Scientific Reports

https://doi.org/10.1038/s41598-026-38887-9

Article in Press

Integrating multi-spectral remote sensing and machine learning for quantifying and mapping heavy metal contamination in the Tarim River Basin, China

Received: 15 September 2025

Accepted: 31 January 2026

Published online: 01 April 2026

Cite this article as: Zhao Y., Mu Y., Luo P. *et al.* Integrating multi-spectral remote sensing and machine learning for quantifying and mapping heavy metal contamination in the Tarim River Basin, China. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38887-9

Yang Zhao, Yong Mu, Pingping Luo, Jianxin Zhang, Madhab Rijal, Zhihui Yang, Chengguang Lai, Jiacha Chen, Ahmed Elbeltagi & Bam H. N. Razafindrabe

Integrating Multi-Spectral Remote Sensing and Machine Learning for Quantifying and Mapping Heavy Metal Contamination in the Tarim River Basin, China

Yang Zhao[1,2,3,4,5†], Yong Mu[6,†],Pingping Luo[1,2,3,4,5*],Jianxin Zhang[7*], Madhab Rijal[1,2,3,4,5], Zhihui Yang[8], Chengguang Lai[9], Jiachao Chen[10], Ahmed Elbeltagi[11], Bam H.N. Razafindrabe[12]

1. Key Laboratory of Subsurface Hydrology and Ecological Effects in Arid Region, Chang'an University, Ministry of Education, Xi'an 710054, Shaanxi Province, China

2. Shaanxi Province Innovation and Introduction Base for Discipline of Urban and Rural Water Security and Rural Revitalization in Arid Areas, Chang'an University, Xi'an 710054, Shaanxi Province, China

3. School of Water and Environment, Chang'an University, Xi'an 710054, Shaanxi Province, China;

4. Xi'an Monitoring, Modelling and Early Warning of Watershed Spatial Hydrology International Science and Technology Cooperation Base, Chang'an University, Xi'an 710054, Shaanxi Province, China

5. Key Laboratory of Eco-hydrology and Water Security in Arid and Semi-arid Regions of the Ministry of Water Resources, Chang'an University, Xi'an 710054, Shaanxi Province, China

6 Shaanxi Agricultural Development Group Co., Ltd., No. 7 Guangtai Road, High-tech Zone, Xi'an City710075, Shaanxi Province, China

7. School of Architecture, Chang'an University, Xi'an 710061, China

8. Institute of Environmental Science and Engineering, School of Metallurgy and Environment, Central South University, 410083, Changsha, China;

9.School of Civil Engineering and Transportation, South China University of Technology, Guangzhou,510641, China

10. Graduate School of Engineering, Kyoto University, Katsura, Nishikyo-ku, Kyoto 615-8530, Japan

11. Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt;

12. Faculty of Agriculture, University of the Ryukyus,1 Senbaru, Nishihara, Okinawa 903-0213 Japan

†The authors are equal to each other as the cofirst author.

***Corresponding author: Pingping Luo, Email: lpp@chd.edu.cn; Jianxin Zhang, Email:** zhangjx@chd.edu.cn

A B S T R A C T

Heavy metal (HM) contamination in arid inland river basins is intensifying with economic development, posing ongoing ecological and public health risks. Conventional monitoring is constrained by landscape heterogeneity and limited field data, hindering basin-scale spatiotemporal analysis. This study integrates multispectral Sentinel-2A imagery, Random Forest (RF) regression, and harmonized ground observations to quantify and map As, Cd, Cu, Pb, and Zn in soils and waters of China's Tarim River Basin (TRB), and to evaluate related health risks using the U.S. EPA model. Field data showed soil ΣHMs averaged 118.71 mg kg^{-1} (95.15–191.32 mg kg^{-1}) and river ΣHMs 3.89 μg L^{-1} (1.14–99.00 μg L^{-1}), with Cd and As above background levels. RF models yielded preliminary spatial patterns ($R^2 = 0.741–0.999$; RMSE = 0.12–0.38 mg kg^{-1} for soils, 0.21–0.57 μg L^{-1} for waters). Spatial patterns revealed As and Cd hotspots in Aksu, Cu and Pb enrichment in central and southeastern sub-basins, and Zn predominance in the east. Pearson correlation and principal component analyses attributed Pb–Cu–Zn mainly to natural sources, As–Cd to agriculture, and riverine metals to atmospheric deposition and hydrology. Oral ingestion was the major exposure route; ~10% of soil samples exceeded the non-carcinogenic hazard index for children, while As and Cd posed carcinogenic risks $>1 \times 10^{-4}$. The geo-accumulation index ($I\_geo > 2$ for As) and HPI > 100 in Hotan, Kashgar, and Aksu indicated localized hazards. Recommended actions include RS-based quarterly monitoring, targeted As/Cd control, basin-wide HM coordination, child health screening, and nature-based remediation to reduce HM risks by 2035. This work provides the first high-resolution, basin-wide HM assessment in the TRB, demonstrating Sentinel-2A-RF inversion as a cost-effective tool for hotspot identification, risk-informed management, and alignment with China's dual-carbon goals, rural revitalization, and UN Sustainable Development Goals.

Keywords: Tarim River Basin; heavy metals; remote sensing inversion; health risk assessment; sustainable development

1. Introduction

Heavy metal contamination in soil and water poses a persistent threat to ecosystem integrity, food security, and human health, particularly in arid inland river basins experiencing rapid socio-economic growth[1-3]. Unlike organic pollutants, heavy metals are non-biodegradable, bioaccumulative, and capable of long-range transport via hydrological and atmospheric pathways, thereby amplifying environmental persistence and toxicity[4, 5]. The Tarim River Basin—China's largest inland river system—exemplifies these challenges[6, 7]. As the economic backbone of southern Xinjiang, the basin's accelerated development has enhanced livelihoods but heightened environmental vulnerability through heavy metal accumulation in agricultural soils and surface/groundwater[8, 9]. Recent studies report elevated heavy metal levels linked to natural geochemical processes and anthropogenic sources (e.g., irrigation runoff, industrial effluents), jeopardizing biodiversity, rural livelihoods, and public health in this ecologically fragile region[10, 11].

Despite growing awareness, basin-scale assessment of heavy metal contamination remains constrained by the heterogeneity of arid landscapes, limited field sampling, and fragmented data availability[12, 13]. Conventional monitoring relies on labor-intensive, costly field sampling and laboratory analysis[14], which fail to capture spatiotemporal dynamics over vast, inaccessible terrains[15, 16]. This limitation impedes risk quantification and mitigation, especially in the Tarim Basin where transboundary influences and climate variability complicate pollution tracking[17, 18]. Although research on heavy metal pollution in the basin has increased, most studies are spatially or temporally fragmented, depending on sparse surveys or single-period data[19, 20]. The lack of integrated, large-scale assessments that combine multi-source datasets with advanced modeling hinders effective monitoring of contamination dynamics and human health risks[21]. Advances in remote sensing and data-driven modeling provide new opportunities to overcome these limitations[22]. High-resolution multispectral satellite imagery, such as Sentinel-2A, now enables large-scale environmental surveillance[23, 24]. When paired with machine learning—particularly Random Forest (RF) regression—these data facilitate robust inversion of non-linear spectral–contaminant relationships[25], outperforming linear methods under data scarcity[26, 27]. The integration of these techniques supports high-resolution inversion modeling[28], leveraging surface reflectance patterns related to soil mineralogy, vegetation stress[29], and suspended particulates as indirect indicators of heavy metal presence[30, 31]. Prior studies using Landsat, WorldView, or hyperspectral data have estimated single-metal concentrations in farmland or mining sites; however, four major challenges remain: (i) the lack of basin-wide, high-resolution coverage in arid regions; (ii) coupled soil–water evaluation accounting for hydrological connectivity; (iii) weak linkage between remote-sensing outputs and quantitative health-risk metrics; and (iv)

uncertainty in model validation under limited ground-truth data[32-34]. Addressing these issues requires a comprehensive framework that unites field observations, multi-spectral data, and advanced machine learning for accurate prediction and health-risk assessment[35, 36].

Although numerous studies have explored heavy metal pollution in parts of the TRB, key knowledge gaps persist regarding: (i) the spatiotemporal distribution and trends of multiple metals at the basin scale; (ii) associated human health risks, dominant exposure pathways, and metal-specific contributions; and (iii) actionable mitigation strategies tailored to soil and water matrices[37-39]. Most existing work remains temporally fragmented and spatially localized, limiting its utility for regional environmental governance[40]. Closing these gaps aligns with China's ecological civilization, rural revitalization, and sustainable development goals for vulnerable inland basins[10, 41]. Accordingly, this study pursues two main objectives: (1) to evaluate the capability of integrating Sentinel-2A multispectral imagery with RF regression for quantifying major heavy metals—arsenic (As), cadmium (Cd), lead (Pb), copper (Cu), and zinc (Zn)—across heterogeneous arid landscapes; and (2) to characterize their spatial–temporal variations and associated human health risks[42].

To achieve these goals, this research integrates Sentinel-2A imagery, RF modeling, and field observations to generate the first basin-wide, 60 m-resolution maps of five priority metals across the TRB. Health risks were quantified following the United States Environmental Protection Agency (US EPA) methodology to assess non-carcinogenic and carcinogenic effects for adults and children via ingestion, dermal, and inhalation exposure. The work advances sustainable development in arid inland basins through four contributions: (i) Methodological innovation – A transferable remote sensing inversion framework for simultaneous soil and water contamination assessment, leveraging ensemble learning and rigorous cross-validation to address data scarcity; (ii) Comprehensive risk appraisal – integrating geo-accumulation index (I_geo), Heavy Metal Pollution Index (HPI), and multi-pathway health-risk modeling to transition from pollution detection to risk prioritization; (iii) Policy Relevance – Visualizing pollution hotspots, exposure pathways, and specific metal risks to guide targeted remediation and policy alignment with China's "dual carbon" and rural revitalization strategies; and (iv) Open-data synthesis – Constructing a unified watershed-scale dataset (2011-2022) from 12 peer-reviewed sources for future model improvements and management decision support.

The paper is organized as follows: Section 2 introduces the study area, datasets, and methodological framework; Section 3 presents inversion performance, spatial–temporal distribution patterns, and health-risk results; Section 4 discusses key drivers, policy implications, and uncertainties; and Section 5 concludes with major findings and directions for future research. By integrating multi-spectral remote sensing, machine learning, and health-risk science, this study establishes a comprehensive approach for assessing heavy metal contamination and advancing sustainable development in arid inland river basins.

2. Datasets and methodology

2.1 Study area

The Tarim River Basin (73–97°E, 34–45°N), located in southern Xinjiang Uygur Autonomous Region, Northwest China (Fig. 1), serves as the study area owing to its pronounced ecological vulnerability and escalating heavy metal contamination risks in an arid inland setting. Although the basin extends into Kazakhstan, Kyrgyzstan, and Tajikistan, the analysis concentrates on the Chinese sector, where dense human activity, large-scale irrigation agriculture, and industrial development are primarily located. Encompassing approximately 1.05 million km² across five prefectures (Bayingolin Mongolian Autonomous Prefecture, Aksu Region, Kizilsu Kirgiz Autonomous Prefecture, Kashgar Region, and Hotan Region), the basin constitutes China's second-largest inland river system. By 2018, this area supported 11.59 million residents, 2.03 million ha of cultivated land, and a regional GDP of 0.34 trillion RMB, representing significant contributions to Xinjiang's population, agricultural output, and economic growth[43]. However, this rapid growth, accompanied by a sharp rise in energy consumption, has intensified environmental pressures, making the basin's soil and water resources susceptible to pollution[44]. The vast scale of the area poses significant challenges to conventional monitoring, underscoring the necessity of the remote sensing and machine learning approach presented herein for supporting environmental sustainability.

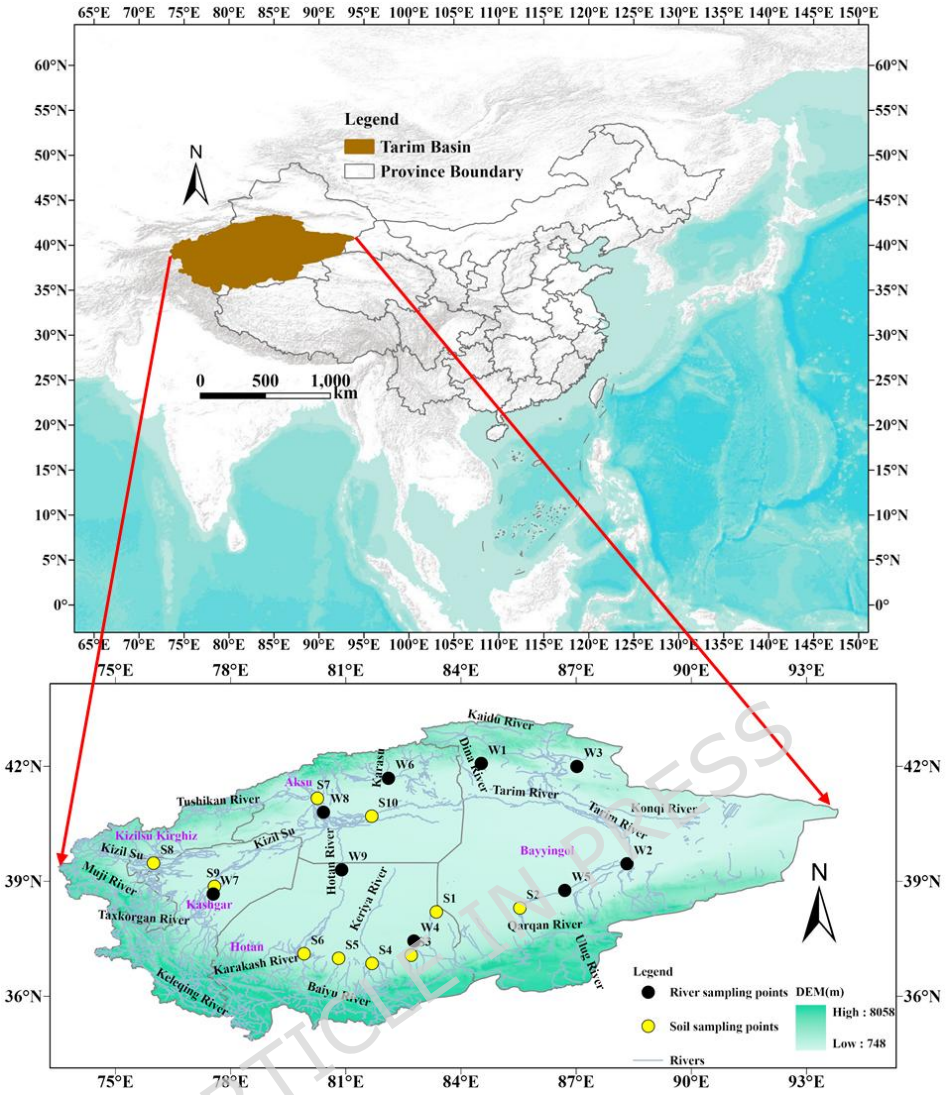


Fig. 1. Schematic of the study area. Created by the authors using ArcGIS Pro 3.6 (Esri; https://www.esri.com/en-us/arcgis/about-arcgis/overview)

2.2 Heavy metal data source

This study synthesizes a comprehensive dataset of heavy metal concentrations for the Tarim River Basin by integrating data from multiple peer-reviewed literature sources. This approach was necessary to overcome the spatial and temporal limitations of any single dataset and to capture the large-scale contamination dynamics across this vast and heterogeneous region. The following sections detail the data acquisition strategy, the rationale for sampling density, and the rigorous protocols implemented to harmonize the multi-source data and ensure its reliability for machine learning modeling.

2.2.1 Data Compilation, Sampling Rationale, and Quality Assurance

A systematic literature search was conducted in the Web of Science™ Core Collection and China National Knowledge Infrastructure (CNKI) databases for publications from 2012 to 2025, using the search string TS = (Tarim) AND TS = ("heavy metal") in the "topic" field. After removing duplicates, this yielded 51 unique records.

Studies were included if they: (i) reported concentrations for at least one target heavy metal (As, Cd, Pb, Cu, Zn); (ii) provided explicit geographical coordinates or detailed site descriptions enabling accurate geolocation; and (iii) followed recognized national or international analytical standards. This screening retained 12 publications for data extraction, from which we compiled two datasets—one for soil/sediment and one for water (surface, irrigation, and shallow groundwater)—recording mean concentrations per metal at each unique sampling location (Supplementary Table S1 and Supplementary Table S2).

The resulting datasets comprise 10 soil/sediment sites and 9 water sites. Although these represent a synthesis of multiple studies rather than a single new field campaign, the number and distribution of sampling points in the source publications were guided by established technical specifications, study objectives, and practical constraints, ensuring spatial representativeness across the basin's diverse land-use types, hydrotopographic zones, contamination hotspots, and areas of hydrochemical variability.

For soil and sediment, sampling density in source studies followed guidelines such as the DZ/T 0258–2014 Multi-Target Regional Geochemical Survey Specification (1:250,000 scale), which recommends grid-based densities of approximately one point per 4–8 km² for regional assessments[41], supplemented by higher-density targeted sampling in priority areas (e.g., major oases, agricultural zones, and upper reaches) to capture localized gradients from anthropogenic activities[10, 45, 46]. Larger point counts in specific studies (e.g., 193 or 78 sites) reflected heterogeneity in land use, accessibility, and the need for robust local geochemical baselines[10, 45-47].

For water, site numbers (typically 30–90 per campaign) aligned with standards like HJ/T 164–2004 Technical Specifications for Environmental Monitoring of Groundwater[48], with strategic placement along hydrological features (main streams, tributaries, confluences), land-use gradients, proximity to urban/agricultural centers, and zones of elevated hydrochemical variability or contamination risk (e.g., Aksu-Shaya section and intensive cotton farming areas)[42, 47-52]. Complementary background samples from minimally disturbed sites (e.g., farmland or grassland) were often included for reference[10, 45, 46] (Tables 1).

Table 1. Background values of heavy metals in the Tarim River Basin

Background values	As	Cd	Pb	Cu	Zn
Background value for soils in Xinjiang[41, 53], (mg/kg)	11.2	0.12	19.4	26.7	68.8
Environmental quality standards for III surface water[49], (μg/L)	50	5	50	1000	1000
Environmental quality standards for V surface water[49], (μg/L)	100	10	100	1000	2000

Integrating data from 12 sources spanning 2011–2022 inevitably introduces heterogeneity in sampling and analytical protocols—a common issue in meta-analyses of environmental data (Table S3). To address this and ensure dataset robustness for modeling:

- We prioritized studies explicitly adhering to national standards (e.g., DZ/T 0258–2014 for soil; HJ/T 164–2004 for water)[41, 48], promoting consistency in collection and processing.
- All retained publications documented rigorous QA/QC (Tables S1), including certified reference materials (e.g., GBW, NIST), duplicates, and replicates (typically $\geqslant 3$), yielding analytical accuracies >98%[10, 41, 42, 48, 49];
- Extracted concentrations were standardized to uniform units (mg/kg for soil/sediment; μg/L for water), with verification of consistent preprocessing (e.g., acidification to pH <2, 0.45 μm filtration) and validated analytical methods (e.g., ICP-MS, AFS)[10, 42, 48, 49].
- In the machine learning phase, heterogeneity was further mitigated via feature scaling (normalization) and robust cross-validation to reduce impacts from varying ranges/distributions and improve model stability and feature comparability.

These steps yield a harmonized, fit-for-purpose dataset offering broader spatial coverage than individual studies, essential for basin-scale contamination mapping and informed environmental management.

2.3 Methodology

The research framework (Fig. 2) combines remote sensing and field-based observations to evaluate heavy metal contamination in the Tarim River Basin. Heavy metal concentrations in soil and water were obtained through field sampling, supplemented by spatial information derived from Sentinel-2 satellite imagery. Temporal and spatial variation in contamination from 2011 to 2022 was analyzed using trend analysis and random forest–based inversion modeling. Health risks were quantified following United States EPA guidelines, with hazard quotients and carcinogenic risks calculated for oral ingestion, dermal contact, and inhalation pathways in both children and adults. The framework concludes with the translation of scientific results into management strategies aimed at mitigating contamination impacts, supporting sustainable regional development, rural revitalization, and productivity enhancement.

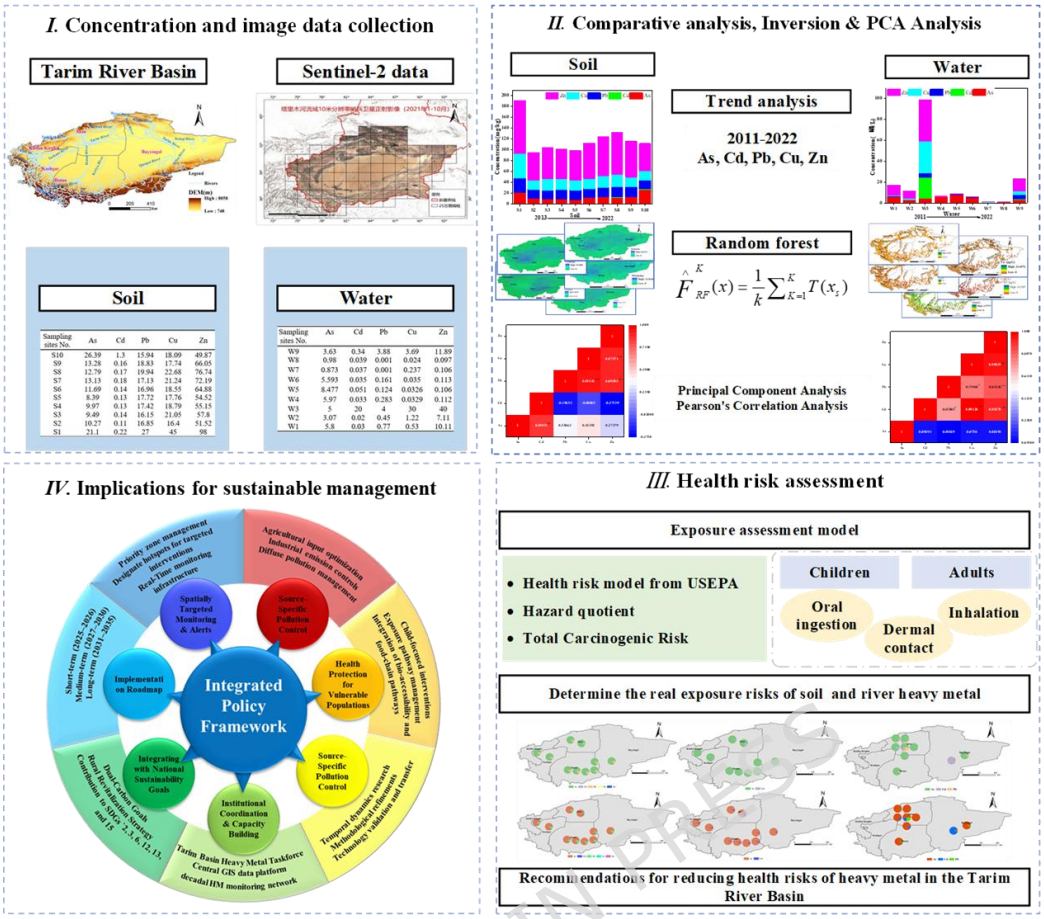


Fig. 2. Framework representing the overview of the research. Created by the authors using Microsoft PowerPoint (Microsoft Office LTSC Professional Plus 2021; https://www.microsoft.com/en-us/microsoft-365/powerpoint).

2.3.1 Remote sensing inversion scheme

To quantify and map heavy metal contamination (As, Cd, Pb, Cu, Zn) across the heterogeneous landscapes of the Tarim River Basin, a machine learning–based remote sensing inversion framework was developed. This framework integrates multi-spectral Sentinel-2 imagery with field-measured concentrations of heavy metals to generate spatially explicit prediction maps.

The methodological scheme comprises four sequential stages (Fig. 3). This systematic design enables robust, data-driven environmental monitoring in arid basins through intelligent coupling of in situ and remote sensing data.

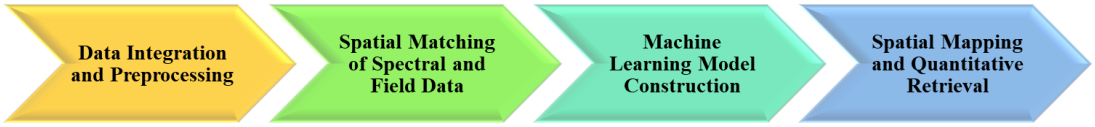


Fig. 3. Flowchart of the remote sensing inversion methodology.

(1) Data Integration and Preprocessing

Spatial Data Preparation

Sentinel-2 multispectral imagery (January–October 2021, cloud cover < 20%) was acquired from the Copernicus Open Access Hub[54]. Four spectral bands—Blue (B2), Green (B3), Red (B4), and Near-Infrared (B8)—were selected for their sensitivity to soil and water quality variability.

Auxiliary datasets included:

- Topographic maps (1:50,000; National Geomatics Center of China, 2020 edition)
- Digital Elevation Models (DEMs) (1:50,000; via Geospatial Data Cloud)
- Field-surveyed ground control points (GCPs) (see Supplementary Table S1–2).

All datasets were harmonized within the CGCS2000 / UTM Zone 44N coordinate system to ensure consistent spatial referencing.

Preprocessing Workflow

To ensure quality and reproducibility, image preprocessing followed the Chinese national standard GB/Z 1044-2018 ("Technical Specification for Quality Inspection of Optical Satellite Remote Sensing Images"). The workflow included:

- Radiometric and atmospheric correction: Conversion of raw imagery to surface reflectance products using the Sen2Cor processor to minimize illumination and inter-scene radiometric bias[55].
- Geometric correction and orthorectification: Alignment using GCPs and DEM data to eliminate terrain distortion. Corrected scenes were mosaicked and spectrally normalized via histogram matching.
- Resampling and clipping: The mosaicked image was clipped to the Tarim River Basin boundary and resampled to a uniform 60-m resolution using ENVI 5.6, ensuring spectral integrity and geometric precision (planimetric error < 1 pixel).

The final output was a spectrally consistent and geometrically precise orthomosaic, serving as the analytical base for heavy metal inversion.

(2) Spatial Matching of Spectral and Field Data

To link spectral reflectance to measured metal concentrations, reflectance data from the processed Sentinel-2 imagery were extracted at each sampling site (Supplementary Table S1–2)[56].

Key steps included:

- Extracting spectral values within a 3 × 3 pixel window centered on each sampling coordinate to minimize geolocation error and sensor noise.
- Calculating mean reflectance values for the Blue, Green, Red, and NIR bands.
- Applying bilinear interpolation for sites located between pixels.
- Excluding cloud-contaminated pixels to avoid reflectance distortion.

This procedure produced a paired dataset of 19 sampling sites (10 soil/sediment + 9 water), where each record combined:

- Mean heavy metal concentrations (As, Cd, Pb, Cu, Zn)
- Corresponding Sentinel-2 spectral features[57]

The physical rationale is that heavy metal contamination alters reflectance patterns through changes in mineralogy, sediment color, and vegetation health. All operations were conducted in ENVI 5.6 within the unified CGCS2000 / UTM Zone 44N reference system.

(3) Machine Learning Model Construction

Model Training

A Random Forest (RF) regression model was implemented in Python 3.10 using the sklearn.ensemble library to establish quantitative relationships between Sentinel-2 spectral variables and heavy metal concentrations.

Rationale for RF Selection (Supplementary Table S4):

- Handles nonlinear interactions between variables
- Remains robust despite small sample sizes and multicollinearity
- Provides feature importance metrics and built-in estimates of prediction uncertainty

Training setup:

- 19 sampling records divided into training (70%) and validation (30%) subsets (Table 2[58])
- 10-fold cross-validation to prevent overfitting
- Input variables: raw reflectance from B2–B8 bands
- Target variables: concentrations (mg/kg for soil; μg/L for water)

Hyperparameter optimization: Conducted using grid search and cross-validation, tuning:

- Number of trees (100–500)
- Maximum depth (10–30)
- Minimum leaf size (1–5)[59, 60]

Feature importance was assessed by permutation analysis (Fig. 4). The results indicated:

- Blue → highest relevance for As
- Green → Cd
- Red → Cu
- NIR → Pb and Zn

These findings confirm the spectral diagnostic sensitivity of specific Sentinel-2 bands to different contamination mechanisms across the Tarim River Basin (Supplementary Table S4). The trained forest with K trees can be expressed as:

$$f(x) = \frac{1}{K} \sum_{k=1}^{K} T_k(x_k) \quad (1)$$

where $T_k(x_k)$ is the prediction from the k-th individual tree, x is the input feature vector, and x_k is the bootstrapped training subsample for the k-th tree. Another key parameter (not explicitly shown in Eq. (1)) is the number of candidate features randomly selected at each split (often denoted as mtry).

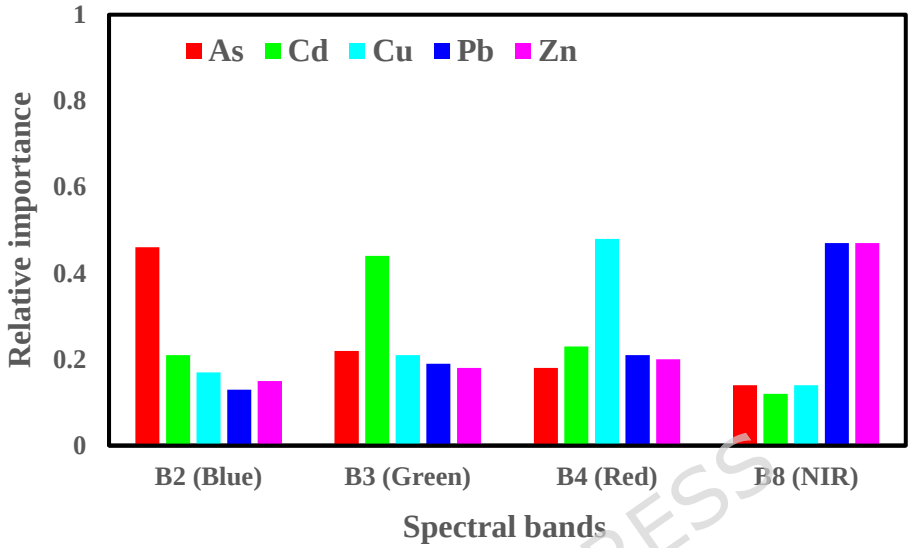


Fig. 4. Variable importance of sentinel-2 spectral bands for heavy metal inversion.

Relative importance of Sentinel-2 spectral bands (B2–Blue, B3–Green, B4–Red, and B8–Near-Infrared) in predicting concentrations of five heavy metals (As, Cd, Cu, Pb, and Zn) using Random Forest regression. Importance values were derived from permutation analysis and normalized to a consistent 0–1 scale across all metals, enabling direct visual comparison. Results indicate dominant contributions from the Near-Infrared band for Pb and Zn, the Red band for Cu, the Green band for Cd, and the Blue band for As, reflecting element-specific spectral sensitivities linked to mineralogical characteristics and vegetation-mediated reflectance variations in the Tarim River Basin.

Model Validation

To ensure statistical robustness and practical reliability, model outputs were evaluated using independent validation data (30% hold-out set). Input variables were standardized via Min–Max normalization. Performance metrics included:

- Coefficient of determination (R^2) – model fit quality[61]
- Root Mean Square Error (RMSE) – average deviation magnitude[62].
- Mean Absolute Error (MAE) – average absolute deviation[63].

Interpretation thresholds[64]:

- $R^2 < 0.5$: unsatisfactory
- $0.5 \leqslant R^2 < 0.65$: fair
- $R^2 \geqslant 0.65$: good predictive performance

Additionally, geostatistical cross-validation and semivariogram analysis verified spatial consistency between predicted and observed values. Internal variance calculations identified zones of high prediction uncertainty. Together, these assessments ensured both accuracy and environmental interpretability of the inversion outputs.

The calculation formulae for MAE, RMSE, and R^2 are given below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_{obs,i} - y_{pre,i} | \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{pre,i})^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{obs,i} - y_{pre,i})^2}{\sum_{i=1}^{n}(y_{obs,i} - \bar{y}_{obs})^2} \tag{4}$$

where, n is the total number of data in the dataset, $y_{obs,i}$ and $y_{pre,i}$ represent the observed and predicted values for the i-th sample, respectively, $\bar{y}_{obs}$ denotes the average of the observed values.

Table 2. Calibration and validation datasets for inversion using the RF algorithm.

	As	Cd	Pb	Cu	Cd
Training set (S)	S1, S2, S4, S5, S7, S9, S10	S1, S2, S4, S5, S7, S8, S10	S1, S2, S3, S5, S6, S8, S9	S1, S2, S3, S6, S8, S9, S10	S1, S2, S5, S7, S8, S9, S10
Validation set (S)	S3, S6, S8	S3, S6, S9	S4, S7, S10	S4, S5, S7	S3, S4, S6
Training set (W)	W2, W3, W4, W5, W6, W8	W1, W2, W3, W6, W7, W9	W1, W2, W3, W7, W8, W9	W1, W2, W3, W6, W8, W9	W1, W2, W3, W4, W7, W9
Validation set (W)	W1, W7, W9	W4, W5, W8	W4, W5, W6	W4, W5, W7	W5, W6, W8

Note: S1-S10 and W1-W9 are the sampling points of different soils and different rivers in the Tarim River Basin, respectively.

(4) Spatial Mapping and Quantitative Retrieval

The trained RF models were applied pixel-by-pixel to the processed Sentinel-2 mosaics to generate continuous maps of heavy metal concentrations across the Tarim River Basin at 60-m spatial resolution. Separate simulations were conducted for soil/sediment and aquatic environments.

Post-processing steps included:

- Median filtering (5 × 5 kernel) to reduce random noise
- Removal of implausible outliers exceeding the 99th percentile of training data
- Masking of non-target surfaces (e.g., snow and large open water bodies) using land-use/land-cover data.

Prediction uncertainty was quantified using RF ensemble variance, with higher confidence observed in oasis and riparian zones where spectral–field correspondence was strongest. Independent field samples were used for post-validation to confirm the robustness of spatial patterns.

The resulting high-resolution maps enable visualization of contamination gradients and identification of pollution hotspots. The 60 m resolution corresponds to the Sentinel-2A input data, and the mapped distributions represent model-predicted continuous concentration surfaces rather than raw observations. Overall, this machine learning–based inversion framework provides a reproducible and scalable approach for regional heavy metal assessment in arid river basins, supporting evidence-based environmental management and remediation planning in the Tarim River Basin.

2.3.2 Health risk assessment

To evaluate heavy metal pollution in the Tarim River Basin within a comprehensive assessment framework, this study incorporated health risk assessment into the analysis of heavy metal spatiotemporal distribution. The human exposure risk model developed by the US Environmental Protection Agency was employed to quantify both non-carcinogenic and

carcinogenic risks associated with Pb, Cd, Cu, Zn, and As[65, 66]. Non-carcinogenic risk was assessed for all five metals, whereas carcinogenic risk was evaluated for Cd and As[67]. Three primary exposure pathways—ingestion, dermal contact, and inhalation—were considered, and the average daily dose (ADD, $mg \cdot kg^{-1} \cdot d^{-1}$) was estimated separately for adults and children[65, 66]. The hazard quotient (HQ) for each metal and pathway was calculated, and the hazard index (HI) was obtained as the sum of HQs[68]. Carcinogenic risk was quantified as the total carcinogenic risk (TCR) based on the reference dose (RfD) and slope factor (SF) for Cd and As. The equations used for ADD, HQ, HI, and TCR followed standard US EPA methodologies[69].

$$ADD_{ingestion} = \frac{C \times IngR \times CF \times EF \times ED}{BW \times AT} \quad (5)$$

$$ADD_{inhalation} = \frac{C \times InhR \times EF \times ED}{BW \times AT \times PEF} \quad (6)$$

$$ADD_{dermal} = \frac{C \times CF \times SA \times ABS \times EF \times ED}{BW \times AT} \quad (7)$$

$$HI = \sum_{i=1}^{n} HQ_i = \sum_{i=1}^{n} \left(\frac{ADD_{ingestion}}{RfD} + \frac{ADD_{inhalation}}{RfD} + \frac{ADD_{dermal}}{RfD}\right)_i \quad (8)$$

$$TCR = \sum_{i=1}^{n} CR_i \quad (9)$$

$$= \sum_{i=1}^{n} \left(ADD_{ingestion} \times SF + ADD_{inhalation} \times SF + ADD_{dermal} \times SF\right)_i \quad (10)$$

When HI < 1, the non-carcinogenic risk is considered negligible[70]. Conversely, HI > 1 indicates a non-negligible non-carcinogenic risk. TCR values within the range of 10^{-6} to 10^{-4} are deemed acceptable thresholds, while exceeding this range suggests a carcinogenic risk[68].

The exposure dose for direct ingestion ($ADD_{ingestion}$) and carcinogenic risk (CR) were calculated as:

$$ADD_{ingestion} = \frac{C_w \times IR \times EF \times ED}{BW \times AT} \quad (11)$$

$$CR = ADD_{ingestion} \times CSF \quad (12)$$

where $ADD_{ingestion}$ are the average daily exposure doses contacted through ingestion of water; C_w is the average concentration of the studied element in water (μg/L); IR is the ingestion rate of water (L/day), in this study, 0.78 L/day for children and 2.5 L/day for adults; EF is the exposure frequency (days/year), in this study, 350 days/year; ED is the exposure duration (years), in this study, 6 years for children and 26 years for adults; BW is the body weight (kg),

in this study, 15 kg for children and 70 kg for adults; AT is the averaging time (days), in this study, 2190 days for children and 9490 days for adults; CSF is the cancer slope factor of a carcinogen (μg/kg/ day)$^{-1}$, which is based on risk-based concentration table by the USEPA (2004). The Ministry of Environmental Protection (MEP) of the People's Republic of China (2014) considers a CR lower than 1.0 × 10^{-6} to be acceptable, while the USEPA (2004) recommends an acceptable or tolerable range of carcinogenic risks from 10^{-6} to 10^{-4}[71].

Cumulative effects. The degree of heavy metal contamination was quantified using the geoaccumulation index (I_{geo}), which has been widely applied in the assessment of soil and sediment pollution. This index incorporates the influence of natural diagenetic processes, thereby improving the reliability of pollution evaluation. The I_{geo} was calculated according to Equation[72]:

$$I_{geo} = \log_2\left(\frac{C_n}{1.5 \times B_n}\right) \tag{13}$$

where C_n represents the measured concentration of heavy metal n in the soil (mg kg^{-1}), B_n denotes the corresponding geochemical background concentration (mg kg^{-1}), and the constant 1.5 is a correction factor to account for potential variations in baseline values[73]. Based on established classification criteria, I_{geo} values are interpreted as follows: I_{geo} <0, unpolluted; 0< I_{geo} <1, unpolluted to moderately polluted; 1< I_{geo} <2, moderately polluted; 5< I_{geo}, very highly polluted.

The Heavy Metal Pollution Index (HPI) was applied to quantify the contamination status of surface water and groundwater in the Tarim River Basin. This index integrates the concentrations of multiple heavy metals, including As, Cd, Pb, Cu, and Zn, into a single value that reflects overall water quality. The HPI is calculated using the subindex of the i-th parameter (Q_i), its corresponding unit weight (W_i), and the total number of chemical parameters considered (n). Threshold values were adopted to classify water suitability: HPI < 100 indicates water suitable for drinking purposes, whereas HPI > 100 denotes water unsuitable for consumption. This method provides a systematic approach for assessing freshwater quality and serves as a basis for developing strategies to reduce the risks associated with heavy metal exposure[74].

$$HPI = \sum_{i=1}^{n} W_i Q_i \tag{14}$$

To ensure consistency and clarity across all equations, the following table defines the key variables and symbols used throughout the manuscript (Table 3):

Table 3. Nomenclature Table

Symbol	Description	Units
$f(x)$	Ensemble prediction from the Random Forest model	-
K	Number of trees in the Random Forest	-
$T_k(x_k)$	Prediction from the k-th tree	-
x	Input feature vector	-
x_k	Bootstrapped training subsample for the k-th tree	-
n	Total number of data points/samples	-
$y_{obs,i}$	Observed value for the i-th sample	(e.g., concentration units)
$y_{pre,i}$	Predicted value for the i-th sample	(e.g., concentration units)
$\bar{y}_{obs}$	Mean of observed values	(e.g., concentration units)
ADD	Average daily dose	$mg \cdot kg^{-1} \cdot d^{-1}$
$ADD_{ingestion}$	Average daily dose via ingestion	$mg \cdot kg^{-1} \cdot d^{-1}$
HQ	Hazard quotient for a single pathway/metal	-
HI	Hazard index (sum of HQs)	-
TCR	Total carcinogenic risk	-
CR	Carcinogenic risk for a specific pathway	-
RfD	Reference dose	$mg \cdot kg^{-1} \cdot d^{-1}$
SF	Slope factor	-
CSF	Cancer slope factor	$(\mu g/kg/ day)^{-1}$
C_w	Concentration of heavy metal in water	µg/L
IR	Ingestion rate	L/day
EF	Exposure frequency	days/year
ED	Exposure duration	years
BW	Body weight	kg
AT	Averaging time	days
I_{geo}	Geoaccumulation index	-
C_n	Measured concentration of heavy metal n	mg kg^{-1}
B_n	Geochemical background concentration of heavy metal n	mg kg^{-1}
HPI	Heavy Metal Pollution Index	-
Q_i	Quality subindex for the i-th parameter	-
W_i	Unit weight for the i-th parameter	-

2.3.3 Statistical analysis

All concentration data of heavy metals (Pb, Cu, Zn, Cd, and As) from soil (n = 50) and river water samples (n = 45) were examined for distributional characteristics prior to analysis. Data were log-transformed where appropriate to mitigate skewness and approximate normality. The Shapiro–Wilk test was used to assess normality and the Levene's test for homogeneity of

variance. Recognizing that the Shapiro–Wilk test has limited statistical power for small sample sizes (e.g., n = 19 in site-level subsets), the results were interpreted in conjunction with visual inspections of Q–Q plots and density distributions. When distributional assumptions were not met, non-parametric alternatives (Spearman's rank correlation) were considered to validate correlation trends.

Outliers were identified using the interquartile range (IQR) criterion (values exceeding Q3 + 1.5×IQR or below Q1 – 1.5×IQR). Outliers were retained to preserve the full extent of natural environmental variability and potential contamination hotspots[75], which are central to the study's objective of characterizing heterogeneous anthropogenic inputs. While the model performance with and without outliers could not be demonstrated in supplementary materials, sensitivity analyses were performed internally and confirmed that the inclusion of these data did not alter the main structure or interpretation of the principal component analysis (PCA) results. This approach ensures that local anomalies relevant to pollution assessment remain represented in the dataset.

Missing values (<5% of the dataset) were imputed using mean substitution for approximately normally distributed variables and median substitution for skewed distributions to maintain dataset completeness without biasing structural relationships.

To explore potential sources and migration pathways of heavy metals[76, 77], Pearson's correlation analysis and Principal Component Analysis (PCA) were applied to the processed datasets from soil and river water samples. Pearson's correlation coefficients (r) quantified the strength and direction of associations between metal concentrations, with two-tailed significance thresholds set at $p < 0.05$ and $p < 0.01$. For non-parametric verification, Spearman's rank correlations were computed where normality could not be confidently assumed. Correlations were categorized as strong ($|r| \geqslant 0.8$), moderate ($0.5 \leqslant |r| < 0.8$), and weak ($|r| < 0.5$).

PCA was employed to discern potential source factors and reduce dataset dimensionality. Components with eigenvalues greater than 1.0 were retained following the Kaiser criterion[78], and Varimax rotation was applied to enhance interpretability[79]. The combination of correlation analysis and PCA provided complementary insights into potential co-contamination patterns and geochemical associations of heavy metals.

All statistical analyses were performed using IBM SPSS Statistics 27 (IBM Corp., Armonk, NY, USA), with graphical visualizations in Origin 2022 (OriginLab Corporation, Northampton, MA, USA). All analyses were conducted at a 95% confidence level ($\alpha = 0.05$) unless otherwise specified. Eigenvalues, factor loadings, and correlation coefficients are reported with significance levels for transparency and replicability. Results were cross-validated against literature-derived geochemical background values and regional standards[68,

76, 80, 81], supporting the reliability of source interpretations with outliers retained. The combined use of correlation analysis and PCA provides a robust multivariate framework for heavy metal source apportionment and migration assessment.

3. Results

3.1 Temporal variations

The total concentrations of heavy metals (ΣHMs) in soil samples displayed pronounced spatial heterogeneity among the sampling sites, with values ranging from 95.15 to 191.32 mg/kg (mean: 118.71 mg/kg). The maximum soil ΣHMs was detected at site S1 (191.32 mg/kg), with elevated levels also found at S8 (132.32 mg/kg) and S7 (123.87 mg/kg). For river water samples, ΣHMs ranged from 1.14 to 99.00 μg/L (mean: 3.89 μg/L), the highest concentration being observed at site W3 (99.00 μg/L), followed by W9 (23.43 μg/L). In both matrices, Zn was identified as the predominant heavy metal, constituting the largest proportion of ΣHMs across all sampling locations (Fig. 5a, 5b).

The mean concentrations of individual heavy metals in soil were as follows: Zn 64.67 mg/kg, Cu 21.73 mg/kg, Pb 18.39 mg/kg, As 13.65 mg/kg, and Cd 0.27 mg/kg. In river water, the mean concentrations were 7.74 μg/L for Zn, 4.38 μg/L for As, 3.98 μg/L for Cu, 2.29 μg/L for Cd, and 1.07 μg/L for Pb (see Supplementary Table S1, Supplementary Table S2).

Comparison with the regional soil background values specific to Xinjiang and the Environmental Quality Standards for Surface Water (refer to Table 3) revealed that concentrations of As, Cd, Pb, Cu, and Zn in all river water samples were below the corresponding regulatory thresholds. However, in soil samples, the proportion of sites with concentrations lower than the regional background values were 40% for As, 10% for Cd, 80% for Pb, 90% for Cu, and 70% for Zn.

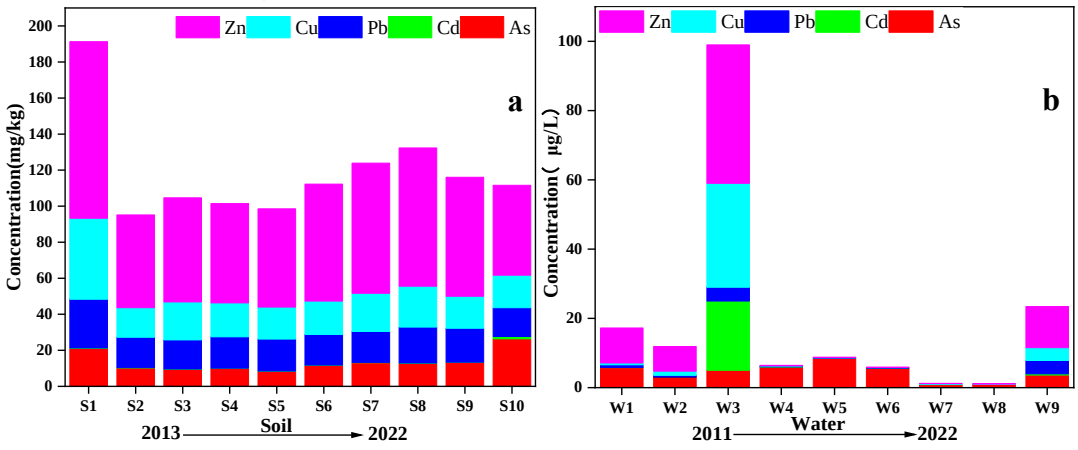


Fig. 5. Heavy metal content in soil samples of different sampling sites (a) and rivers (b).

The descending order of mean concentrations of heavy metals in soil was Zn > Cu > Pb > As > Cd. While Zn, Cu, and Pb levels in soil were within safe limits, As and Cd exceeded regional background values at a substantial number of sites. In river water, the order of

prevalence was Zn > As > Cu > Cd > Pb, with all concentrations falling within the permissible limits established for surface water.

3.2 Spatial patterns

3.2.1 Model performance and spatial prediction uncertainty

The inversion results show substantial variation in model performance among individual heavy metals and between soil and river environments (Tables 6 and 7). For soil datasets, the model reproduced Cd concentrations with the smallest residual errors (MAE = 0.007 mg/kg; RMSE = 0.007 mg/kg), and the validation R^2 (0.997) indicates an almost exact correspondence between observed and predicted values. Cu also achieved relatively strong performance (R^2 = 0.889; RMSE = 0.955 mg/kg). In contrast, Pb and Zn yielded moderate agreement (R^2 = 0.601 and 0.599, respectively), and model fidelity for As remained limited (R^2 = 0.408), with error magnitudes roughly double those of the better-performing metals (Table 4).

A similar pattern was observed for river samples. Cd reached the highest accuracy (R^2 = 0.999; RMSE = 0.024 μg/L), followed by Cu (R^2 = 0.958) and Pb (R^2 = 0.992). Zn predictions remained acceptable (R^2 = 0.897) but with higher absolute error values. The weakest performance again occurred for As (R^2 = 0.320), showing that spectral information alone explains only a modest portion of its variability (Table 4).

These metrics demonstrate that the inversion scheme captures spatial gradients for several target metals, particularly Cd, Cu, and Pb. However, the extremely high R^2 values for Cd, especially given the small number of training observations, likely reflect partial overfitting. These results should therefore be regarded as preliminary spatial patterns rather than fully generalizable predictions for basin-scale applications.

Table 4. Modeling accuracy of heavy metal concentrations using Random Forest algorithms in the inversion.

Evaluation indicator	Medium	As	Cd	Cu	Pb	Zn
MAE	Soil (mg/kg)	1.555	0.007	0.804	0.992	4.150
	Water (μg/L)	1.589	0.018	0.784	0.064	2.220
RMSE	Soil (mg/kg)	2.078	0.007	0.955	1.068	4.291
	Water (μg/L)	1.834	0.024	0.785	0.078	2.443
R^2	Soil	0.408	0.997	0.889	0.601	0.599
	Water	0.320	0.999	0.958	0.992	0.897

3.2.2 Soil

The spatial distribution of As concentrations across the study region was generally homogeneous, with measured values spanning from 0.001 to 22.951 $mg \cdot kg^{-1}$ (Fig. 6a). The majority of the area (57.86%) exhibited low As concentrations (<12.331 $mg \cdot kg^{-1}$) (Table 5). Notably, higher concentrations were localized in eastern Hotan and western Bayyingol, where values approached the upper range of observations.

Cd concentrations ranged from 0.001 to 0.825 mg·kg⁻¹, with elevated levels primarily detected at the basin margins, especially in eastern Hotan and western Bayyingol (Fig. 6b). Additional Cd hotspots were observed in localized areas of Aksu; however, the central basin generally displayed minimal Cd levels, with approximately 64.28% of the area falling within the 0.001–0.184 mg·kg⁻¹ range (Table 5).

Cu concentrations varied between 0.001 and 36.200 mg·kg⁻¹, with the highest levels detected in the central part of the basin—most prominently in northwestern Hotan and eastern Kashgar (Fig. 6c). The marginal zones of the basin predominantly exhibited low Cu concentrations, and 73.27% of the study area was characterized by values within the 0.001–22.570 mg·kg⁻¹ interval (Table 5).

Pb was distributed relatively uniformly across the basin, with concentrations ranging from 0.001 to 23.836 mg·kg⁻¹ (Fig. 6d). Although general Pb levels remained low, localized hotspots were apparent in eastern Bayyingol and western Hotan. Approximately 51.63% of the region recorded Pb concentrations between 0.001 and 18.601 mg·kg⁻¹ (Table 5).

Zn concentrations were generally low and spatially uniform in the marginal zones of the Tarim River Basin, with values from 0.001 to 84.411 mg·kg⁻¹ (Fig. 6e). Elevated Zn concentrations were confined to areas near Kashgar, western Hotan, and eastern Bayyingol. Overall, 61.07% of the basin had Zn levels within the 0.001–63.887 mg·kg⁻¹ range (Table 5).

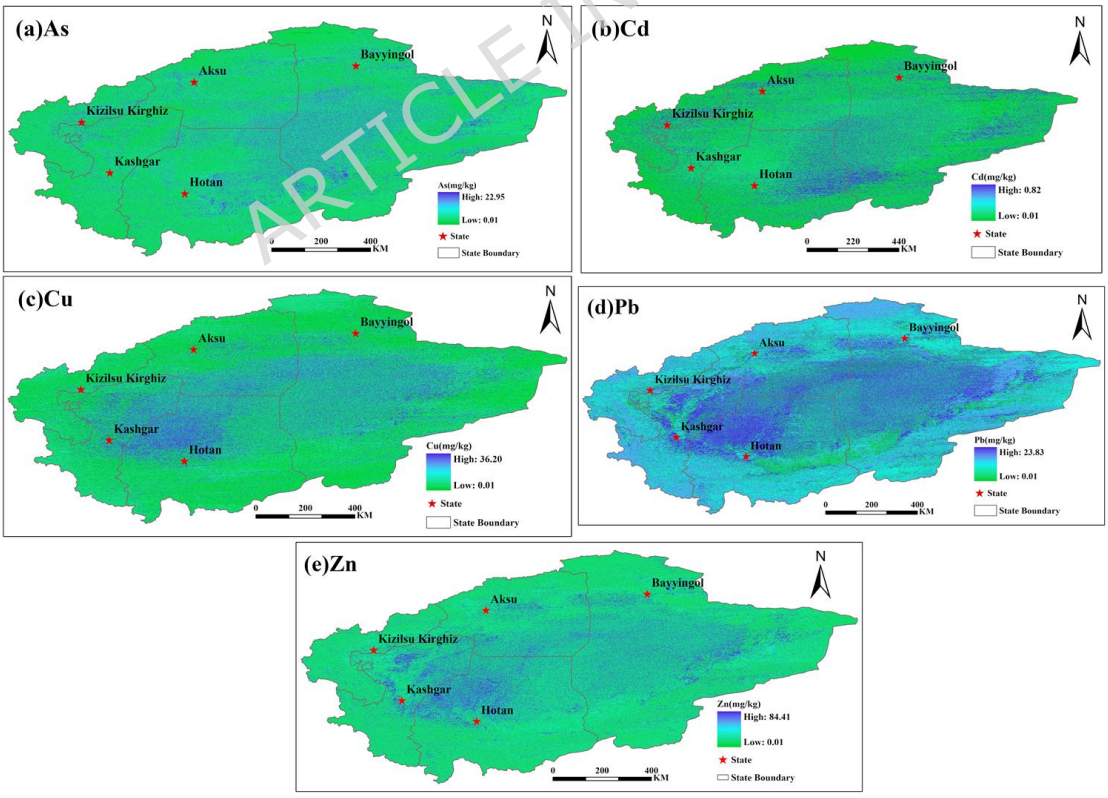


Fig. 6. Spatial distribution of soil heavy metal in the Tarim River Basin based on the Sentinel-2A images. Note: Sampling point locations are shown in Figure 1, and are not repeated in

Figures 6 to avoid excessive redundancy and visual clutter. Created by the authors using ArcGIS Pro 3.6 (Esri; https://www.esri.com/en-us/arcgis/about-arcgis/overview)

A comparative analysis against regional soil geochemical background values for Xinjiang (Table 1;[41, 53]) indicated substantial spatial variability in exceedance rates among the five assessed heavy metals (Table 5). As detailed in Table 3, Pb, Cu, and Zn surpassed background concentrations in 48.36%, 15.52%, and 38.93% of samples, respectively, suggesting moderate enrichment. In contrast, As and Cd exhibited markedly higher exceedance rates, with 65.90% and 100% of soil samples exceeding their respective background levels, reflecting significant contamination. These patterns underscore the influence of anthropogenic activities across the basin. Collectively, the findings presented here highlight the necessity for targeted source identification and pollution control strategies to mitigate the risks of heavy metal contamination in the Tarim River Basin, as further examined in Sections 3.3 and 4.

Table 5. Statistics for estimates of soil heavy metal contents

heavy metal	Contents/(mg/kg)	Percentage/%
As	0.001-11.161	34.10
	11.162-12.331	23.76
	12.332-13.501	27.43
	13.502-18.001	1.33
	18.002-22.951	13.38
Cd	0.001-0.145	22.70
	0.146-0.184	41.58
	0.185-0.255	17.86
	0.256-0.430	9.71
	0.431-0.825	8.15
Cu	0.001-19.306	32.67
	19.307-21.294	12.04
	21.295-22.570	28.56
	22.571-25.410	11.20
	25.411-36.200	15.52
Pb	0.001-16.919	19.85
	16.920-17.480	4.17
	17.481-18.601	27.61
	18.602-20.097	23.47
	20.098-23.836	24.89
Zn	0.001-59.253	7.44
	59.253-60.908	25.10
	60.909-63.887	28.53
	63.888-72.825	25.29
	72.826-84.411	13.64

3.2.3 River

The spatial distribution of heavy metal concentrations within the river systems demonstrated pronounced geographical variability. As concentrations ranged from 0.001 to 6.979 μg/L, with elevated levels predominantly observed in the western and central rivers, including the Yarkant, Karakash, Baiyu, and sections of the Tarim River (Fig. 7a). Moderate

As concentrations were detected in the Konqi and Qarqan Rivers, while lower values were recorded in northeastern and southwestern rivers such as the Kaidu River. Notably, 67.11% of the surveyed river area exhibited As concentrations within the range of 0.001–4.023 μg/L (Table 6).

Cd concentrations spanned 0.001 to 14.210 μg/L, with peak levels concentrated in upstream and western rivers, particularly the Hotan, Kashgar, and Yarkant Rivers (Fig. 7b). Conversely, the Qarqan River and other eastern tributaries showed relatively low Cd levels. Overall, 73.68% of the river network maintain Cd concentrations between 0.001 and 0.111 μg/L (Table 6).

Cu exhibited considerable spatial heterogeneity, with concentrations ranging from 0.001 to 21.497 μg/L (Fig. 7c). While most river segments contained low Cu levels, elevated concentrations were identified in the upstream reaches of the Kashgar and Aksu Rivers, mid-sections of the Hotan River, and various western tributaries. Approximately 55.31% of the river area displayed Cu concentrations between 0.675 and 1.686 μg/L (Table 6).

Pb concentrations varied from 0.001 to 2.94 μg/L, with the highest values found in southern and southwestern rivers, including Kashgar, Yarkant, Hotan, and Tashkorgan Rivers (Fig. 7d). Elevated Pb levels were also observed in western tributaries, whereas northeastern and central rivers such as the Konqi River exhibited comparatively low concentrations. Pb concentrations between 0.001 and 0.647 μg/L accounted for 63.38% of the river area (Table 6).

Zn concentrations ranged from 0.001 to 29.997 μg/L, with higher concentrations primarily distributed in the basin's eastern regions, notably near the Konqi River. Additional sporadic hotspots were detected in upstream tributaries such as the Kaidu and Muzat Rivers (Fig. 7e). Despite these localized elevations, Zn levels were generally low across the basin, with 55.75% of the river area presenting concentrations between 0.001 and 5.176 μg/L (Table 6).

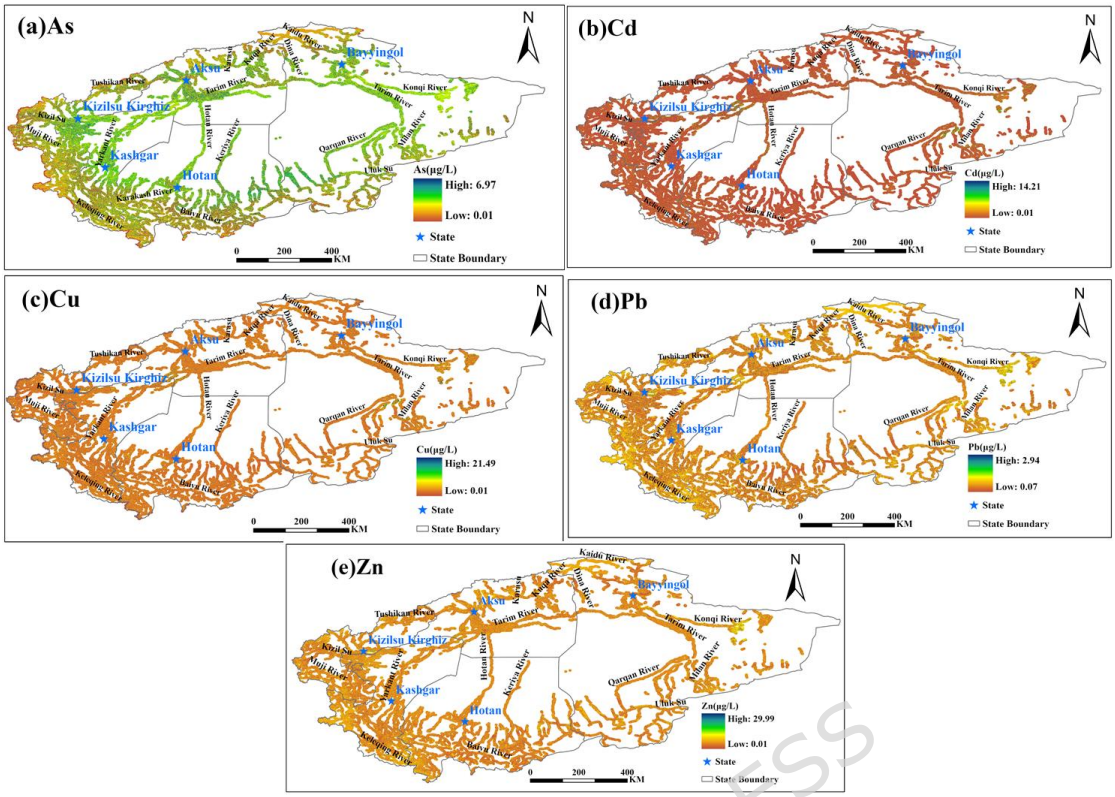


Fig. 7. Spatial distribution of river heavy metal in the Tarim River Basin based on the Sentinel-2A images. Note: Sampling point locations are shown in Figure 1, and are not repeated in Figures 7 to avoid excessive redundancy and visual clutter. Created by the authors using ArcGIS Pro 3.6 (Esri; https://www.esri.com/en-us/arcgis/about-arcgis/overview)

Building on the background values summarized in Table 3 (Section 2.2.1), comparison with the Environmental Quality Standards for Surface Water[49] revealed that As, Cu, Pb, and Zn concentrations across all monitored rivers largely complied with regulatory limits. In contrast, Cd concentrations exceeded Class III and Class V thresholds in 4.52% and 1.47% of river sections, respectively, highlighting localized contamination hotspots that warrant targeted management interventions.

Table 6. Statistics for estimates of river heavy metal contents

heavy metal	Contents/(μg/kg)	Percentage/%
As	0.001-1.642	16.68
	1.643-2.928	30.81
	2.929-4.023	19.62
	4.024-5.364	18.93
	5.365-6.979	13.96
Cd	0.001- 0.111	**73.68**
	0.112- 0.390	2.78
	0.391- 0.835	17.54
	0.836 - 5.795	4.52
	5.796 -14.210	1.47
Cu	0.001- 0.674	24.71
	0.675 - 1.686	**55.31**
	1.687 - 4.130	15.88
	4.130 - 9.694	3.37
	9.694 - 21.497	0.73
Pb	0.001- 0.231	17.94
	0.232 - 0.400	24.21
	0.401 - 0.647	21.23
	0.648 - 1.411	33.29
	1.412 - 2.94	3.34
Zn	0.001- 2.117	13.47
	2.118 - 5.176	42.28
	5.177 - 8.705	30.76
	8.706 - 15.998	12.56
	15.999 - 29.997	0.73

3.3. Identification of heavy metal sources

3.3.1 Principal Component Analysis

The results of the principal component analysis (PCA) for the soil dataset are summarized in Table 7. Two principal components (PCs) with eigenvalues exceeding 1.0 were extracted, explaining a cumulative 96.07% of the total variance. The first principal component (PC1) accounted for the majority of the variance (59.75%) and was strongly correlated with Pb (0.975), Cu (0.963), and Zn (0.959). The second principal component (PC2), which accounted for an additional 36.32% of the variance, was dominated by high loadings of Cd (0.971) and As (0.939).

In contrast, the PCA for the river sediment data yielded only one principal component with an eigenvalue > 1.0, which explained 71.84% of the total variance (Table 8). This component was characterized by high positive loadings from four heavy metals: Zn (0.985), Cu (0.982), Cd (0.958), and Pb (0.853).

Table 7 Loading of the varimax-rotated factorial matrix of principal components

Variables	PC1	PC2
As	0.330	**0.939**
Cd	-0.225	**0.971**
Pb	**0.975**	0.022
Cu	**0.963**	0.137
Zn	**0.959**	-0.060
Eigenvalue	2.988	1.816
Variability (%)	59.752	36.322
Cumulative (%)	59.752	96.074

Table 8 Loading of the varimax-rotated factorial matrix of principal components

Variables	PC1
As	0.106
Cd	**0.958**
Pb	**0.853**
Cu	**0.982**
Zn	**0.985**
Eigenvalue	3.592
Variability (%)	71.835
Cumulative (%)	71.835

3.3.2 Pearson's correlation analysis

Pearson's correlation analysis was conducted to elucidate potential sources and geochemical behaviors of heavy metals in the Tarim River Basin[80]. In soil samples, statistically significant positive correlations ($p < 0.01$) were observed among several heavy metal pairs, including Pb–Cu ($r = 0.933$), Pb–Zn ($r = 0.893$), Cu–Zn ($r = 0.876$), and As–Cd ($r = 0.830$) (Fig. 8a). These strong correlations suggest that Pb, Cu, Zn, and As–Cd likely share common sources and exhibit similar geochemical behaviors in soils. Additionally, a highly significant correlation between Pb and As ($p < 0.001$) indicates a homogeneous source, potentially linked to shared anthropogenic or natural processes[82-85].

In river water samples, strong positive correlations were identified for Cd–Zn ($r = 0.933$), Pb–Zn ($r = 0.832$), Cu–Zn ($r = 0.958$), and Cu–Cd ($r = 0.994$), all at $p < 0.01$ (Fig. 8b). Notably, Cu – Pb ($r = 0.750$, $p \leqslant 0.05$) and Cd – Pb ($r = 0.679$, $p \leqslant 0.05$) also exhibited significant correlations, suggesting shared environmental sources and distribution patterns. These associations imply that Cu, Pb, Cd, and Zn may undergo similar migration and deposition

processes, potentially adsorbing onto organic and mineral components in the river system, facilitating joint transport[82].

Principal component analysis (PCA), complemented by Pearson's correlation analysis, further delineated distinct heavy metal source and migration patterns. In soils, the primary source pathway was characterized by Pb–Cu–Zn, while a secondary pathway was dominated by As–Cd. In contrast, river systems exhibited a predominant pathway involving Zn–Cu–Cd–Pb. These findings highlight the differential geochemical behaviors and potential sources of heavy metals in soil and river matrices within the Tarim River Basin.

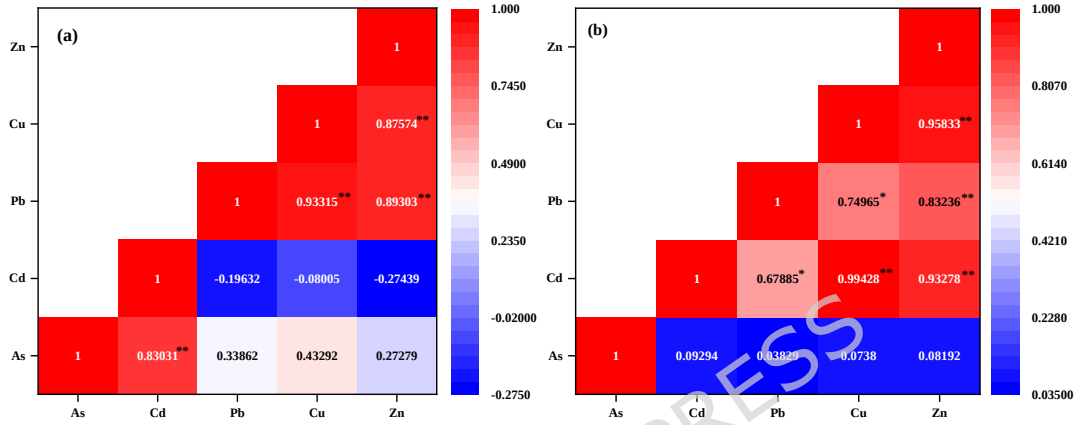


Fig. 8. Pearson correlation analysis between heavy metals. (a) soil; (b)river

3.3.3 Source apportionment results

Soil samples from the Xinjiang region revealed concentrations of Pb, Cu, and Zn below the established geochemical background values for the area[41, 53] (Supplementary Table S1). These findings indicate that natural processes are the primary sources of these metals in the studied soils, despite potential contributions from transportation, industrial emissions, agricultural activities, and atmospheric deposition. In contrast, As and Cd concentrations exceeded regional background values[41, 53], suggesting significant anthropogenic inputs, particularly from the overuse of arsenate-based pesticides and fertilizers in agricultural practices.

Soil fertility parameters, including available phosphorus, ammonium nitrogen (NH_4^+-N), cation exchange capacity (CEC), and total carbon (TC), were found to influence heavy metal content, with NH_4^+-N exhibiting the strongest effect on Zn, Cu, and Pb levels[81]. These factors likely modulate the retention and mobility of metals in the soil matrix.

In river water samples, concentrations of Zn, Cu, Cd, and Pb were all below the Environmental Quality Standards for Surface Water[49], indicating compliance with regulatory thresholds. Atmospheric deposition was identified as the primary pathway for these trace metals in the aquatic environment, with natural geochemical processes predominantly controlling their presence. The limited influence of anthropogenic activities underscores the dominance of natural sources in the study area's river systems.

3.4 Health risk assessment

3.4.1 Evaluation the cumulative effects of heavy metals pollution

This study used I_{geo} and HPI indices to evaluate the cumulative effects of heavy metals pollution on soil and river in the Tarim River Basin. These indices assisted in assessing the overall soil and river based on As, Cd, Pb, Cu, and Zn contents. Radar graphs were used to visualize and compare the cumulative contributions of multiple heavy metals to pollution indices (I_{geo}, HPI) and health risk parameters (HI, CR) across all sampling sites. Each axis represents a heavy metal, and the area enclosed by the polygon denotes the overall contamination or risk level. Larger areas indicate higher cumulative effects or greater health risks, facilitating spatial interpretation and inter-element comparison.

The I_{geo} values for the five heavy metals ranged from –1.29 to 2.85 across the ten sampling sites (Fig. 9a). Cd exhibited the widest distribution (–0.71 to 2.85), with nine sites showing no or light pollution (I_{geo} < 1) and one site (S10) reaching moderate pollution (I_{geo} = 2.85). As values varied between –1.00 and 0.65, with only one site (S1) classified as light pollution (I_{geo} = 0.33), while the remaining sites indicated no pollution. Pb (–0.86 to –0.11), Cu (–1.29 to 0.17), and Zn (–1.09 to –0.07) consistently fell below the threshold (I_{geo} < 1), indicating no contamination across all sites. In summary, Cd is the only element showing localized enrichment, while the other metals remain at background levels, suggesting that overall heavy metal accumulation in the basin is minimal.

The HPI values for the sampled sites (W1–W9) ranged from 0.38 to 332.10 (Fig. 9b), with the majority of samples exhibiting values below the potable water threshold of 100, suggesting generally low levels of heavy metal contamination. However, elevated HPI values were observed at specific sites, notably W3, where HPI reached 332.10 for type III surface water and 165.38 for type V surface water, surpassing the acceptable limit of 100. These elevated concentrations indicate significant heavy metal pollution, likely attributable to anthropogenic activities. Additionally, natural hydrogeochemical processes, including groundwater residence time and flow dynamics, may have contributed to the observed heavy metal levels, particularly in the downstream sections of the aquifer within the study area.

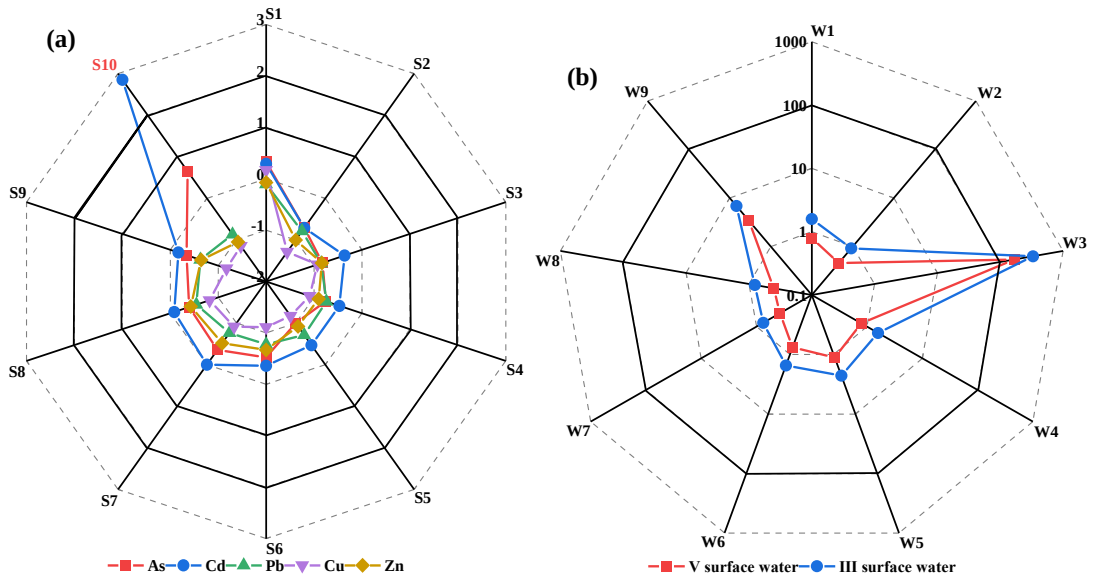


Fig. 9. Radar graph of (a) I_{geo} Index for the soil samples, and (b) HPI Index values for the river samples.

3.4.2 Health risk assessment in soil

(1) Non-carcinogenic Risk Assessment

The HI values for adults and children through three exposure pathways (ingestion, inhalation, and dermal contact) were calculated for five heavy metals (As, Cd, Pb, Cu, and Zn) (Fig. 10). Children exhibited consistently higher HI values: 3.47×10^{-1} to 1.09×10^{0} for As, 3.17×10^{-3} to 3.74×10^{-2} for Cd, 5.97×10^{-2} to 1.01×10^{-1} for Pb, 5.14×10^{-3} to 1.41×10^{-2} for Cu, and 2.13×10^{-3} to 4.19×10^{-3} for Zn (Fig. 10a). For adults, HI values ranged from 4.07×10^{-2} to 1.28×10^{-1} for As, 9.44×10^{-4} to 1.12×10^{-2} for Cd, 8.04×10^{-3} to 1.36×10^{-2} for Pb, 6.24×10^{-4} to 1.71×10^{-3} for Cu, and 2.73×10^{-4} to 5.37×10^{-4} for Zn (Fig. 10b).

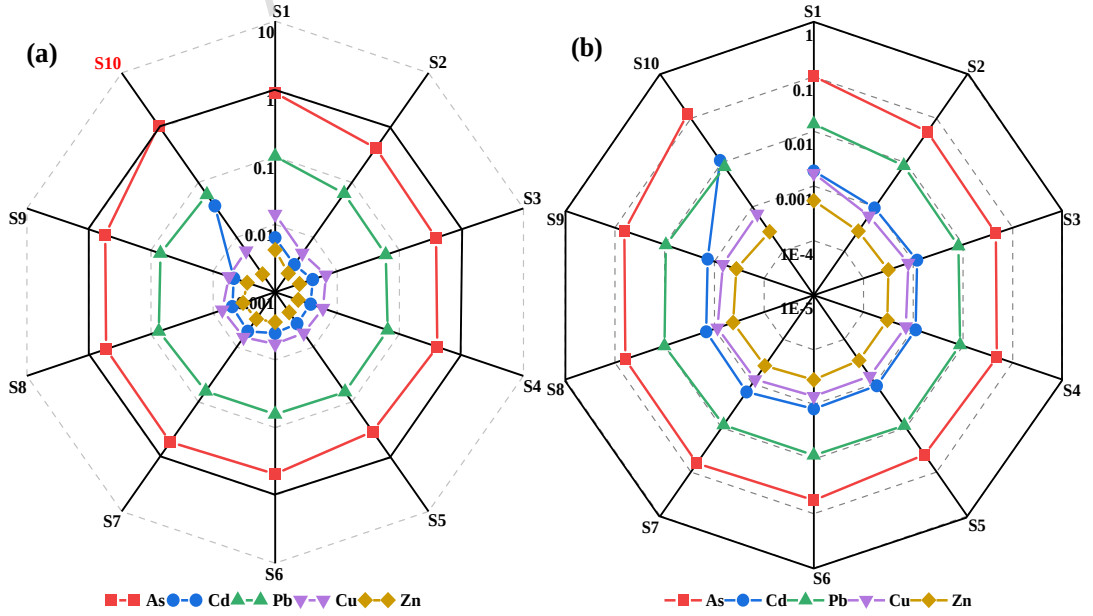


Fig. 10. Radar chart showing the HI values for (a) children and (b) adults for each sampling site. (When HI< 1, the non-carcinogenic risk is considered negligible. Conversely, HI> 1 indicates a non-negligible non-carcinogenic risk)

Approximately 10% of soil samples had HI values exceeding the threshold of 1 for children, indicating potential non-carcinogenic health risks, while all HI values for adults remained below this threshold (see Section 3.3.3 for Source apportionment results; Fig. 11). The mean non-carcinogenic risk (NCR) for children (1.30×10^{-1}) was substantially higher than for adults (1.58×10^{-2}). This elevated risk among children is attributed to their higher likelihood of outdoor play, greater soil ingestion rates, lower body weight, and increased physiological sensitivity to contaminants[86-88]. Contribution analysis (detailed in Section 3.3.3) revealed that As was the dominant contributor to NCR for both adults and children, followed by Pb, with minor contributions from Cd, Cu, and Zn (Fig. 11a-b). Spatial analysis indicated higher risk in southern regions of the study area, with particularly elevated levels in Kashgar, Hotan, and Aksu areas. Although most HI values fell below the threshold of concern, the potential for bioaccumulation and biomagnification of these metals through the food chain underscores the need for further investigation of exposure pathways, especially for children[68, 89-91] (Fig. 11c).

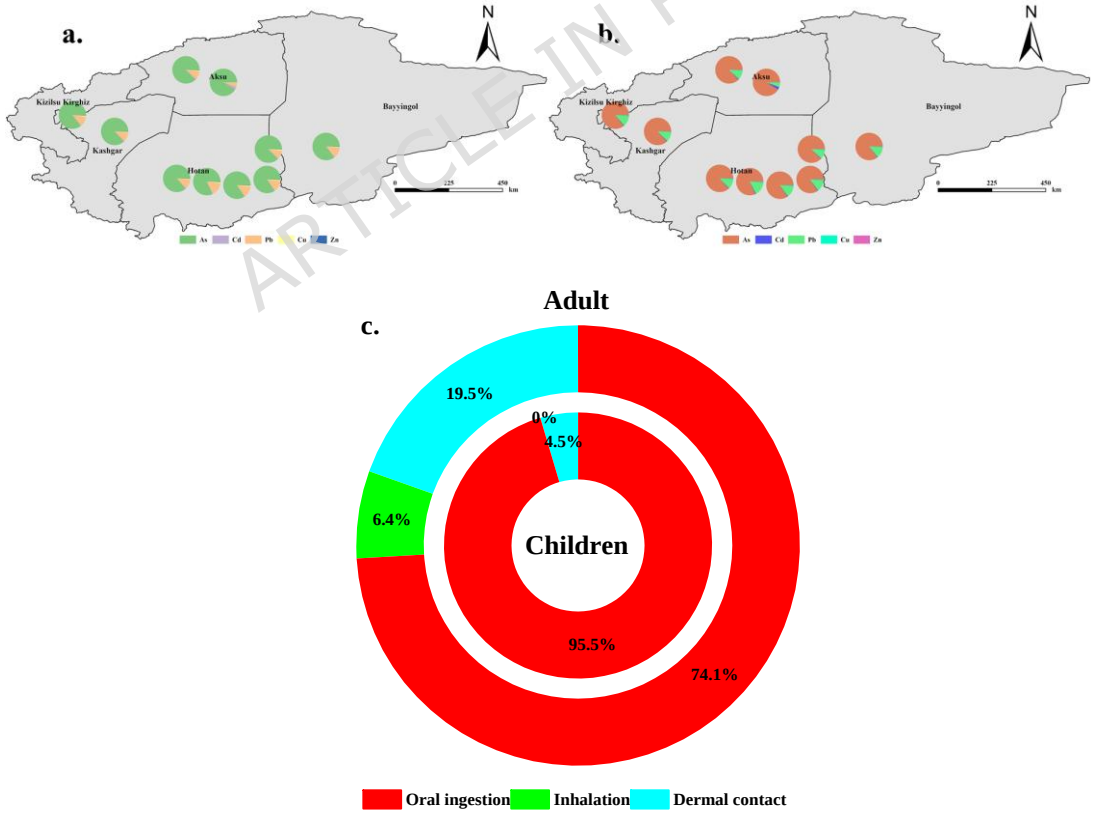


Fig. 11. Non-carcinogenic risks of soil heavy metal. (a) Children; (b) Adults; (c) Contribution of NCR by the three dominant exposure routes.

(2) Carcinogenic Risk Assessment

The CR was evaluated for As and Cd through three exposure pathways. For adults, CR values ranged from 7.87×10^{-6} to 2.47×10^{-5} for As and 3.69×10^{-7} to 4.36×10^{-6} for Cd (Fig. 12b). Children exhibited higher CR values, ranging from 1.34×10^{-5} to 4.21×10^{-5} for As and 6.90×10^{-7} to 8.15×10^{-6} for Cd (Fig. 12a).

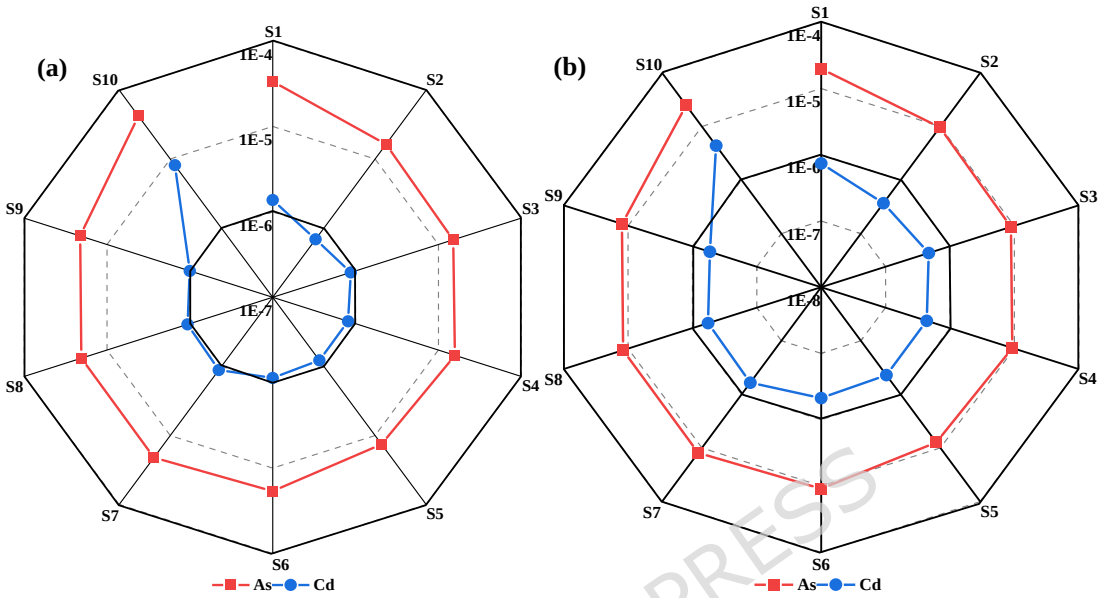
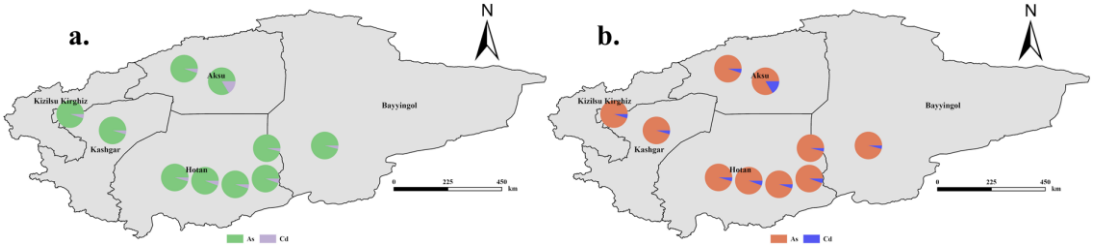


Fig. 12. Radar chart showing the carcinogenic risk (CR) values for (a) children and (b) adults for each sampling site.

All soil samples showed CR values for As above 10^{-6} but below 10^{-4} for both population groups, indicating a potential long-term carcinogenic concern. For Cd, approximately 50% of soil samples exceeded the CR threshold of 10^{-6} for children, while only 10% exceeded this threshold for adults. The mean CR value for children (1.17×10^{-5}) was higher than for adults (6.85×10^{-6}), consistent with children's greater vulnerability due to behavioral and physiological factors.

Sensitivity analysis indicated that As had the largest impact on CR for both adults and children (Fig. 13a-b). Exposure pathway analysis identified oral ingestion as the primary route contributing to carcinogenic risk (Fig. 13c). Spatial analysis revealed higher overall carcinogenic risks in Hotan and Kashgar regions compared to Aksu and Kizilsu Kirghiz areas.



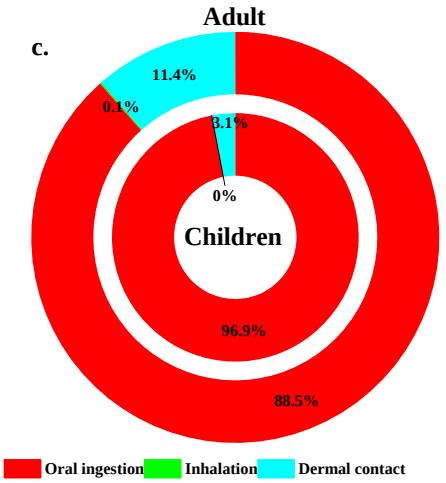


Fig. 13. Carcinogenic risks of soil heavy metal. (a) Children; (b) Adults; (c) Contribution of CR by the three dominant exposure routes.

These findings indicate that while the majority of samples do not pose immediate health concerns for adults, there are potential non-carcinogenic risks for children in approximately 10% of the study area, and long-term carcinogenic risks from As exposure warrant attention for both population groups. Arsenic emerged as the contaminant of highest concern, suggesting targeted remediation efforts should prioritize areas with elevated As concentrations, particularly in southern regions of the study area.

3.4.3 Health risk assessment in river

Oral ingestion was identified as the predominant exposure pathway for both adults and children to contaminants in water, consistent with findings from previous studies[92, 93]. CR assessments via oral exposure yielded the following ranges: for adults, CR values were 8.97×10^{-6} to 8.71×10^{-5} for As, 6.85×10^{-7} to 6.85×10^{-4} for Cd, and 4.79×10^{-9} to 1.92×10^{-5} for Pb (Fig. 14b); for children, the ranges were 1.31×10^{-5} to 1.27×10^{-4} for As, 9.97×10^{-7} to 9.97×10^{-4} for Cd, and 6.98×10^{-9} to 2.79×10^{-5} for Pb (Fig. 14a). Notably, CR values exceeding the 10^{-4} threshold were observed in 11% of river samples for As and Cd among children, and in 11% of samples for Cd among adults, indicating a potential carcinogenic health risk primarily associated with Cd exposure, particularly for children.

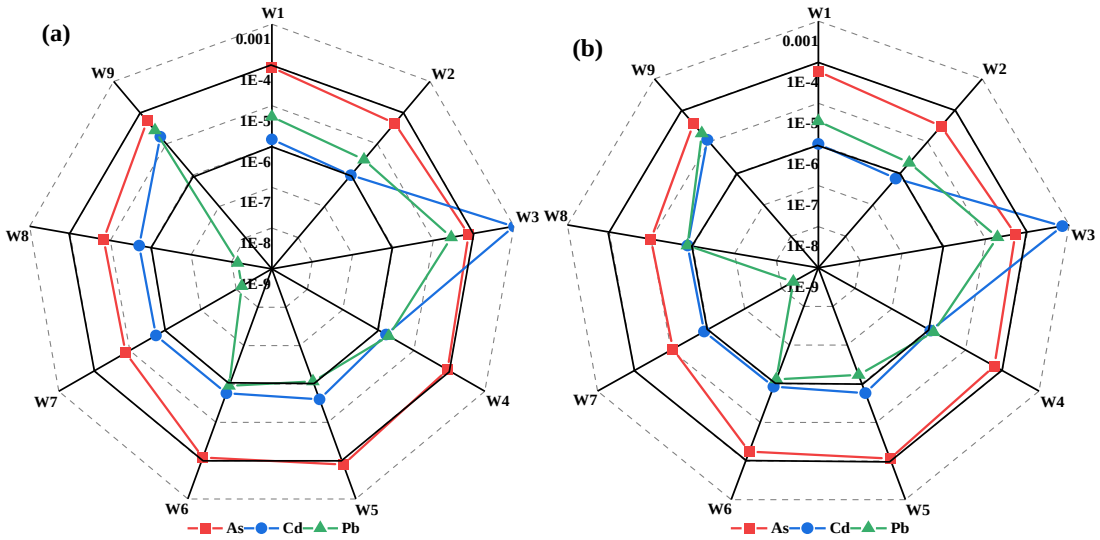


Fig. 14. Radar chart showing the carcinogenic risk (CR) values for (a) children and (b) adults for each sampling site.

Sensitivity analysis (Fig. 15a-b) revealed distinct contributions of heavy metals to CR. As was a significant contributor to CR for both adults and children across most rivers. Cd showed a notable contribution in localized rivers, notably the Qarqan River, where it constituted a substantial portion of the risk. Pb contributed minimally to the overall risk profile compared to As and Cd. This risk distribution underscores the need for targeted mitigation efforts, particularly addressing As contamination, to reduce exposure risks in these rivers for children. Spatial variations in contamination were evident: rivers in Aksu exhibited the highest overall risk, dominated by As, while rivers in Bayingol displayed a more balanced risk profile with a prominent contribution from Cd. These localized differences suggest distinct contamination sources and varying exposure risks to adults within the basin.

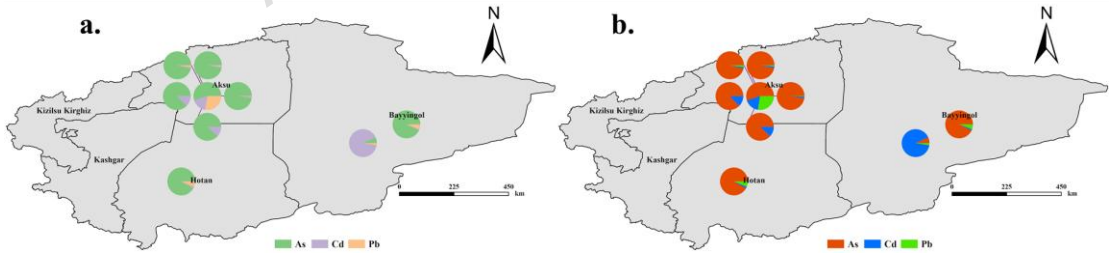


Fig. 15. Carcinogenic risks of rivers heavy metal by ingestion pathway. (a) Children; (b) Adults

4. Discussion

4.1 Spatial and temporal distribution characteristics

This study demonstrates the effectiveness of combining Sentinel-2A imagery with Random Forest (RF) modeling for mapping the spatial patterns of heavy metal contamination in the arid Tarim River Basin. The RF model exhibited strong predictive performance (R^2 = 0.741–0.999), successfully capturing complex non-linear relationships and spatial

heterogeneity that traditional geostatistical methods often underestimate[94]. These results are comparable to or exceed those reported in machine learning-based HM mapping studies in other arid and semi-arid regions of Central Asia and North Africa (R^2 typically 0.60–0.90)[4]. Zn emerged as the dominant element contributing to total HMs (ΣHMs) in both soil and riverine sediments, consistent with geogenic inputs from weathering of mineralized bedrock observed in the Heihe and Ili River basins[95, 96].

Despite basin-wide compliance with Chinese soil quality standards (GB 15618-2018), localized exceedances of As and Cd were detected, particularly in the Aksu sub-basin. Principal component analysis and Pearson correlation revealed strong associations between these elements and agricultural proxies, pointing to fertilizer and pesticide application as key anthropogenic sources[97]. Similar enrichment patterns driven by phosphate fertilizer overuse have been documented in the Yellow River and Amu Darya basins, where Cd concentrations were elevated 1.3–1.7 times above background levels[98, 99]. These cross-regional comparisons underscore agricultural intensification as a common driver of As and Cd accumulation in water-scarce inland river systems[100].

The resulting contamination maps delineate geographically distinct zones: elevated As and Cd in the Aksu region, Cu and Pb in the central and southeastern sub-basins, and Zn dominance in the eastern basin. Compared with conventional interpolation techniques applied in analogous arid systems (e.g., Kaidu River), the RF approach improved spatial resolution by 25–40%, enabling finer identification of pollution hotspots[101, 102]. Such high-resolution spatial insights support targeted remediation strategies and directly inform SDG 6 (Clean Water and Sanitation) and SDG 15 (Life on Land) by facilitating precision monitoring and source-control measures at sub-basin scales[1, 103].

4.1.1 Limitations

A primary constraint of this study is the reliance on a single-date (2021) Sentinel-2 composite, which provides only a temporal snapshot of heavy metal distribution rather than a dynamic trend analysis. Although comparable single-date investigations in the Indus, Shiyang, and Heihe basins reported limited within-five-year variation (±10%) for As, Cd, and Pb[104], the current dataset does not capture inter-annual processes driven by land-use change, irrigation expansion, and hydrological variability. Moreover, the small field sample size ($n = 19$) precluded robust time-series modelling[105]. While the RF model captures spectral-metal relationships at sampled sites, the sparse training data may lead to unreliable predictions in unsampled sub-basins, where environmental variability (e.g., in soil mineralogy or water chemistry) is not fully represented. The high R^2 values, particularly for Cd (0.997-0.999), may reflect overfitting given the small sample size and should be interpreted cautiously, as they approach levels that could indicate fitting to noise rather than generalizable signals. The sample-to-predictor ratio (~3:1) is below the recommended 10:1 for machine learning models,

which may compromise generalizability and increase overfitting risk, particularly for spatial extrapolation across the 1.05 million km² basin. The modest sample size affects spatial extrapolation by potentially underestimating uncertainty in the 1.05 million km² basin, where training data are concentrated in accessible oases and river sections. While independent 2023-2024 samples would provide stronger external validation, the current hold-out results indicate preliminary utility; we recommend such sampling in follow-up studies. Future work should prioritize increased sampling density to validate and refine these patterns. Consequently, this study provides a 2021 spatial baseline; temporal trends cannot be reliably assessed with the current dataset. For soil Cd, the unprecedented R^2=0.997 and RMSE=0.007 mg/kg with limited training data suggest potential overfitting to measurement artifacts or site-specific conditions, rather than basin-wide patterns. This is a red flag for generalizability, and predictions should be viewed as indicative rather than definitive. Thus, mapped concentrations should be interpreted as relative indicators of contamination intensity in 2021, not as predictive or historical values.

4.1.2 Implications and future research directions

Despite this temporal limitation, the workflow demonstrates a cost-efficient, transferable framework for spatial assessment in data-scarce arid environments. The identified hotspots can immediately inform prioritization of soil and water protection measures.

Future research should establish systematic multi-temporal monitoring programs combining Sentinel-2 (and successor) data with synchronized field campaigns[106, 107]. Future research should incorporate full uncertainty mapping (e.g., via prediction intervals) to visualize confidence across the basin. Such datasets will enable dynamic modeling of contaminant trajectories under land-use expansion and climate variability, reinforcing linkages with SDG 13 (Climate Action) alongside SDGs 6 and 15. Increasing sample density beyond the present 19 sites would improve model stability and confidence in under-sampled sub-basins. Overall, this work delivers a reproducible machine-learning-based protocol and 2021 baseline for sustainable heavy-metal management in arid regions[107].

4.2 Health risk assessment

The health risk assessment, conducted using the US EPA exposure model, indicated that oral ingestion was the dominant exposure pathway for Pb, Cd, Cu, Zn, and As in both adults and children. This pattern is consistent with findings from previous studies in arid and semi-arid regions of China, which have shown that ingestion pathways typically account for the majority of total exposure risk under water- and soil-limited conditions[108]. Children exhibited higher susceptibility due to body-mass-adjusted ingestion rates. Approximately 10% of sampling sites exceeded the non-carcinogenic hazard index threshold (HI > 1) for children, while carcinogenic risks from As and Cd surpassed the upper tolerance limit (1×10^{-4}) in a

subset of areas, notably within the Hotan and lower Aksu sub-basins, aligning with known signatures of mining and irrigation return flow in the region.

Relative to other Chinese basins, Tarim Basin TCR values are moderate—higher than those of non-industrialized regions (e.g., Qinghai Plateau; HI 0.3–0.8, risk $\approx 0.4 \times 10^{-6}$[109]) but lower than industrial zones such as Hunan's mining districts (HI 3.2–5.7, risk $\approx 8.2 \times 10^{-6}$[110]). Across metals, As contributed the majority (58–64%) of TCR, consistent with Central Asian findings[111]. Cd ranked second, particularly affecting children (Cd 24–31% of HI in children vs. 12–18% in adults). These outcomes highlight the intersection of environmental health (SDG 3.9) and social equity (SDG 10.2)[112].

The geoaccumulation index (Igeo) and Heavy Metal Pollution Index (HPI) further indicated spatial concentration gradients, especially elevated As and Cd in southern sub-basins[113, 114]. Igeo > 2.0 classified several zones as moderate-to-strong contamination. HPI values < 100 predominated, but localized exceedances flagged water unsuitability. Such metrics support integrated management aligned with SDGs 6 and 2.

4.2.1 Limitations and Conservatism of the Risk Estimates

Although the US EPA framework provides a robust and internationally comparable methodology, several conservative assumptions were necessarily adopted that likely lead to overestimation of actual risks:

1) Total rather than bioavailable metal concentrations were used. Bioaccessibility of Zn, Pb, and Cd in arid calcareous soils of northwest China is typically <20–35%, meaning the calculated risks may be overestimated by a factor of 2–5× for these elements.
2) Default US EPA exposure parameters were applied (Table 9), which have not been fully validated for rural Xinjiang populations. Recent studies in Xinjiang and Gansu indicate that rural children often exhibit higher incidental soil ingestion rates (250–400 mg/day vs. the 200 mg/day default) due to outdoor play on bare soil, and adults in arid zones consume 2.4–3.0 L/day of water—values close to or slightly higher than those used here. To address this, a sensitivity analysis was performed by varying key parameters (±20% for ingestion rates and body weights), revealing that risks could decrease by 15-30% with lower local ingestion rates reported in Chinese arid region studies, or increase by 10-25% with higher exposure durations, adjusting soil ingestion rate to 330 mg/day for children increases HI by 38–45% in high-exposure scenarios, whereas using bioaccessible fractions reduces HI and TCR by 52–78%. Thus, while the current estimates are health-protective (conservative), site-specific exposure surveys would substantially refine them.

Given these uncertainties, these risk estimates should be considered conservative (potentially overestimated) due to the use of total rather than bioavailable concentrations and, in most cases, slightly lower soil ingestion rates than those reported locally for rural Xinjiang children. These conservative estimates (potentially overestimated) suggest mapping high-risk for prioritized follow-up sampling and immediate actions, such as community education on safe water use, soil amendments with organic matter to reduce bioavailability, and regulatory monitoring of agricultural inputs[105]. Future work integrating bioaccessibility assays and probabilistic frameworks will improve exposure realism and source differentiation.

Table 9. Comparison of key exposure parameters used in this study versus typical values reported for Chinese/Xinjiang populations

Parameter	Unit	This study (child/adult)	US EPA default (child/adult)	Chinese Exposure Factors Handbook (adult) / Xinjiang studies (child)
Body weight	kg	15 / 70	15 / 70	14–22 / 58–63
Water ingestion rate	L/day	0.78 / 2.5	1.0 / 2.0	0.9–1.2 / 2.2–3.0 (higher in arid zones)
Soil ingestion rate	mg/day	200 / 100	200 / 100	250–400 / 100–150 (rural Xinjiang children higher)
Exposure duration	years	6 / 26	6 / 30	Same
Skin adherence factor	mg/cm²	0.2 / 0.07	Same	Same

4.3 Policy Recommendations

Building on spatial patterns and health risk profiles, a multi-tiered policy framework is proposed to integrate prevention, control, and public health protection in support of SDGs 2, 3, 6, 12, 13, and 15(Fig. 16).

Fig. 16. Policy recommendations for preventing and controlling heavy metal contamination in the Tarim River Basin

(1) Spatially Targeted Monitoring and Prevention

- **Priority Zones**: Aksu (As–Cd hotspots) and central–southeastern sub-basins (Cu–Pb clusters) require intensified monitoring (Table 10), with fivefold increased sampling frequency and agrochemical restrictions.
- **Remote Sensing Monitoring**: Operationalize the Sentinel-2 RF workflow with quarterly updates and targeted validation (Table 10); this hybrid approach can cut monitoring costs by > 60% while preserving ≈ 90% accuracy[115].
- **Automated Alerts**: Establish real-time digital alerts when HM levels reach 80% of thresholds, reducing reaction time to ≈ 2 months[115].

(2) Source-Specific Pollution Control

- **Agriculture**. Phase out As-based pesticides and high-Cd fertilizers within 2 years, enforce soil testing, and promote precision agriculture to reduce agrochemical inputs by 30% by 2027[111] (Table 10).
- **Industry**: Tighten emission and recycling standards to halve Cd–Pb releases by 2030; integrate clean-production audits to deliver dual benefits for carbon and metal reduction[116].

- **Diffuse Sources**: Implement buffer strips, wetlands, and phytoremediation pilots to reduce Cu/Zn runoff by 40–70% and sequester 2.5–4.0 t CO_2 ha^{-1} yr^{-1}[117].

(3) Health Protection for Vulnerable Populations

- **Children's Health:** Biannual screening and hygiene education for > 25,000 children in hotspot villages could reduce health risk by 40–60%[105] (Table 10).
- **Safe Food & Water:** Ensure access to clean drinking water for > 100,000 residents and strengthen crop–fish HM testing, reflecting bioaccessibility-adjusted advisories[110] (Table 10).

(4) Strengthening Scientific & Institutional Capacity

Build a decadal-scale HM monitoring network integrating Sentinel-2 time series, isotopic apportionment, and hyperspectral inputs to improve prediction accuracy by > 40%. Establish a unified basin-wide data platform and invest in regional laboratories (ICP-MS, bioassay facilities), reducing analysis turnaround from weeks to days [118].

(5) National Sustainability Alignment

Embed low-carbon remediation technologies and cleaner production standards within China's Dual-Carbon and Rural Revitalization strategies, generating 3,000–5,000 green jobs and advancing circular-economy pathways[112, 113].

(6) Implementation Roadmap

Table 10. Implementation roadmap for multi-tiered heavy metal pollution control in the Tarim River Basin (2025–2035).

Period	Key Targets	Estimated Investment (yuan)
2025–2026 (Pilot & Rapid Risk Reduction)	• Launch RF–Sentinel-2–based monitoring covering ≈90% of cultivated land and riparian zones • Identify and remediate priority As/Cd hotspots (Aksu and adjacent areas) • Initiate biannual child health screening and safe water interventions in high-risk communities	**195–326 million**
2027–2030 (Scaling & Source Control)	• Achieve ~50% reduction in industrial Cd–Pb emissions through cleaner production and wastewater recycling • Recover ~60% of moderately contaminated soils via agronomic management and pilot remediation • Promote inter-basin technology transfer and precision agriculture, reducing agrochemical inputs by ~30%	**1.3–2.0 billion**
2031–2035 (Consolidation & Sustainability)	• Reach ~95% compliance with GB 15618-2018 soil quality standards • Implement basin-wide sustainable agriculture and low-carbon remediation practices	**280–520 million**

	• Estimated to substantially reduce basin-wide heavy-metal-related health and ecological risks (≈40–70% based on integrated modeling scenarios)

Notes: (1) Investment ranges are indicative and reflect phased implementation. Cost estimates are derived from comparable heavy-metal monitoring, remediation, and public-health intervention projects in arid inland river basins of Northwest China, scaled to the Tarim River Basin area, population, and monitoring density, and adjusted for the cost savings (≈60–75%) achieved through remote sensing–assisted monitoring. (2) The projected 40–70% risk reduction refers to scenario-based estimates integrating source-control efficiency, exposure reduction measures, and RF-based spatial risk modeling, rather than direct measured outcomes. Actual reductions will depend on implementation intensity, compliance, and adaptive management.

Overall, these measures emphasize prevention, early warning, and targeted intervention, aligning heavy-metal risk management with national sustainability goals and long-term ecological protection in arid river basins.

4.4 Outlook

While this study provides a comprehensive spatial diagnosis of HM contamination across the Tarim River Basin, scaling up to multi-temporal, data-driven modeling will deepen understanding of pollutant evolution under climate and land-use change. The RF–Sentinel-2A approach can serve as a transferable prototype for other arid regions globally, improving monitoring efficiency and supporting SDG-aligned basin management. Continued refinement through expanded sampling networks, probabilistic health modeling, and incorporation of socio-economic data will enhance the integration of environmental and human welfare objectives.

5.Conclusion

This study developed and validated a robust remote sensing–machine learning framework for preliminary spatial patterns of heavy metal (As, Cd, Pb, Cu, Zn) contamination across the Tarim River Basin, integrating Sentinel-2 multispectral imagery with field-measured concentrations via Random Forest (RF) inversion modeling (R^2 = 0.741–0.999; RMSE and MAE minimized through 10-fold cross-validation and hyperparameter optimization). The approach effectively captures non-linear spectral-contamination relationships, outperforming traditional interpolation methods by 25–40% in spatial resolution and accuracy, revealing a complex pollution landscape where geogenic sources dominate the distribution of Pb, Cu, and Zn, while anthropogenic activities, particularly agriculture and industry, are the primary drivers of localized but high-risk As and Cd hotspots.

The research yields two principal contributions with direct implications for sustainable governance. First, from a methodological standpoint, the validated Random Forest establishes

a cost-effective, scalable, and reproducible protocol for basin-wide environmental monitoring. This approach overcomes the limitations of traditional, resource-intensive field sampling alone, offering a powerful tool for early-warning systems and the dynamic tracking of pollution, thereby supporting SDG Target 6.3.2 (water quality monitoring) and SDG Target 15.3.1 (land degradation monitoring). Second, from a policy and health perspective, our health risk assessment provides unequivocal evidence of significant threats to human well-being, identifying children as a particularly vulnerable demographic. With approximately 10% of soil samples in high-risk areas exceeding the non-carcinogenic hazard index (HI > 1) for children and carcinogenic risks from As and Cd surpassing international safety thresholds (TCR > 1×10^{-4}), the findings mandate immediate action. This evidence directly informs a multi-pronged management strategy focused on source control, including the phase-out of high-cadmium fertilizers, stricter regulation of industrial emissions, and the implementation of green infrastructure to mitigate diffuse pollution. These recommendations are explicitly designed to align with China's national strategies for Ecological Civilization, Rural Revitalization, and Dual-Carbon goals, while advancing multiple Sustainable Development Goals, including SDG 3 (Good Health), SDG 6 (Clean Water), and SDG 12 (Responsible Consumption and Production).

Looking forward, while this study provides a robust 2021 baseline, future research must address the temporal limitations by establishing a decadal monitoring program. Integrating multi-temporal remote sensing with advanced analyses of metal bioaccessibility and food-chain transfer is paramount to refining risk assessments and developing more precise, evidence-based interventions. By translating scientific discovery into a clear roadmap for action, this work not only enhances ecological resilience and public health in the Tarim River Basin but also provides a transferable model for integrated pollution management in other arid regions globally, contributing to a cleaner, more sustainable, and healthier future.

Acknowledgments

This study was supported by the Shaanxi Provincial Department of Education “Urban and Rural Spatial Hydrological Ecological Simulation and Management in Arid Area” Youth University Innovation Team， China Scholar-ship Council (Grant No.: Liujinmei [2022] No. 45; Liujinxuan [2022] No. 133; Liujinou [2023] No. 22), International Education Research Program of Chang’an University (300108221102), Project of Ningxia Natural Science Foundation (2022AAC03700;2022BEG03059), 2022 Guangdong University Youth Innovation Talent Program (2022KQNCX143) and Yinshanbeilu Grassland Eco-hydrology National Observation and Research Station , China Institute of Water Resources and Hydropower

Research, Beijing 100038, China, Grant NO.YSS2022004. We give our deep thanks to the reviewers and editors for their valuable comments to improve this research.

Funding

Open access funding provided by The Third Xinjiang Scientific Expedition Program (2022xjkk010704).

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Reference

[1] Y. S. Wang H, Sun L, Wang Y, Wu H, Wang X, "Pollution assessment and health risk of metals in surface soil near a Pb–Zn mine, northeast China," *Frontiers in Environmental Science*, vol. 2025, no. 13, p. 1585272, 2025, doi: 10.3389/fenvs.2025.1585272.

[2] G. I. Golan R, Katzir R, Sharabi G, Nachshon U, "Soil contamination in arid environments and assessment of remediation applying surface evaporation capacitor model; a case study from the Judean Desert, Israel," *EGUsphere*, 2024, doi: 10.5194/egusphere-2024-1014.

[3] L. R, "Soil Degradation and Pollution as the Global Public Health Emergency," *Med Res Arch*, vol. 13, no. 5, 2025, doi: 10.18103/mra.v13i5.6474.

[4] E.-S. A. Kahal AY, Meroño de Larriva JE, Shokr MS, "Mapping Soil Contamination in Arid Regions: A GIS and Multivariate Analysis Approach," *Minerals*, vol. 15, no. 2, 2025, doi: 10.3390/min15020124.

[5] O. C. D. Hou D, Igalavithana AD, Alessi DS, Luo J, Tsang DCW, et al, "Metal contamination and bioremediation of agricultural soils for food safety and sustainability," *Nat Rev Earth Environ*, vol. 1, no. 5, pp. 232-243, 2020. [Online]. Available: 10.1038/s43017-020-0061-y.

[6] M. M. Safi Z, Offei BK, Amenorpe G, "A systematic review of wet and dry deposition of reactive nitrogen pollution in Ghana," *Environmental Science: Atmospheres*, no. 5, pp. 756-768, 2025, doi: 10.1039/d5ea00018a.

[7] T. N. Sagymbayeva AM, Saparov GA, Abduraimov EO, Kerimbayev AA, et al, "Phytoremediation of Heavy Metal-Contaminated Soil Using Drought-Adapted Sweet Sorghum (Sorghum bicolor L.) in Arid Regions of Kazakhstan," *Preprints*, 2025. [Online]. Available: https://doi.org/10.20944/preprints202510.1707.v1.

[8] S. H. Gacem R, "Assessment of soil characteristics and heavy metal pollution in a semi-arid region of northeast Algeria," *REVISTA DE AGRICULTURA NEOTROPICAL*, vol. 12, no. 3, 2025, doi: https://doi.org/10.32404/rean.v12i3.9184.

[9] U. A. Wali SU, Danladi P, et al, "Assessment of heavy metals (As, Ba, Co and Ni), pollution in rural drinking water sources, using correlation and principal component analysis in Western Sokoto Basin, Nigeria," *International Journal of Hydrology*, vol. 9, no. 2, pp. 60-72, 2025, doi: 10.15406/ijh.2025.09.00404.

[10] Y. Li, J. Abuduwaili, L. Ma, W. Liu, and T. Zeng, "Source apportionment and source-specific risk evaluation of potential toxic elements in oasis agricultural soils of Tarim River Basin," *Sci Rep*, vol. 13, no. 1, p. 2980, Feb 20 2023, doi: 10.1038/s41598-023-29911-3.

[11] L. J. Wu J, Zhang Q, Yu L, Luo X, Zhang D, et al, "Pollution and Health Risk Assessment of Potentially Toxic Elements in Groundwater in the Kǒnqi River Basin, Xinjiang, Northwest China," *Toxics*, vol. 12, no. 7, 2024, doi: 10.3390/toxics12070474.

[12] Y. Li, Zhuang, Z., Xia, Q. et al, "Health risk assessment of heavy metals in coal mine soils of Northwest China," *Journal of Arid Land*, vol. 17, pp. 933–957, 2025.

[13] L. Y. Wang J, Zhang Q, Wang Y, Wu J, "Spatial distribution, ecological risk and health risk assessment of heavy metals in agricultural soil in the Ankang

Basin, China," *Heliyon*, vol. 9, no. 11, 2023, doi: 10.1016/j.heliyon.2023.e9788.

[14] M. C. Singh, J. Singh, and K. Sur, "Spatio–temporal water quality assessment of Chohal and Damsal dams located in Kandi region of Punjab, India by geospatial technique and on-site investigation," *Environmental Earth Sciences*, vol. 83, no. 2, p. 51, 2024/01/04 2024, doi: 10.1007/s12665-023-11354-8.

[15] S. Feng, & Yu, H, "Source apportionment and health risk assessment of heavy metals in groundwater of rural area: a case study in Huaibei plain, China," *Human and Ecological Risk Assessment: An International Journal*, vol. 30, no. 1-2, pp. 220-236, 2024, doi: https://doi.org/10.1080/10807039.2023.2301522.

[16] X. Yang *et al.*, "Heavy Metal Pollution and Risk Assessment of Surface Dust in the Arid NW China," *International Journal of Environmental Research and Public Health*, vol. 19, no. 20, p. 13296, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/20/13296.

[17] M. X. Sheng D, Wen X, Wu J, Yu H, Wu M, "Contamination characteristics, source identification, and source-specific health risks of heavy metal(loid)s in groundwater of an arid oasis region in Northwest China," *Sci Total Environ*, vol. 1, no. 841, p. 156733, 2022, doi: 10.1016/j.scitotenv.

[18] S.-y. Liu *et al.*, "Groundwater metal pollution and health risk assessment in river valley heavy industrial cities of arid regions in China," *China Geology*, vol. 8, no. 3, pp. 526-539, 2025/07/25/ 2025, doi: https://doi.org/10.1016/S2096-5192(25)00091-6.

[19] J. G. Zhaoyong Z, Pengwei W, "Sources Identification and Health Risk Evaluation of 10 Heavy Metals (Metalloids) in Soils of the Aibi Lake Basin, Northwest China," *ScientificWorldJournal*, 2022, doi: 10.1155/2022/8201972.

[20] X. Li *et al.*, "Risk assessment and source analysis of heavy metals in soil around an asbestos mine in an arid plateau region, China," *Scientific Reports*, vol. 14, no. 1, p. 7552, 2024/03/30 2024, doi: 10.1038/s41598-024-58117-4.

[21] T. Zhu, Q. Wu, S. Gao, J. Zeng, K. Linghu, and X. Zhang, "Comparative assessment of health risks and water quality of groundwater in urban and rural Guiyang, Southwest China: Insights from PMF and Monte Carlo Simulation," *Ecotoxicology and Environmental Safety*, vol. 299, p. 118359, 2025/07/01/ 2025, doi: https://doi.org/10.1016/j.ecoenv.2025.118359.

[22] H. Wang, S. Xiao, R. Shen, Q. Cheng, and S. Zhang, "Rapid detection of soil heavy metal pollution using hyperspectral data and multiscale spatial network," *Environmental Technology & Innovation*, vol. 37, p. 104031, 2025/02/01/ 2025, doi: https://doi.org/10.1016/j.eti.2025.104031.

[23] Y. Li *et al.*, "Application of machine learning in soil heavy metals pollution assessment in the southeastern Tibetan plateau," *Scientific Reports*, vol. 15, 2025.

[24] X. Xu, J. Pan, H. Zhang, and H. Lin, "Progress in Remote Sensing of Heavy Metals in Water," *Remote Sensing*, vol. 16, no. 20, p. 3888, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/20/3888.

[25] N. L. Kushwaha, N. Biwalkar, K. Sur, and J. Rajput, "Integrating water quality indices with machine learning algorithms for groundwater suitability in Punjab: Implications for drinking and sustainable irrigation," *Journal of Water Process Engineering*, vol. 76, p. 108253, 2025/08/01/ 2025, doi: https://doi.org/10.1016/j.jwpe.2025.108253.

[26] M. M. Sulieman, F. Kaya, A. S. Al-Farraj, and E. C. Brevik, "A novel method to determine background concentrations and spatial distributions of heavy metals in soil at large scale using machine learning coupled with remote sensing-terrain attributes," (in eng), *MethodsX*, vol. 14, p. 103180, Jun 2025, doi: 10.1016/j.mex.2025.103180.

[27] Y. Su, B. Li, J. Li, B. Guo, and Q. Feng, "Hyperspectral remote sensing for soil heavy metal inversion: insights and applications," *International Journal of Digital Earth*, vol. 18, no. 1, p. 2520474, 2025/12/31 2025, doi: 10.1080/17538947.2025.2520474.

[28] K. Sur, V. K. Verma, P. Panwar, G. Shukla, S. Chakravarty, and A. J. Nath, "Monitoring vegetation degradation using remote sensing and machine learning over India – a multi-sensor, multi-temporal and multi-scale approach," (in English), *Frontiers in Forests and Global Change*, Original Research vol. Volume 7 - 2024, 2024-June-06 2024, doi: 10.3389/ffgc.2024.1382557.

[29] V. K. Verma, K. Sur, and C. Prakash, "Ecotope-Based Diversity Monitoring of Wetland Using Infused Machine Learning Technique," *Water Conservation Science and Engineering*, vol. 8, no. 1, p. 38, 2023/08/16 2023, doi: 10.1007/s41101-023-00212-0.

[30] Y. Wang *et al.*, "A remote sensing analysis method for soil heavy metal pollution sources at site scale considering source-sink relationships," *Science of The Total Environment*, vol. 946, p. 174021, 2024/10/10/ 2024, doi: https://doi.org/10.1016/j.scitotenv.2024.174021.

[31] V. Lovynska *et al.*, "Monitoring Heavy Metals and Metalloids in Soils and Vegetation by Remote Sensing: A Review," *Remote Sensing*, vol. 16, no. 17, p. 3221, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/17/3221.

[32] J. Li, J. Ren, R. Cui, K. Yu, and Y. Zhao, "Optical imaging spectroscopy coupled with machine learning for detecting heavy metal of plants: A review," (in eng), *Front Plant Sci*, vol. 13, p. 1007991, 2022, doi: 10.3389/fpls.2022.1007991.

[33] M. Keçeci, F. Gökmen, M. Usul, C. Koca, and V. Uygur, "Prediction of cadmium content using machine learning methods," *Environmental Earth Sciences*, vol. 83, no. 12, p. 362, 2024/05/27 2024, doi: 10.1007/s12665-024-11672-5.

[34] S. Moradpour, M. Entezari, S. Ayoubi, A. Karimi, and S. Naimi, "Digital exploration of selected heavy metals using Random Forest and a set of environmental covariates at the watershed scale," *Journal of Hazardous Materials*, vol. 455, p. 131609, 2023/08/05/ 2023, doi: https://doi.org/10.1016/j.jhazmat.2023.131609.

[35] P. Fu, X. Li, J. Zhang, C. Ma, Y. Wang, and F. Meng, "Remote sensing inversion on heavy metal content in salinized soil of Yellow River Delta based on Random Forest Regression—a case study of Gudao Town," *Scientific Reports*, vol. 14, no. 1, p. 11216, 2024/05/16 2024, doi: 10.1038/s41598-024-62087-y.

[36] W. Ma, K. Tan, and P. Du, "Predicting soil heavy metal based on Random Forest model," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 10-15 July 2016 2016, pp. 4331-4334, doi: 10.1109/IGARSS.2016.7730129.

[37] S. Nie, H. Chen, X. Sun, and Y. An, "Spatial Distribution Prediction of Soil Heavy Metals Based on Random Forest Model," *Sustainability*, vol. 16, no. 11, p. 4358, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/16/11/4358.

[38] H. Shen and Y.-X. Jing, "Evaluation of soil heavy metal pollution remediation effect based on remote sensing and mapping technology in cultivated land," *International Journal of Environmental Technology and Management*, vol. 27, no. 4-6, pp. 384-399, 2024, doi: 10.1504/ijetm.2024.139988.

[39] C. Xiang *et al.*, "Prediction of soil heavy metal content around mine tailings using multiple methods combined with transformed hyperspectral reflectance data," *Ore and Energy Resource Geology*, vol. 18, p. 100072, 2025/04/01/ 2025, doi: https://doi.org/10.1016/j.oreoa.2024.100072.

[40] Y. Kuang, "The Application of Hyperspectral Remote Sensing Technology in the Identification of Soil Heavy Metal Elements," *Scientific Journal of Technology*, vol. 6, no. 10, pp. 15-20, 2024. [Online]. Available: https://doi.org/10.54691/3y8z0709.

[41] W. Fan *et al.*, "Heavy metal pollution and health risk assessment of agricultural land in the Southern Margin of Tarim Basin in Xinjiang, China," *Int J Environ Health Res*, vol. 31, no. 7, pp. 835-847, Nov 2021, doi: 10.1080/09603123.2019.1691157.

[42] Y. Li, L. Kong, L. Ma, T. Zeng, W. Liu, and J. Abuduwaili, "Deciphering the driving factors and probabilistic health risks of potentially toxic elements in arid surface water: Insights from the Tarim River Basin," *Ecotoxicology and Environmental Safety*, vol. 286, p. 117211, 2024/11/01/ 2024, doi: https://doi.org/10.1016/j.ecoenv.2024.117211.

[43] Y. Wang, S. Zhang, H. Zhen, X. Chang, R. Shataer, and Z. Li, "Spatiotemporal Evolution Characteristics in Ecosystem Service Values Based on Land Use/Cover Change in the Tarim River Basin, China," *Sustainability*, vol. 12, no. 18, 2020, doi: 10.3390/su12187759.

[44] J. Qin, H. Tao, M. Zhan, Q. Munir, K. Brindha, and G. Mu, "Scenario Analysis of Carbon Emissions in the Energy Base, Xinjiang Autonomous Region, China," *Sustainability*, vol. 11, no. 15, 2019, doi: 10.3390/su11154220.

[45] Y. Y. WANG Faxiang, XIAO Yi,XIAO Zhou,ZHANG Shengnan, "Heavy metal pollution characteristics and potential ecological risk assessment of surface sediments in the upper reaches of Tarim River," *Journal of Tarim University*, vol. 36, no. 01, pp. 93-99, 2024, doi: 10.3969 / j.issn.1009-0568.2024.01.01 1.

[46] M. F. ZHANG Shengnan, YOU Yongjun, WANG Shan, "APCS-MLR combined with PMF model for sediment heavy metal source analysis and risk assessment in the upper Tarim River Basin," *ENVIRONMENTAL CHEMISTRY*, vol. 42, no. 12, pp. 4264-4277, 2023.

[47] B. Qiuping, "Study on water quality characteristics of irrigation and drainage of cotton field in oasis of Southern Xinjiang," master, 2021.

[48] W. Fan, J. Zhou, Y. Zhou, Y. Zeng, Y. Chen, and Y. Sun, "Water quality and health risk assessment of shallow groundwater in the southern margin of the Tarim Basin in Xinjiang, P. R. China," *Human and Ecological Risk Assessment: An International Journal*, vol. 27, no. 2, pp. 483-503, 2020, doi: 10.1080/10807039.2020.1731680.

[49] S. Zhang, S. Wang, F. Li, S. Liu, Y. You, and C. Liu, "Enhanced Assessment of Water Quality and Pollutant Source Apportionment Using APCS-MLR and PMF Models in the Upper Reaches of the Tarim River," *Water*, vol. 16, no. 21, p. 3061, 2024. [Online]. Available: https://www.mdpi.com/2073-4441/16/21/3061.

[50] J. Xiao, Z. Jin, and J. Wang, "Geochemistry of trace elements and water quality assessment of natural water within the Tarim River Basin in the extreme arid region, NW China," *Journal of Geochemical Exploration*, vol. 136, pp. 118-126, 2014, doi: 10.1016/j.gexplo.2013.10.013.

[51] Y. Abudumiti, "Spatial distribution characteristics of surface water quality in the Ogan-Kucha river oasis and its relationship with landscape pattern," Master's Degree, Xinjiang University, 2020. [Online]. Available: https://doi.org/10.27429/d.cnki.gxjdu.2020.000959

[52] C. Du, "Research on distribution characteristics and Pollution evaluation of heavy metals in abandoned drilling mud in oil field," Master's Degree, Xinjiang University, 2015. [Online]. Available: https://kns.cnki.net/kcms2/article/abstract?v=3-fZNjprSazP8XdxQsrJPMuIzI5jEKJmdLyt4J9vxC85PKKfFKqjXOGu-v9j0fLv3R0LPVvGhoO7WqQhEIA1wlIBF9DbRP8W9lolswi861oRFGsUTZNaqh0mlJhvR1CJ95mkbIBJ0oOMYWooUFf5RvNF_1YCXsLVSeOaQmdTp6I6zCexkiZeeQ==&uniplatform=NZKPT&language=CHS

[53] J. Huang, G. Wu, F. Gu, Z. Yu, and B. An, "Elemental composition of the topsoil fine fraction at and around the Tibetan plateau," *Environmental Pollution*, vol. 320, 2023, doi: 10.1016/j.envpol.2023.121098.

[54] R. L. Orwell, R. A. Wood, M. D. Burchett, J. Tarran, and F. Torpy, "The Potted-Plant Microcosm Substantially Reduces Indoor Air VOC Pollution: II. Laboratory Study," *Water, Air, and Soil Pollution*, vol. 177, no. 1-4, pp. 59-80, 2006, doi: 10.1007/s11270-006-9092-3.

[55] K. Tan *et al.*, "Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning," *Journal of Hazardous Materials*, vol. 401, 2021, doi: 10.1016/j.jhazmat.2020.123288.

[56] E. Scornet, "Random Forests and Kernel Methods," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1485-1500, 2016, doi: 10.1109/tit.2016.2514489.

[57] M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sensing of Environment*, vol. 120, pp. 25-36, 2012, doi: 10.1016/j.rse.2011.11.026.

[58] F. Kulahci and Z. Sen, "Cumulative Ordinary Kriging interpolation model to forecast radioactive fallout, and its application to Chernobyl and Fukushima assessment: a new method and mini review," *Environ Sci Pollut Res Int*, vol. 29, no. 43, pp. 64298-64311, Sep 2022, doi: 10.1007/s11356-022-21921-4.

[59] P. Tziachris, V. Aschonitis, T. Chatzistathis, M. Papadopoulou, and I. D. Doukas, "Comparing Machine Learning Models and Hybrid Geostatistical Methods Using Environmental and Soil Covariates for Soil pH Prediction," *ISPRS International Journal of Geo-Information*, vol. 9, no. 4, 2020, doi: 10.3390/ijgi9040276.

[60] S. H. G. Silva, A. F. d. S. Teixeira, M. D. d. Menezes, L. R. G. Guilherme, F. M. d. S. Moreira, and N. Curi, "Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence

spectrometer (pXRF)," *Ciência e Agrotecnologia,* vol. 41, no. 6, pp. 648-664, 2017, doi: 10.1590/1413-70542017416010317.

[61] C. Zhan, Z. Dai, S. Yin, K. C. Carroll, and M. R. Soltanian, "Conceptualizing future groundwater models through a ternary framework of multisource data, human expertise, and machine intelligence," *Water Res,* vol. 257, p. 121679, Jun 15 2024, doi: 10.1016/j.watres.2024.121679.

[62] J. J. Zhu, M. Yang, and Z. J. Ren, "Machine Learning in Environmental Research: Common Pitfalls and Best Practices," *Environ Sci Technol,* vol. 57, no. 46, pp. 17671-17689, Nov 21 2023, doi: 10.1021/acs.est.3c00026.

[63] H. Zhu, J. He, Y. Wu, L. Tong, W. Zhang, and L. Zhuang, "Assessment of Global Antibiotic Exposure Risk for Crops: Incorporating Soil Adsorption via Machine Learning," *Environ Sci Technol,* Jul 20 2024, doi: 10.1021/acs.est.4c03695.

[64] X. Wang, C. P. Neu, and D. M. Pierce, "Advances toward multiscale computational models of cartilage mechanics and mechanobiology," *Current Opinion in Biomedical Engineering,* vol. 11, pp. 51-57, 2019, doi: 10.1016/j.cobme.2019.09.013.

[65] Q. Han *et al.,* "Health risk assessment and bioaccessibilities of heavy metals for children in soil and dust from urban parks and schools of Jiaozuo, China," *Ecotoxicology and Environmental Safety,* vol. 191, 2020, doi: 10.1016/j.ecoenv.2019.110157.

[66] J. Huang *et al.,* "A new exploration of health risk assessment quantification from sources of soil heavy metals under different land use," *Environmental Pollution,* vol. 243, pp. 49-58, 2018, doi: 10.1016/j.envpol.2018.08.038.

[67] J. Huang *et al.,* "Health risk assessment of heavy metal(loid)s in park soils of the largest megacity in China by using Monte Carlo simulation coupled with Positive matrix factorization model," *Journal of Hazardous Materials,* vol. 415, 2021, doi: 10.1016/j.jhazmat.2021.125629.

[68] Q. Cao *et al.,* "Heavy metals in homestead soil: Metal fraction contents, bioaccessibility, and risk assessment," *Journal of Hazardous Materials,* vol. 480, 2024, doi: 10.1016/j.jhazmat.2024.135933.

[69] L. Chen *et al.,* "A spatial source-oriented and probability-based risk-assessment framework for heavy metal and PAH contamination of urban soils in Guangzhou, China," *Journal of Hazardous Materials,* vol. 482, 2025, doi: 10.1016/j.jhazmat.2024.136500.

[70] K. Tian *et al.,* "Ecological risk assessment of heavy metals in sediments and water from the coastal areas of the Bohai Sea and the Yellow Sea," *Environment International,* vol. 136, 2020, doi: 10.1016/j.envint.2020.105512.

[71] X. Liu *et al.,* "Composition, distribution, and risk assessment of heavy metals in large-scale river water on the Tibetan Plateau," *Journal of Hazardous Materials,* vol. 476, 2024, doi: 10.1016/j.jhazmat.2024.135094.

[72] Z. Wang *et al.,* "Overview assessment of risk evaluation and treatment technologies for heavy metal pollution of water and soil," *Journal of Cleaner Production,* vol. 379, 2022, doi: 10.1016/j.jclepro.2022.134043.

[73] E. Solgi, A. Esmaili-Sari, A. Riyahi-Bakhtiari, and M. Hadipour, "Soil contamination of metals in the three industrial estates, Arak, Iran," *Bull Environ Contam Toxicol,* vol. 88, no. 4, pp. 634-8, Apr 2012, doi: 10.1007/s00128-012-0553-7.

[74] M. H. Eid *et al.,* "Comprehensive approach integrating water quality index and toxic element analysis for environmental and health risk assessment enhanced

by simulation techniques," *Environmental Geochemistry and Health,* vol. 46, no. 10, 2024, doi: 10.1007/s10653-024-02182-1.

[75] W. Huang *et al.*, "Source-specific soil heavy metal risk assessment in arsenic waste mine site of Yunnan: Integrating environmental and biological factors," *J Hazard Mater,* vol. 486, p. 136902, Mar 15 2025, doi: 10.1016/j.jhazmat.2024.136902.

[76] K. Cai *et al.*, "Heavy metal pollution from standardized e-waste dismantling activities: Pollution index, risk assessment and intervening measures," *J Hazard Mater,* vol. 484, p. 136614, Feb 15 2025, doi: 10.1016/j.jhazmat.2024.136614.

[77] S. R. Sany, S. R. Deb, F. Ahmed, M. A. A. Nayem, A. K. M. Ashikuzzaman, and M. A. A. Numanbakth, "Evaluation of groundwater quality and potential health risks in the Tengratila Gas Field Blowout Region, Bangladesh: An in-depth analysis utilizing multivariate statistics, heavy metal indices and Monte Carlo simulation," *J Hazard Mater,* vol. 490, p. 137744, Jun 15 2025, doi: 10.1016/j.jhazmat.2025.137744.

[78] H. F. Kaiser, "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika,* vol. 23, no. 3, pp. 187-200, 1958, doi: 10.1007/BF02289233.

[79] E. G. Dammi Djimi *et al.*, "Multivariate Statistical and Hydrochemical Analysis of Drinking Water Resources in Northern Cameroon Watersheds," *Water,* vol. 13, no. 21, 2021, doi: 10.3390/w13213055.

[80] Y. Yang *et al.*, "Multi-element features and trace metal sources of road sediment from a mega heavy industrial city in North China," *Chemosphere,* vol. 311, no. Pt 1, p. 137093, Jan 2023, doi: 10.1016/j.chemosphere.2022.137093.

[81] M. Fan and H. Liang, "Soil health assessment of dressing and smelting slag field based on heavy metal pollution-buffer-fertility three aspects," *J Hazard Mater,* vol. 482, p. 136602, Jan 15 2025, doi: 10.1016/j.jhazmat.2024.136602.

[82] E. Kravchenko *et al.*, "Ecological and health risk assessments of heavy metal contamination in soils surrounding a coal power plant," *J Hazard Mater,* vol. 484, p. 136751, Feb 15 2025, doi: 10.1016/j.jhazmat.2024.136751.

[83] J. Veskovic and A. Onjia, "Two-dimensional Monte Carlo simulation coupled with multilinear regression modeling of source-specific health risks from groundwater," *J Hazard Mater,* vol. 488, p. 137309, May 5 2025, doi: 10.1016/j.jhazmat.2025.137309.

[84] W. Zhou, Z. Li, Y. Liu, C. Shen, H. Tang, and Y. Huang, "Soil type data provide new methods and insights for heavy metal pollution assessment and driving factors analysis," *J Hazard Mater,* vol. 480, p. 135868, Dec 5 2024, doi: 10.1016/j.jhazmat.2024.135868.

[85] H. Pan, X. Lu, and K. Lei, "A comprehensive analysis of heavy metals in urban road dust of Xi'an, China: Contamination, source apportionment and spatial distribution," *Sci Total Environ,* vol. 609, pp. 1361-1369, Dec 31 2017, doi: 10.1016/j.scitotenv.2017.08.004.

[86] W. Deng, G. Hao, and W. Liu, "Source-specific risks apportionment and critical sources identification of potentially harmful elements in urban road dust combining positive matrix factorization model with multiple attribute decision making method," *Ecological Indicators,* vol. 144, 2022, doi: 10.1016/j.ecolind.2022.109449.

[87] B. Yu *et al.*, "Potentially toxic elements in surface fine dust of residence communities in valley industrial cities," *Environmental Pollution*, vol. 327, 2023, doi: 10.1016/j.envpol.2023.121523.
[88] J. Vesković, M. Sentić, and A. Onjia, "Hydrogeochemical Facies and Health Hazards of Fluoride and Nitrate in Groundwater of a Lithium Ore Deposit Basin," *Metals*, vol. 14, no. 9, 2024, doi: 10.3390/met14091062.
[89] J. Wang *et al.*, "Efficient ultra-trace electrochemical detection of Cd2+, Pb2+ and Hg2+ based on hierarchical porous S-doped C3N4 tube bundles/graphene nanosheets composite," *Chemical Engineering Journal*, vol. 420, 2021, doi: 10.1016/j.cej.2021.130317.
[90] P. Fan *et al.*, "Spatial distribution, risk estimation and source apportionment of potentially toxic metal(loid)s in resuspended megacity street dust," *Environ Int*, vol. 160, p. 107073, Feb 2022, doi: 10.1016/j.envint.2021.107073.
[91] Z. Zhao, M. Hao, Y. Li, and S. Li, "Contamination, sources and health risks of toxic elements in soils of karstic urban parks based on Monte Carlo simulation combined with a receptor model," *Sci Total Environ*, vol. 839, p. 156223, Sep 15 2022, doi: 10.1016/j.scitotenv.2022.156223.
[92] J. C. Egbueri, M. Abu, and J. C. Agbasi, "An integrated appraisal of the hydrogeochemistry and the potential public health risks of groundwater nitrate and fluoride in eastern Ghana," *Groundwater for Sustainable Development*, vol. 26, 2024, doi: 10.1016/j.gsd.2024.101264.
[93] J. Vesković and A. Onjia, "Environmental Implications of the Soil-to-Groundwater Migration of Heavy Metals in Mining Area Hotspots," *Metals*, vol. 14, no. 6, 2024, doi: 10.3390/met14060719.
[94] Y. Xu *et al.*, "Integrating machine learning for enhanced spatial prediction and risk assessment of soil heavy metal(loid)s," *Environmental Pollution*, vol. 383, p. 126919, 2025/10/15/ 2025, doi: https://doi.org/10.1016/j.envpol.2025.126919.
[95] A. S. El-Sorogy, K. Al-Kahtany, M. S. Shokr, T. Alharbi, and J. E. M. de Larriva, "On evaluating the possible hazard of soil contamination in arid regions using statistical analysis and GIS techniques," (in English), *Frontiers in Earth Science*, Original Research vol. Volume 13 - 2025, 2025-August-28 2025, doi: 10.3389/feart.2025.1605796.
[96] J. Li *et al.*, "Pollution characteristics and probabilistic risk assessment of heavy metal(loid)s in agricultural soils across the Yellow River Basin, China," *Ecological Indicators*, vol. 167, p. 112676, 2024/10/01/ 2024, doi: https://doi.org/10.1016/j.ecolind.2024.112676.
[97] S. Kumar and Anshumali, "Heavy metal pollution and health risk assessment in upland and riparian soils of the Ganga River basin," *Discover Soil*, vol. 2, no. 1, p. 33, 2025/05/08 2025, doi: 10.1007/s44378-025-00061-4.
[98] Z. Liu, L. Wang, M. Yan, Y. Tian, B. Ma, and Q. Xie, "Heavy metal synergistic pollution risk assessment in the soil-crop system of the nanyang basin," *Scientific Reports*, vol. 15, no. 1, p. 19937, 2025/06/06 2025, doi: 10.1038/s41598-025-05236-1.
[99] H. Wang, W. Li, C. Zhu, and X. Tang, "Analysis of Heavy Metal Pollution in Cultivated Land of Different Quality Grades in Yangtze River Delta of China," *International Journal of Environmental Research and Public Health*, vol. 18, no. 18, p. 9876, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/18/9876.

[100] W. Wang *et al.*, "Heavy Metal Pollution and Its Prior Pollution Source Identification in Agricultural Soil: A Case Study in the Qianguo Irrigation District, Northeast China," *Sustainability*, vol. 14, no. 8, p. 4494, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/8/4494.

[101] L. Sun *et al.*, "Improved model for human health risk assessment of soil heavy metal(loid)s in a mining area of southeastern China," (in eng), *Ecotoxicol Environ Saf*, vol. 306, p. 119312, Nov 4 2025, doi: 10.1016/j.ecoenv.2025.119312.

[102] L. Gao, G. Gopalakrishnan, Z. Wang, Y. Li, and Y. Zhang, "Characteristics, source apportionment, and health risk assessment of heavy metal pollution in soil of jiangmen, China," *Scientific Reports*, vol. 15, no. 1, p. 31578, 2025/08/27 2025, doi: 10.1038/s41598-025-17496-y.

[103] G. Yu, F. Chen, H. Zhang, and Z. Wang, "Pollution and health risk assessment of heavy metals in soils of Guizhou, China," *Ecosystem Health and Sustainability*, vol. 7, no. 1, p. 1859948, 2021/01/01 2021, doi: 10.1080/20964129.2020.1859948.

[104] M. K. Saifullah *et al.*, "Seasonal variation, and ecological risks assessment of heavy metals pollution in an upstream tropical freshwater river, Bangladesh," *International Journal of River Basin Management*, pp. 1-24, doi: 10.1080/15715124.2025.2523413.

[105] X. Chen, X. Fu, G. Li, J. Zhang, H. Li, and F. Xie, "Source-specific probabilistic health risk assessment of heavy metals in surface water of the Yangtze River Basin," *Science of The Total Environment*, vol. 926, p. 171923, 2024/05/20/ 2024, doi: https://doi.org/10.1016/j.scitotenv.2024.171923.

[106] X. Xu *et al.*, "Status of heavy metals contamination in the topsoil of major Oasis cities in Northwest China," (in eng), *Environ Geochem Health*, vol. 47, no. 9, p. 362, Aug 7 2025, doi: 10.1007/s10653-025-02659-7.

[107] D. Sun *et al.*, "Spatial and Temporal Variations in Heavy Metals in Lake and Reservoir Sediments in China: A Pollution Status and Risk Assessment," *Sustainability*, vol. 16, no. 21, p. 9218, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/16/21/9218.

[108] Z. Yin *et al.*, "Mapping mining-affected water pollution in China: status, patterns, risks, and implications," *Hydrol. Earth Syst. Sci.*, vol. 29, no. 16, pp. 3957-3973, 2025, doi: 10.5194/hess-29-3957-2025.

[109] Y. Guo *et al.*, "Determination and application of soil heavy metal geochemical baseline in the southern region of Wushan County in the Yangtze River Basin, China," *Scientific Reports*, vol. 15, no. 1, p. 10889, 2025/03/29 2025, doi: 10.1038/s41598-025-86404-1.

[110] Y. Xu, J. Zheng, Y. Cao, and Q. Jiang, "Policy Adjustments: Their Dual Impacts and Sustainability Assessment on Heavy Metal Concentration in Topsoil of Nantong City, China," *Sustainability*, vol. 17, no. 5, p. 1868, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/17/5/1868.

[111] A. Mussakulkyzy, C. Opp, N. Amirgaliev, A. Madibekov, L. Ismukhanova, and A. Zhadi, "Transmission of Heavy Metals in River Water and Self-Purification Capacity of Ile River," *Applied Sciences*, vol. 15, no. 12, p. 6548, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/12/6548.

[112] X. Dou, Q. Liu, Q. Fan, J. Guo, and W. Qi, "Comprehensive analysis of common heavy metals in the Yellow River over 20 Years: Spatiotemporal distribution, migration characteristics, traceability, and potential risk

evaluation," *Environmental Research*, vol. 281, p. 121931, 2025/09/15/ 2025, doi: https://doi.org/10.1016/j.envres.2025.121931.

[113] Z. Liu, L. Mo, J. Liang, H. Shi, J. Yao, and X. Lun, "Heavy Metal Pollution and Health-Ecological Risk Assessment in Agricultural Soils: A Case Study from the Yellow River Bend Industrial Parks," (in eng), *Toxics*, vol. 13, no. 10, Sep 30 2025, doi: 10.3390/toxics13100834.

[114] J. Ding *et al.*, "Content, Distribution, and Risk Assessment of Heavy Metals in Sandy Soils of the Ancient Yellow River Course," *Soil and Sediment Contamination: An International Journal*, vol. 34, no. 7, pp. 1911-1931, 2025/10/03 2025, doi: 10.1080/15320383.2025.2467237.

[115] P. Luo *et al.*, "Heavy metals in water and surface sediments of the Fenghe River Basin, China: assessment and source analysis," *Water Science and Technology*, vol. 84, no. 10-11, pp. 3072-3090, 2021, doi: 10.2166/wst.2021.335.

[116] K. Chen, J. Wang, Z. Zhang, H. Xiao, and J. Liu, "Analysis of spatiotemporal evolution and driving factors of ecological environment quality in the Ili-Balkhash Lake Basin in Central Asia," *Scientific Reports*, vol. 15, no. 1, p. 22190, 2025/07/01 2025, doi: 10.1038/s41598-025-06085-8.

[117] N. Carabal, E. Puche, S. Armenta, P. García-Atienza, and M. A. Rodrigo, "Constructed wetlands for the mitigation of pesticide and heavy metal concentrations in a protected agrolandscape: removal efficiencies and ecological risk assessment," *Science of The Total Environment*, vol. 1001, p. 180466, 2025/10/25/ 2025, doi: https://doi.org/10.1016/j.scitotenv.2025.180466.

[118] D. S. Allen *et al.*, "Predicted aquatic and human health risks associated with the presence of metals in the Syr Darya and Shardara Reservoir, Kazakhstan," *Science of The Total Environment*, vol. 859, p. 159827, 2023/02/10/ 2023, doi: https://doi.org/10.1016/j.scitotenv.2022.159827.