is not showing due to lack of measurements*