remote sensing

MDPI

Article

Assessing the Potential of EMIT Hyperspectral Data Combined with DEM-Derived Terrain Variables for Predicting Soil As, Cu and Zn Concentrations in a Mountainous Region of Southwest China

Guangping Qie [1,2,3], Minzi Wang [4], Ziping Pan [4], Zongdi Sun [5], Wenjin Xie [5], Zhiyi Liu [1] and Guangxing Wang [3,*]

[1] Department of Tourism Management, Moutai Institute, Renhuai 551801, China; qieguangping@mtxy.edu.cn or guangpingqie@siu.edu (G.Q.)
[2] Department of Geography and Environmental Resources, Southern Illinois University at Carbondale, Carbondale, IL 62901, USA
[3] School of Earth Systems and Sustainability, Southern Illinois University at Carbondale, Carbondale, IL 62901, USA
[4] School of Resource and Environment, Moutai Institute, Renhuai 551801, China; wangminzi@mtxy.edu.cn (M.W.)
[5] School of Business Administration, Moutai Institute, Renhuai 551801, China
* Correspondence: gxwang@siu.edu

Highlights

What are the main findings?

- EMIT hyperspectral predictors showed moderate potential for regional-scale soil As mapping, while Cu and Zn predictions were weaker under spatial cross-validation.
- Terrain information improved the representation of As-related topographic controls, but terrain-only information was insufficient for Cu and Zn.

What are the implications of the main findings?

- EMIT-based heavy metal mapping should be interpreted as an indirect prediction approach rather than a direct geochemical inversion method.
- Future work should integrate lithology, land use, anthropogenic source indicators, and more systematic field sampling to improve model transferability.

Abstract

Spaceborne imaging spectroscopy has created new opportunities for monitoring soil properties at regional scales. Its use for predicting soil heavy metal concentrations in mountainous environments, however, remains insufficiently tested, especially when EMIT hyperspectral data are used. In this study, EMIT Level-2A surface reflectance data were integrated with DEM-derived terrain variables to estimate soil arsenic (As), copper (Cu), and zinc (Zn) concentrations in Renhuai, Guizhou Province, Southwest China. Only soil samples falling within valid EMIT coverage were used for element-specific modeling, resulting in 139 samples for As, 136 for Cu, and 130 for Zn. To reduce redundancy among predictors, EMIT spectral variables and terrain factors were screened before model construction. Random forest and XGBoost models were then tested using repeated spatial cross-validation. The best-performing model for As combined EMIT predictors with elevation and achieved a validation R^2 of 0.460. Model performance was considerably weaker for Cu, with a validation R^2 of 0.188. For Zn, the model failed to outperform the mean-based benchmark, producing a negative validation R^2 of −0.028. The spatial prediction maps and residual

Check for updates

Academic Editors: Weichao Sun, Kun Shang and Naichen Xing

Received: 2 June 2026
Revised: 2 July 2026
Accepted: 3 July 2026
Published: 5 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

patterns suggested that the EMIT-based prediction showed moderate potential for As, limited predictive value for Cu, and exploratory rather than reliable mapping capability for Zn under the current sample and predictor conditions.

Keywords: EMIT; hyperspectral remote sensing; soil heavy metals; arsenic; copper; zinc; DEM-derived terrain variables; random forest; XGBoost; spatial cross-validation

1. Introduction

Soil heavy metal contamination is a persistent environmental concern because heavy metals and metalloids can accumulate in soils, affect ecosystem functions, enter food chains, and pose potential risks to human health [1,2]. Elements such as arsenic (As), copper (Cu), and zinc (Zn) are of particular interest because they are widely distributed in terrestrial environments and may originate from both natural parent materials and anthropogenic activities [3]. Although Zn and Cu are essential micronutrients at low concentrations, excessive accumulation may affect soil quality and crop safety [4], while As is a toxic metalloid of high environmental concern. Reliable spatial information on soil heavy metal concentrations is therefore important for environmental assessment [5], land management [6], and pollution prevention [7], especially in regions where terrain complexity and human activities jointly shape soil geochemical patterns [8,9].

Conventional assessment of soil heavy metal contamination still relies largely on field sampling followed by laboratory chemical analysis [10]. Such approaches are reliable and can produce accurate point-based measurements [10]. However, they are also time-consuming, expensive, and difficult to apply at large spatial scales, especially in heterogeneous landscapes where soil properties may vary sharply over short distances [10]. As a result, they often fail to capture the continuous spatial variability of heavy metal concentrations [11]. In recent years, hyperspectral remote sensing has become an increasingly useful tool for soil heavy metal monitoring. Its dense spectral information can reflect subtle differences in land surface and soil characteristics [12], and when integrated with machine learning methods, it provides a practical way to estimate heavy metal concentrations and generate spatial prediction maps [9,11,13,14]. However, the remote sensing of soil heavy metals remains challenging because most heavy metals do not have strong, direct, and unique absorption features in the VNIR and SWIR regions [15]. Their spectral responses are commonly indirect and mediated by soil color, iron oxides, clay minerals, organic matter, moisture, texture, vegetation cover, and land surface conditions [8,9].

The Earth Surface Mineral Dust Source Investigation (EMIT) mission provides new opportunities for spaceborne imaging spectroscopy [16]. EMIT records visible-to-shortwave-infrared reflected radiation from the International Space Station and provides Level-2A surface reflectance products with 285 spectral bands covering approximately 381–2493 nm [17]. Although EMIT was designed primarily to improve knowledge of mineral dust source composition, its dense visible-to-shortwave infrared spectral information is also potentially useful for broader land surface applications, including soil and mineral background characterization. For soil heavy metal prediction, EMIT may contribute spectral proxies associated with iron oxides, clay minerals, bare soil exposure, vegetation cover, and moisture conditions. Nevertheless, its moderate spatial resolution, irregular coverage, and sensitivity to land surface conditions require careful evaluation before it can be used for soil geochemical mapping in complex terrain.

Terrain variables play an important role in predicting soil heavy metal concentrations in mountainous areas [18]. DEM-derived attributes, including elevation, slope, aspect com-

ponents, topographic position, terrain roughness, wetness-related indices, and flow-related metrics, help describe how topography influences soil formation, erosion, deposition, surface runoff, soil moisture redistribution, and material transport [19–21]. These processes are particularly relevant in the mountainous landscapes of Southwest China. Steep slopes, deeply incised valleys, active soil erosion, and heterogeneous land-use patterns are common in this region. Under such conditions, terrain information can provide useful environmental covariates that complement hyperspectral reflectance data. In many cases, these topographic controls may capture spatial variation in soil heavy metals that cannot be fully explained by surface reflectance alone.

Machine learning methods are widely used in soil and environmental mapping because they can model nonlinear relationships and interactions among predictors [13,22–24]. Random forest (RF) is a robust ensemble method suitable for nonlinear and high-dimensional environmental prediction [25], whereas XGBoost is a gradient boosting framework with strong regularization and predictive capability in structured data problems [26]. However, model evaluation in spatial prediction requires careful validation. Random cross-validation can overestimate predictive performance when nearby samples occur in both training and validation folds [27], and spatially structured cross-validation is therefore recommended when the objective is to assess transferability to unsampled locations [23].

Although hyperspectral remote sensing has attracted increasing attention in soil heavy metal monitoring, several issues remain unresolved [28]. First, the potential of spaceborne EMIT hyperspectral data for estimating soil As, Cu, and Zn concentrations has not been sufficiently tested. Second, it is still uncertain whether EMIT-derived spectral predictors alone can provide enough information for heavy metal prediction in complex mountainous environments. Third, the contribution of DEM-derived terrain variables to EMIT-based prediction may differ among elements, but this element-specific effect has not yet been fully examined.

Therefore, this study assessed the potential and limitations of EMIT hyperspectral data combined with DEM-derived terrain variables for predicting soil As, Cu, and Zn concentrations in a mountainous region of Southwest China. The objectives were to: (1) construct EMIT-derived spectral predictor sets, including interpretable spectral features and screened hyperspectral bands; (2) integrate DEM-derived terrain variables to account for topographic controls; (3) compare RF and XGBoost models under different predictor combinations using repeated spatial cross-validation; and (4) generate spatial prediction and validation residual maps within the valid EMIT coverage area. The study emphasizes regional-scale predictive potential, model transferability, and residual behavior rather than direct geochemical inversion or formal uncertainty quantification.

2. Study Area, Datasets and Methods

2.1. Study Area

This study was conducted in Renhuai, Guizhou Province, Southwest China, a typical mountainous region characterized by complex terrain, strong topographic variation, and heterogeneous land surface conditions. The elevation within the study area ranges from approximately 290 to 1632 m, with mountains, hills, valleys, and slope systems forming a highly fragmented landscape (Figure 1). Such topographic complexity can influence soil formation, erosion, runoff, sediment redistribution, and the spatial accumulation or depletion of soil elements. Therefore, Renhuai provides a suitable case for assessing whether hyperspectral information from EMIT, when combined with DEM-derived terrain variables, can support the spatial prediction of soil heavy metals in complex mountainous environments.

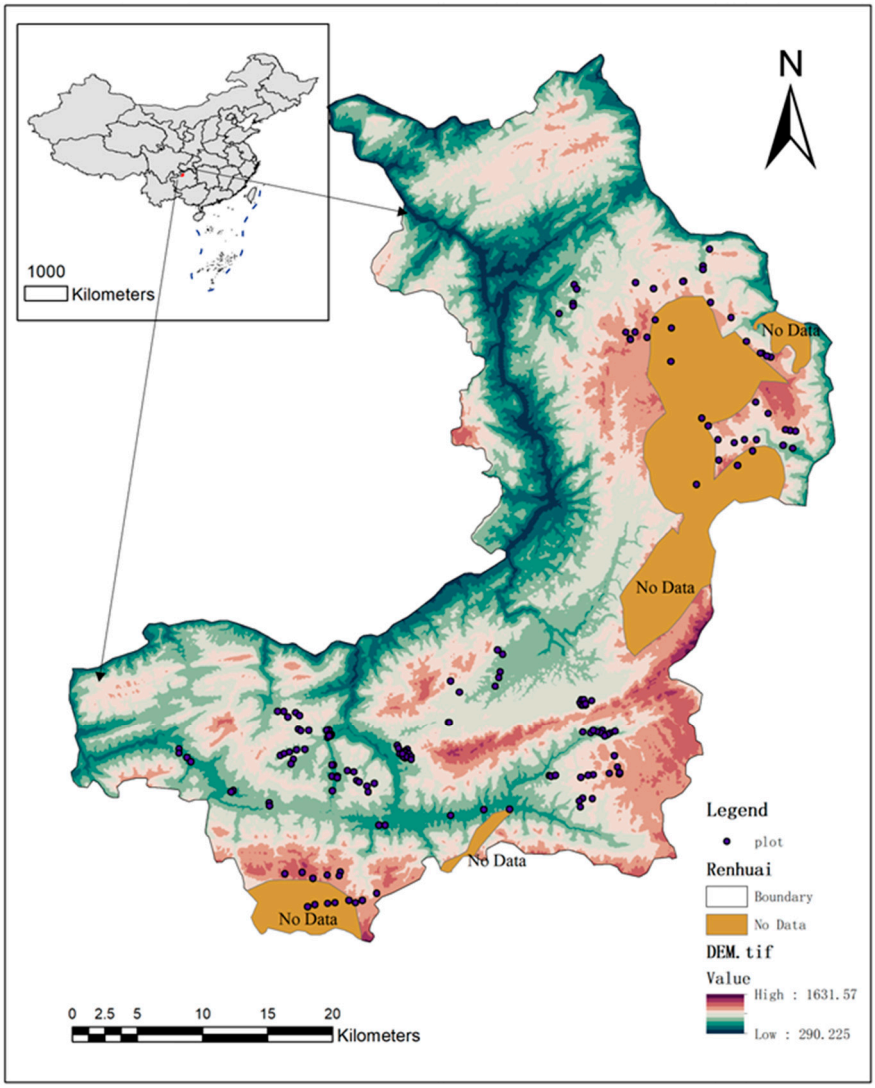


Figure 1. Location of the study area, terrain background, sampling sites, and valid EMIT coverage. The brown polygons indicate areas without valid EMIT data and were excluded from subsequent modeling and mapping.

The study area was initially defined according to the administrative boundary of Renhuai. However, the available EMIT observations did not cover the entire administrative region. For this reason, the final modeling and mapping domain was limited to areas with valid EMIT data. As illustrated in Figure 1, some parts of Renhuai lacked valid EMIT observations and were therefore classified as "No Data". These areas were not used in predictor extraction, model calibration, or spatial prediction. To maintain spatial consistency between hyperspectral and terrain predictors, all DEM-derived terrain variables were further clipped by the valid EMIT coverage mask. Consequently, the final prediction maps for soil As, Cu, and Zn were produced only within the overlapping area between the Renhuai administrative boundary and the valid EMIT coverage.

2.2. Soil Sampling and Heavy Metal Measurements

Soil sampling was conducted mainly between 2024 and 2025 using a field-accessible and environmentally guided purposive sampling strategy rather than a purely random sampling design. The sampling sites were selected to represent the major terrain and landscape conditions within the study area, including valley bottoms, slope areas, upland surfaces, and different elevation gradients. In addition, representative land-use settings

and bare or sparsely vegetated surfaces suitable for EMIT-based spectral extraction were considered where field conditions allowed. The final site selection was therefore based on both environmental representativeness and practical field accessibility. In mountainous areas such as Renhuai, completely random sampling is difficult to implement because many locations are restricted by steep terrain, poor road access, land-use permission, and field safety conditions. This strategy was necessary because fieldwork in mountainous regions such as Renhuai is strongly constrained by terrain accessibility, road conditions, land use, and field safety. As a result, the sampling sites were not evenly distributed across space, and local clustering occurred in several accessible parts of the study area, as shown in Figure 1.

For each target element, including As, Cu, and Zn, 158 soil samples were initially collected across the study area. At each sampling site, 3–5 subsamples were taken from the 0–20 cm surface soil layer within a circular area of approximately 50 m in diameter. The 0–20 cm depth was selected because it represents the surface soil layer most directly influenced by land-surface processes, atmospheric deposition, agricultural disturbance, and soil–vegetation interactions. This depth is also more consistent with the information captured by EMIT surface reflectance, which mainly represents near-surface land conditions rather than deeper soil horizons. These subsamples were fully mixed in the field and then reduced to approximately 1.5 kg to form one composite sample for each sampling unit. The composite sampling design was intended to represent the local soil condition around each sampling site rather than a single point-scale measurement. Because the spatial support of the composite sample was broadly comparable to the approximately 60 m EMIT pixel size, the field measurement was treated as a local average corresponding to the surrounding EMIT observation. Nevertheless, the sampling support and EMIT pixel footprint were not perfectly identical, and no explicit mixed-pixel removal was conducted in this study.

In the laboratory, all soil samples were naturally air-dried at room temperature. Visible stones, plant residues, and other impurities were removed by hand. The dried samples were then gently crushed using a wooden hammer or rod and passed through a 20-mesh nylon sieve. After sieving, each sample was quartered to approximately 500 g and stored in a sealed polyethylene bottle. For heavy metal analysis, approximately 100 g of each prepared sample was further ground with an agate mortar until it passed through a 100-mesh sieve, corresponding to a particle size of 0.149 mm. The concentrations of As, Cu, and Zn were determined by inductively coupled plasma-mass spectrometry (ICP-MS) according to the Chinese environmental standard HJ 1315-2023 [xx]. Briefly, 0.2000 g of the prepared soil powder was accurately weighed to the nearest 0.0001 g and transferred into a polytetrafluoroethylene digestion vessel. A mixed acid solution containing 6 mL of HNO_3, 2 mL of HCl, and 1 mL of HF was added for pre-digestion for 20 min. The samples were then digested in a microwave digestion system at 180 °C. After digestion, the vessels were placed on an acid evaporation system and heated at 140 °C until the remaining solution was concentrated to approximately 1 mL. The digest was subsequently transferred to a 50 mL volumetric flask and diluted to volume with ultrapure water. The final solutions were analyzed using an ICP-MS instrument (NexION 1000G, PerkinElmer, Shelton, CT, USA).

Quality assurance and quality control were implemented throughout the analytical procedure. All element measurements followed national standard methods. The method detection limits were 0.2 mg kg^{-1} for As, 0.7 mg kg^{-1} for Cu, and 5 mg kg^{-1} for Zn. For every 50 samples, five duplicate samples and two certified national soil reference materials, GSS-4a and GSS-5a, were included to monitor analytical precision and accuracy. More than 90% of the duplicate analyses met the relative error requirements. The measured values of the certified reference materials were also within the allowable ranges of the

recommended values. These quality-control results indicate that the analytical data were reliable for subsequent modeling.

Before model development, the soil sample locations were intersected with the valid EMIT coverage mask. Samples falling outside the valid EMIT coverage area were excluded, because complete hyperspectral predictor information was required for model calibration and validation. After this filtering step, 139 samples were retained for As, 136 for Cu, and 130 for Zn. These retained samples were used for model training, validation, and spatial prediction. Each retained sample also had valid DEM-derived terrain predictors, ensuring consistency among the soil measurements, EMIT spectral variables, and terrain covariates. Descriptive statistics for the retained As, Cu, and Zn samples are presented in Table 1. Before model construction, the distributions of both raw and ln-transformed heavy metal concentrations were examined. Although RF and XGBoost do not require normally distributed response variables, the ln-transformation was used as a pragmatic preprocessing step to reduce the leverage of strongly right-skewed concentration distributions and high-concentration observations under the relatively small sample size. Therefore, the transformation was not intended to satisfy parametric normality assumptions. The ln-transformation reduced the skewness of As, Cu, and Zn and was adopted for model calibration, while model evaluation metrics were calculated after back-transforming predictions to the original concentration scale (Figure A1). Specifically, predictions generated on the ln-transformed scale were back-transformed to the original concentration scale using direct exponentiation, namely predicted concentration = exp(predicted ln-concentration). No smearing correction or other bias-correction procedure was applied. We acknowledge that this simple retransformation may introduce retransformation bias because the exponentiated mean prediction on the log scale may not equal the expected value on the original scale, particularly when residual variance is non-negligible. Therefore, the original-scale predictions and accuracy metrics should be interpreted with this limitation in mind.

Table 1. Descriptive statistics of soil As, Cu, and Zn concentrations for retained samples within the valid EMIT coverage area. Concentration unit: mg kg^{-1}.

Metal	*n*	Min	Max	Median	Mean	SD	CV	Skewness	Response Used
As	139	4.201	53.101	15.003	16.411	8.2	0.50	1.93	ln(value)
Cu	136	12.05	139.003	38.407	42.309	20.6	0.49	1.75	ln(value)
Zn	130	49.04	229.006	83.002	88.895	27.8	0.31	1.62	ln(value)

Because the sampling design was not spatially random and the retained samples were unevenly distributed, spatial cross-validation was used for model evaluation. This validation strategy was adopted to reduce the potential overestimation of model performance caused by the spatial clustering of nearby samples. Accordingly, the prediction results should be interpreted as spatial estimates within the valid EMIT coverage area rather than as an extrapolation to the entire administrative boundary of Renhuai.

2.3. EMIT Hyperspectral Data

2.3.1. EMIT L2A Surface Reflectance Data

The EMIT Level-2A surface reflectance product was used as the primary hyperspectral remote sensing data source in this study. Due to the mountainous terrain and frequent cloudy conditions in Guizhou, it was difficult to obtain sufficient high-quality EMIT observations within a short acquisition period; therefore, usable EMIT scenes acquired between 27 December 2023 and 21 February 2026 were selected for subsequent compositing and modeling. This relatively long acquisition window was adopted as a practical compromise rather than on the assumption that surface reflectance remains completely stable

over time. The study area is characterized by a mountainous and humid climate with frequent cloudy conditions, which substantially limited the availability of usable EMIT observations. Even under the current temporal window, valid EMIT coverage remained incomplete in some parts of the study area. The data were accessed and processed in Google Earth Engine using the NASA/EMIT/L2A/RFL image collection [4,9]. EMIT provides visible-to-shortwave infrared surface reflectance observations with a spatial resolution of approximately 60 m. In this study, a total of 285 reflectance bands were used, covering wavelengths from 381.01 to 2492.92 nm, with full width at half maximum values ranging from 8.41 to 8.81 nm. A compact summary of the EMIT spectral bands used in this study is provided in Table A2.

EMIT scenes intersecting the Renhuai study area were first filtered according to spatial coverage, acquisition availability, and pixel-level quality information. After scene screening, eight EMIT scenes acquired between 27 December 2023 and 21 February 2026 were retained for subsequent compositing and modeling. The selected scenes had a mean valid-pixel ratio of 70.31%, with valid ROI coverage ranging from 1.28% to 53.70%. The cloud flag ratio and water flag ratio of the selected scenes ranged from 0.00% to 3.54% and from 0.00% to 0.26%, respectively, indicating generally acceptable observation quality for constructing the hyperspectral predictor dataset. To reduce the influence of date-specific surface disturbances, such as residual cloud contamination, short-term soil moisture variation, vegetation phenology, and dust-related effects, a pixel-level quality mask and multi-scene median compositing strategy were applied. The median composite was intended to suppress transient spectral noise while retaining the relatively stable regional spectral signal available from valid EMIT observations. The acquisition dates, image IDs, orbit and scene numbers, and key quality metrics of the selected EMIT scenes are listed in Table A1. Since the valid EMIT observations did not fully cover the entire study area, a valid EMIT coverage mask was explicitly generated. This mask was subsequently used to constrain the extraction of soil sample spectral values, DEM-derived terrain predictors, and final spatial prediction maps, thereby ensuring spatial consistency among soil observations, hyperspectral variables, and auxiliary environmental covariates. Accordingly, the resulting models and maps should be interpreted as regional-scale predictive estimates within valid EMIT coverage rather than as direct geochemical inversion under temporally invariant reflectance conditions.

2.3.2. EMIT Quality Masking and Scene Selection

Pixel-level quality masking was performed using the aggregate_flag band. Pixels with aggregate_flag = 0 were considered valid observations, whereas pixels affected by clouds, cirrus, water, spacecraft-related artifacts, or dilated cloud areas were excluded. For each EMIT scene, the valid-pixel ratio was calculated within the study area as:

$$\text{Valid} - \text{pixelratio} = \frac{N_{aggregate_{flag}=0}}{N_{covered\ pixels}} \quad (1)$$

where $N_{aggregate_{flag}=0}$ is the number of valid pixels and $N_{covered\ pixels}$ is the number of EMIT pixels intersecting the study area. Scenes with valid-pixel ratios greater than or equal to the predefined threshold were retained for compositing. The selected scenes were then masked using aggregate_flag = 0, and the median value of each reflectance band was calculated to generate a cloud-screened multi-scene EMIT reflectance composite.

In addition to the reflectance composite, a valid observation count layer was produced by summing the number of valid observations at each pixel. This layer was used to describe the spatial availability of EMIT observations and define the valid EMIT coverage mask.

2.3.3. EMIT Spectral Feature Construction

Two types of EMIT spectral predictors were constructed. First, selected physically interpretable spectral features were derived from representative visible, red-edge, near-infrared, and SWIR bands. These features were designed to represent soil color, vegetation cover, bare soil exposure, iron oxide-related spectral behavior, clay mineral-related spectral behavior, and moisture-related conditions. The selected spectral features included vegetation and soil indices such as NDVI, red-edge NDVI, NDMI, BSI, iron oxide ratios, redness indices, clay-related SWIR ratios, and absorption-depth approximations around 2200 nm and 2300 nm. To improve reproducibility, all candidate EMIT-derived spectral indices and spectral features investigated before feature selection are listed in Table A4, including their formulas or spectral definitions, related wavelength regions, and intended physical interpretation.

Second, the full 285-band EMIT reflectance composite was retained as the raw hyperspectral predictor set. Given the strong collinearity among adjacent hyperspectral bands, raw band selection was performed separately for As, Cu, and Zn rather than directly including all 285 bands in the final model. This strategy allowed the modeling framework to retain element-specific spectral information while reducing predictor redundancy and overfitting risk. The complete spectral range and band-group information of the EMIT predictors are summarized in Table A2. The term "element-specific feature screening" in this study refers to a filter-based candidate-variable screening step conducted separately for As, Cu, and Zn. Details of the missing-value filtering, near-zero-variance filtering, correlation-based redundancy screening, and terrain-variable screening are provided in Section 2.6.

2.4. DEM-Derived Terrain Variables

The DEM data used in this study were obtained from the National Basic Geographic Information Database provided by the National Geomatics Center of China (https://www.ngcc.cn/dlxxzy/gjjcdlxxsjk/, accessed on 4 February 2026), and were used to derive terrain variables for subsequent modeling. DEM and DEM-derived terrain variables were incorporated to characterize the influence of topography on the spatial distribution of soil heavy metals. This consideration is particularly important in mountainous regions, where terrain conditions regulate surface runoff, erosion intensity, sediment transport, soil moisture redistribution, slope processes, and the local accumulation or depletion of soil materials [10,11]. In this study, the terrain predictor set included elevation, slope, aspect, topographic position index, terrain ruggedness index, topographic wetness index, stream power index, specific catchment area, slope length–steepness factor, flow accumulation, curvature-related metrics, and local slope variability.

Representative DEM-derived variables are shown in Figure A2. The elevation layer reflects the pronounced topographic gradient across the study area, while slope, roughness, and terrain ruggedness index highlight areas with strong local relief and complex terrain fragmentation. Aspect exhibits substantial directional variability, indicating heterogeneous slope orientations across the landscape. The topographic position index further distinguishes local ridge-like and valley-like terrain positions, which may influence soil redistribution and the spatial accumulation of heavy metals.

To ensure that the terrain predictors were suitable for statistical modeling, several preprocessing steps were performed. Aspect was transformed into sine and cosine components to avoid the circularity problem associated with directional angles. Highly skewed variables, including flow accumulation, specific catchment area, stream power index, and slope length–steepness factor, were transformed using logarithmic or log-like functions to reduce the influence of extreme values. All DEM-derived predictors were subsequently

resampled, reprojected, and aligned to the EMIT composite in terms of spatial resolution, coordinate reference system, and spatial extent. These were also clipped by the valid EMIT coverage mask to ensure spatial consistency among terrain variables, hyperspectral predictors, and soil sampling locations.

2.5. Predictor Extraction, Modeling Database Construction, and Feature Set Design

All raster predictors were aligned to the same coordinate reference system, spatial resolution, and spatial extent. EMIT reflectance bands, selected spectral features, and DEM-derived terrain variables were stacked into predictor layers. Soil sample points were then overlaid on the raster predictor stack, and predictor values were extracted at each sampling location. This point-based extraction assumes that the composite soil sample represents the local surface condition captured by the corresponding EMIT pixel. However, because EMIT pixels may contain mixed land-surface components, such as soil, vegetation, roads, or built-up surfaces, pixel mixing may introduce uncertainty into the relationship between the measured soil concentrations and spectral predictors.

The extracted values were organized into element-specific modeling tables for As, Cu, and Zn. Each table contained the observed concentration, geographic coordinates, EMIT spectral variables, selected raw EMIT bands, and DEM-derived terrain variables. Records with missing response values or invalid predictor values were removed or imputed according to the model preprocessing procedure. Median imputation was performed within the training data during cross-validation to avoid information leakage from validation folds. After the modeling tables were constructed, three feature sets were defined to evaluate the relative contributions of EMIT spectral predictors and DEM-derived terrain variables. FS1 represented selected EMIT spectral predictors, including element-specific EMIT raw bands and derived spectral features retained after preprocessing and redundancy screening. FS2 represented selected DEM-derived terrain predictors that satisfied the element-specific terrain screening criterion. If no terrain variable met the criterion for a given metal, FS2 was excluded from model calibration for that metal. FS3 combined selected EMIT spectral predictors and selected DEM-derived terrain predictors. If FS2 was empty, FS3 was equivalent to FS1.

This combined section clarifies that predictor extraction, modeling-table construction, and feature-set definition were consecutive steps within the same modeling workflow. The hierarchical feature-set design allowed spectral and topographic contributions to be evaluated separately while avoiding inconsistent feature-set definitions.

2.6. Element-Specific Candidate-Variable Screening

Element-specific candidate-variable screening was conducted separately for As, Cu, and Zn before model calibration. This step was designed to reduce dimensionality and redundancy in the high-dimensional EMIT predictor space, rather than to represent a fully nested feature-selection procedure within each spatial cross-validation fold. First, predictors with high missing-value proportions were removed. Specifically, a predictor was removed if more than 20% of its values were missing within the retained element-specific modeling dataset. In the present dataset, no EMIT spectral feature or DEM-derived terrain variable exceeded this threshold, as summarized in Table 2. Second, zero- and near-zero-variance predictors were excluded. Third, redundancy among highly correlated predictors was reduced using target-guided correlation screening. Specifically, highly correlated predictor pairs were identified using an absolute pairwise correlation threshold of $|r| \geq 0.95$. When two predictors exceeded this threshold, the predictor with the stronger absolute Spearman correlation with the ln-transformed response variable was retained. Fourth, element-specific candidate sets were finalized for model calibration.

Table 2. Summary of predictor preprocessing for EMIT spectral variables and DEM-derived terrain variables.

Predictor Group	Candidate Variables	Removed by Missing-Value Filter	Removed by Zero/NZV Filter	Retained Predictors	Purpose in Modeling
Raw EMIT reflectance bands	285	0	29	256	High-dimensional hyperspectral information and raw-band screening
Selected EMIT spectral features	34	0	0	34	Physically interpretable spectral baseline predictors
DEM-derived terrain variables	9	0	0	9	Topographic controls on erosion, runoff, and material redistribution
Total	328	0	29	299	Final predictor pool before element-specific modeling

For terrain variables, the strongest terrain-related association was first identified for each element using Spearman correlation with the ln-transformed response. Terrain variables were retained only when the association remained statistically supported after adjustment for multiple testing. This rule retained DEM_elevation for As but excluded the weak and non-significant terrain candidates for Cu and Zn. Here, "adjustment for multiple testing" refers to the Benjamini–Hochberg false discovery rate correction applied to the *p*-values of the terrain–response correlation tests for each metal. This criterion was used to reduce the chance of retaining weak terrain predictors solely because of multiple univariate comparisons.

This procedure reduced the dimensionality of the EMIT and DEM predictor pool while preserving element-specific spectral and topographic information for subsequent RF and XGBoost modeling. Because the number of retained samples was limited and the EMIT hyperspectral bands were highly collinear, a filter-based screening strategy was adopted to improve model parsimony and reduce overfitting risk before model calibration. We acknowledge that this strategy mainly evaluates univariate predictor–response associations and may not fully capture nonlinear interactions among spectral and terrain variables. Therefore, the retained predictor sets should be interpreted as compact candidate sets for regional prediction rather than as an exhaustive representation of all potentially useful multivariate relationships.

2.7. Machine Learning Models

Two machine learning algorithms were used for soil heavy metal prediction: random forest and XGBoost. Random forest is an ensemble learning method based on bootstrap aggregation and random feature selection, and it is widely used for environmental prediction because of its robustness to nonlinear relationships and multicollinearity [22]. XGBoost is a gradient boosting framework that builds additive regression trees and has been widely used for structured prediction tasks because of its scalability and regularization capability [23].

For each element, RF and XGBoost models were calibrated using FS1, FS2, and FS3. Therefore, the modeling framework allowed for comparisons among different feature sets and between the two machine learning algorithms. Model hyperparameters were specified separately for RF and XGBoost. For RF models implemented using the ranger package, the number of trees was fixed at 500, and the splitting rule was fixed as variance-based regression splitting. The mtry parameter was tuned over candidate values defined according to the number of predictors in each feature set, and min.node.size was tuned over 3, 5, and 10. For XGBoost models, the regression objective was fixed as reg:squarederror, while nrounds, max_depth, eta, subsample, colsample_bytree, and min_child_weight were tuned using a limited grid search. Candidate values were 50, 100, and 200 for nrounds; 2, 3, and 4 for max_depth; 0.01, 0.05, and 0.10 for eta; 0.7 and 0.9 for subsample; 0.7 and

0.9 for colsample_bytree; and 1, 3, and 5 for min_child_weight. Hyperparameter selection was performed using only the training data within each spatial cross-validation split, and the held-out validation fold was not used during tuning. The parameter combination with the lowest training-only cross-validation RMSE was selected for validation prediction. All predictors were processed consistently across models to ensure fair comparison.

2.8. Spatial Cross-Validation and Accuracy Assessment

Spatial cross-validation was used to evaluate model performance. This approach was adopted because random cross-validation can overestimate predictive performance when spatially close samples occur in both training and validation sets. Spatially structured validation provides a more conservative and realistic assessment of spatial prediction performance in unsampled areas [23].

A fivefold spatial cross-validation was repeated nine times. In each repetition, sample points were spatially clustered into five folds based on projected coordinates. Each sample was predicted once per repetition, resulting in $n \times 9$ validation predictions for each metal. The final model performance was summarized across all repetitions.

Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), ratio of performance to deviation (RPD), and ratio of performance to interquartile distance (RPIQ). These metrics were calculated on the original concentration scale. For models trained on logarithmically transformed responses, predictions were back-transformed before accuracy assessment.

The evaluation metrics were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{4}$$

$$\mathrm{RPD} = \frac{SD_y}{RMSE} \tag{5}$$

$$\mathrm{RPIQ} = \frac{IQR_y}{RMSE} \tag{6}$$

where y_i is the observed concentration, $\hat{y}_i$ is the predicted concentration, $\overline{y}$ is the mean observed concentration, SD_y is the standard deviation of observed concentrations, and IQR_y is the interquartile range of observed concentrations.

2.9. Spatial Prediction Mapping

For each metal, the best-performing model was selected based on spatial cross-validation results. The selected model was then recalibrated using all available samples for that metal and applied to the raster predictor stack to generate spatial prediction maps. The final prediction maps were restricted to the valid EMIT coverage area to ensure that all mapped pixels contained both EMIT spectral information and DEM-derived terrain variables.

Separate spatial prediction maps were generated for As, Cu, and Zn. The maps were used to identify broad spatial patterns of soil heavy metal concentrations within the valid EMIT coverage area. The final outputs included raster prediction layers and map figures suitable for interpretation and publication.

2.10. Software

EMIT image preprocessing, quality masking, scene selection, and image export were conducted in Google Earth Engine. Raster processing, terrain variable integration, predictor extraction, feature selection, machine learning modeling, spatial cross-validation, accuracy assessment, and mapping were performed in R 4.4.1. The main R packages included terra, sf, dplyr, ggplot2, ranger, xgboost, and related packages for data processing and visualization.

3. Results

3.1. Predictor Preprocessing and Redundancy Diagnostics

Predictor preprocessing reduced the original candidate predictor pool before element-specific modeling (Table 2). The raw EMIT reflectance dataset contained 285 bands, of which 29 were removed by zero- or near-zero-variance filtering, leaving 256 preprocessed raw bands. The 34 selected EMIT spectral features and 9 DEM-derived terrain variables had complete values within the retained modeling domain. In total, 299 predictors were available before element-specific feature screening.

Correlation diagnostics showed strong redundancy among the candidate predictors, particularly within the raw EMIT reflectance bands (Table 3). The raw EMIT dataset contained 8689 highly correlated band pairs, with a mean absolute correlation coefficient of 0.967 and a maximum value of 1.000. The strongest correlation occurred between EMIT_B119 and EMIT_B120, which correspond to adjacent shortwave-infrared bands at approximately 1.26–1.27 μm, based on the EMIT band-center information summarized in Table A2. This strong correlation reflects the high spectral similarity between adjacent hyperspectral bands. Although RF and XGBoost are relatively robust to correlated predictors, strong redundancy can reduce model parsimony and complicate the interpretation of variable importance. Therefore, zero- and near-zero-variance filtering and correlation-based screening were applied before model calibration.

Table 3. High-correlation diagnostics for major predictor sets.

Predictor Set	Highly Correlated Pairs	Mean |r|	Maximum |r|	Strongest Pair
Raw EMIT reflectance bands	8689	0.967	1.000	EMIT_B119-EMIT_B120
Selected EMIT features + DEM variables	61	0.966	0.999	EMIT_SEL_015-EMIT_SEL_016
DEM-derived terrain variables	9	0.945	0.969	DEM_roughness-DEM_TRI

3.2. Correlation-Based Feature Screening for Model Development

Correlation-based feature screening was conducted to construct compact, element-specific candidate variable sets for As, Cu, and Zn. After preprocessing, 290 EMIT spectral predictors (256 raw EMIT bands plus 34 derived spectral features) and 9 DEM-derived terrain variables were screened for each metal. The number of retained EMIT predictors differed by element, reflecting differences in the strength and redundancy of spectral proxies. Terrain predictors were retained only when their adjusted correlation with the transformed response was statistically supported (Table 4).

Table 4. Summary of element-specific candidate variable screening for As, Cu, and Zn.

Metal	EMIT Predictors Screened	EMIT Predictors Retained	DEM Variables Screened	DEM Variables Retained	Final Candidate Predictors	Selection Rule
As	290	62	9	1	63	Correlation screening + adjusted p for DEM
Cu	290	13	9	0	13	Correlation screening + adjusted p for DEM
Zn	290	7	9	0	7	Correlation screening + adjusted p for DEM

The top-ranked terrain variable differed among the three metals (Table 5). DEM_elevation showed the strongest terrain-related association with As, with a Spearman correlation coefficient of 0.437 and an adjusted p value of 7.01×10^{-7}; it was therefore retained for model development. In contrast, DEM_TPI for Cu and DEM_aspect_cos for Zn showed weak and non-significant adjusted correlations and were not retained. Thus, Table 5 reports the top-ranked terrain association for each metal, whereas Table 4 reports the final number of terrain variables retained after screening. Only DEM_elevation for As was retained in the final feature set; the top-ranked terrain variables for Cu and Zn were excluded. These results suggest a clearer topographic signal for As than for Cu and Zn.

Table 5. Top-ranked DEM-derived terrain variable for each heavy metal based on Spearman correlation.

Metal	Top-Ranked Terrain Variable	*n*	Spearman r	|r|	Adjusted p	Model Status
As	DEM_elevation	139	0.437	0.437	7.01×10^{-7}	Retained
Cu	DEM_TPI	136	−0.175	0.175	0.370	Excluded
Zn	DEM_aspect_cos	130	0.193	0.193	0.195	Excluded

Note: The terrain variable with the strongest absolute Spearman correlation for each metal. "Retained" indicates that the variable passed the adjusted-significance criterion and was included in the model-ready feature set, whereas "Excluded" indicates that the variable was reported for diagnostic comparison only and was not included in subsequent model calibration.

The top-ranked raw EMIT bands also differed by metal (Table 6). For As, the strongest bands were concentrated in the EMIT_B064-EMIT_B074 range, with negative Spearman correlations of approximately −0.44. Cu showed stronger spectral associations than the other two metals, with important bands including EMIT_B128-EMIT_B132 and EMIT_B188-EMIT_B191. Zn shared several top-ranked bands with Cu, particularly in the EMIT_B188-EMIT_B191 and EMIT_B130-EMIT_B142 ranges. These element-specific differences indicate that the spectral proxies associated with As, Cu, and Zn were not identical.

Table 6. Selected raw EMIT bands for each heavy metal based on absolute Spearman correlation.

Metal	*n*	Main Top-Ranked EMIT Band Groups	Representative Top Bands	Spearman r Range	Adjusted p Range
As	139	EMIT_B064-EMIT_B074	B069, B072, B071, B074	−0.442 to −0.440	8.40×10^{-7}–8.40×10^{-7}
Cu	136	EMIT_B128-EMIT_B132; EMIT_B188-EMIT_B191	B132, B128, B131, B130, B129	−0.558 to −0.494	4.05×10^{-10}–2.62×10^{-8}
Zn	130	EMIT_B130-EMIT_B142; EMIT_B188-EMIT_B191	B141, B191, B190, B189, B188	−0.415 to −0.383	6.10×10^{-5}–1.74×10^{-4}

Note: Table 6 reports the dominant top-ranked band groups and representative bands.

Overall, preprocessing and screening reduced predictor redundancy and generated compact, element-specific feature sets for RF and XGBoost modeling. The screening results also indicate that the relative contribution of terrain information was element-specific.

3.3. Feature Set Construction and Modeling Readiness

Based on the element-specific screening results summarized in Table 4, three feature sets were constructed for subsequent RF and XGBoost modeling. Because the definitions of FS1, FS2, and FS3 have already been provided in the Methods section, the original feature-set definition table was removed to avoid redundancy. FS1 represented selected EMIT spectral predictors, FS2 represented selected DEM-derived terrain predictors, and FS3 combined the non-empty EMIT and terrain predictor sets. This design was used to evaluate whether DEM-derived terrain information provided additional explanatory information beyond the EMIT spectral predictors. To improve reproducibility, the final model-ready

feature sets after candidate-variable screening are additionally summarized in Table A5, including the retained predictor composition for each metal and feature-set combination.

Only variables satisfying the feature-screening criterion were carried forward into the model-ready feature sets. DEM_elevation was retained as the terrain predictor for As, whereas no DEM-derived terrain variable met the adjusted-correlation criterion for Cu or Zn. Consequently, FS2 was excluded for Cu and Zn, and FS3 was equivalent to FS1 for these two metals in terms of predictor composition (Table 7).

Table 7. Final model-ready feature-set composition after candidate-variable screening.

Metal	Feature Set	Samples (*n*)	EMIT Predictors	DEM Predictors	Total Predictors	Complete Cases	Status
As	FS1	139	62	0	62	139	Used
As	FS2	139	0	1	1	139	Used
As	FS3	139	62	1	63	139	Used
Cu	FS1	136	13	0	13	136	Used
Cu	FS2	136	0	0	0	0	Excluded
Cu	FS3	136	13	0	13	136	Used (equivalent to FS1)
Zn	FS1	130	7	0	7	130	Used
Zn	FS2	130	0	0	0	0	Excluded
Zn	FS3	130	7	0	7	130	Used (equivalent to FS1)

All non-empty feature sets had complete predictor values for the retained samples within the valid EMIT coverage area. The model-ready feature sets contained 63 predictors for As under FS3, 13 predictors for Cu under FS1/FS3, and 7 predictors for Zn under FS1/FS3. These compact feature sets were used in the subsequent RF and XGBoost model comparison.

3.4. Training–Validation Diagnostics and Residual Characteristics

The training–validation diagnostic results of the selected final models are shown in Table 8. The As model, based on FS3 and RF, achieved the best validation performance, with a validation R^2 of 0.460. However, the training–validation R^2 gap indicated a moderate decline in predictive performance from the training data to the spatial validation folds. The Cu model, based on FS1 and RF, showed a weak validation performance, with a validation R^2 of 0.188. The Zn model, based on FS1/FS3 and XGBoost, showed the weakest generalization ability and did not outperform the mean-based benchmark, with a slightly negative validation R^2. For Zn, FS3 was equivalent to FS1 because no DEM-derived terrain variable was retained after feature screening.

Table 8. Training–validation diagnostic metrics of the selected final models.

Metal	FS	Algorithm	Predictors	*n*	nCV	R^2_train	R^2_val	ΔR^2	RMSE	MAE	RPD	RPIQ
As	FS3	RF	63	139	1251	0.759	0.460	0.299	6.031	4.178	1.366	0.757
Cu	FS1	RF	13	136	1224	0.445	0.188	0.257	15.275	9.000	1.343	1.114
Zn	FS1/FS3	XGBoost	7	130	1170	0.290	−0.028	0.318	23.347	14.766	1.188	0.990

Note: nCV is the total number of validation predictions across the nine repeated spatial cross-validation runs. ΔR^2 was calculated as training R^2 minus validation R^2. RMSE and MAE were calculated on the original concentration scale. FS1 = selected EMIT spectral predictors; FS3 = selected EMIT spectral predictors plus selected DEM-derived terrain predictors. For Zn, FS3 was equivalent to FS1 because no DEM-derived terrain variable passed the screening criterion.

Residual summaries are provided in Table A3. The mean residuals were negative for all three metals, indicating a general tendency toward underestimation. Zn showed the largest residual variability, followed by Cu and As. The presence of large negative residuals

for Cu and Zn suggests that high-concentration samples were difficult to predict accurately. These results indicate that the selected models were more suitable for capturing broad regional patterns than local concentration extremes.

3.5. Spatial Prediction and Validation Residual Patterns

Based on the selected final models, spatial prediction maps of soil As, Cu, and Zn concentrations were generated within the valid EMIT coverage area (Figure 2). The three metals showed distinct spatial patterns. The predicted As concentrations exhibited strong local variability, with relatively high values mainly distributed in the southern and southeastern parts of the mapped area, as well as several localized high-value patches in the eastern region. In contrast, lower As concentrations were more widely distributed across the northern and central parts of the study area.

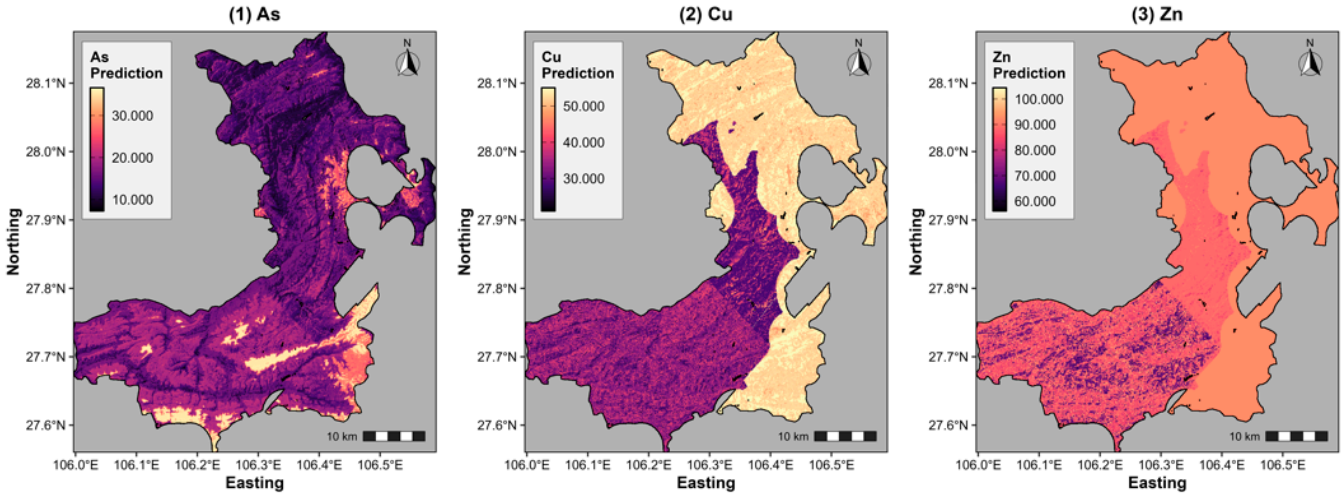


Figure 2. Spatial prediction maps of soil As, Cu, and Zn concentrations within the valid EMIT coverage area. Concentrations are expressed in mg kg^{-1}. Grey areas indicate regions without valid EMIT observations and were excluded from prediction.

The predicted Cu concentrations displayed a more distinct large-scale spatial gradient than those of As. Relatively high Cu values were mainly distributed in the northern and eastern parts of the valid EMIT coverage area. In contrast, lower values were mostly concentrated in the central-western and southwestern regions. The predicted Zn map showed a smoother pattern when compared with As and Cu. Higher Zn concentrations were mainly found in the eastern and northeastern parts of the study area, whereas lower values tended to occur in the west and southwest. Taken together, these prediction maps suggest that the selected models captured broad spatial heterogeneity in soil heavy metal concentrations. However, the spatial patterns were clearly element-specific, with noticeable differences among As, Cu, and Zn.

Validation residual maps were also examined to evaluate the spatial distribution of model errors (Figure 3). These residual maps were used as diagnostic tools for identifying spatial error patterns, rather than as formal uncertainty maps or quantitative uncertainty estimates. For As, the residuals were generally smaller than those observed for Cu and Zn. This agrees with the stronger validation performance of the As model reported in Table 8. Both positive and negative residuals appeared across the validation samples, suggesting that the As model did not exhibit a consistent spatial bias. Even so, several clusters of negative residuals were found in the southwestern part of the study area, indicating local underestimation in that region.

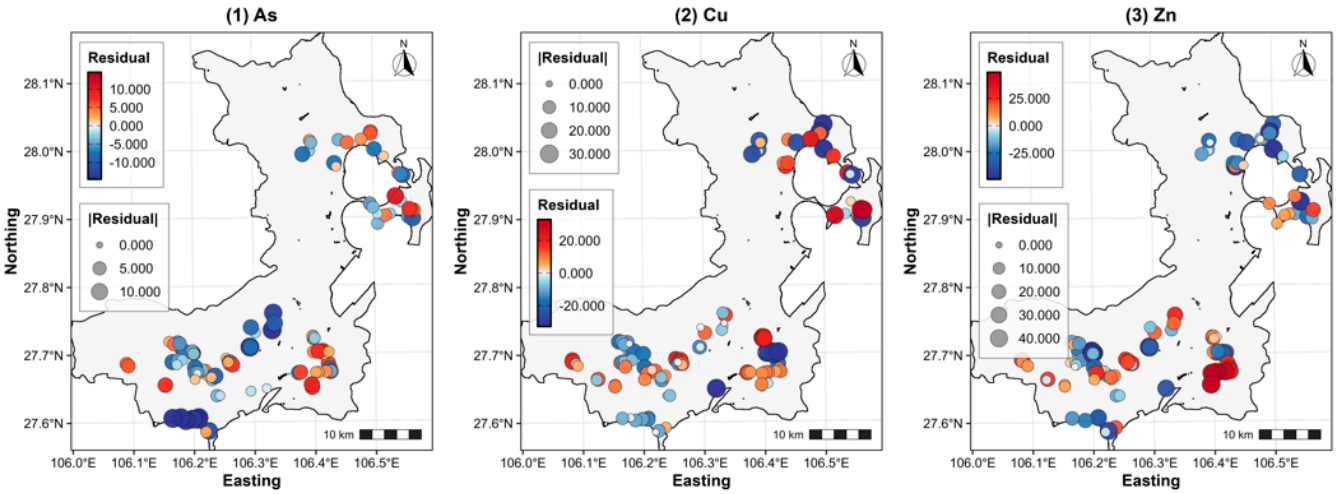


Figure 3. Spatial distribution of validation residuals for the selected final models of As, Cu, and Zn. Residuals were calculated as predicted minus observed concentrations and are expressed in mg kg^{-1}. Positive residuals indicate overestimation, whereas negative residuals indicate underestimation.

For Cu and Zn, the residuals were larger and showed more obvious spatial clustering. The Cu residual map revealed both overestimation and underestimation across different parts of the validation area. Relatively large residuals were mainly observed in the eastern and southern sampling clusters. Zn showed an even wider range of residual magnitudes, which is consistent with the residual statistics presented in Table A3 and the weak validation performance reported in Table 8. In particular, several large negative residuals indicated that the final model tended to underestimate areas with high Zn concentrations.

Overall, the spatial pattern of residuals suggests that the selected models were better at capturing regional-scale gradients than local concentration extremes. The relatively small and less clustered residuals for As further support its stronger predictive performance. In contrast, the larger and more spatially clustered residuals for Cu and Zn indicate that the current predictor set was not sufficient to explain their local variability. Additional covariates, such as land use, geological background, and anthropogenic source-related factors, may therefore be needed to improve future model performance.

4. Discussion

4.1. Potential and Limitations of EMIT Hyperspectral Data for Soil Heavy Metal Prediction

This study explored the potential of EMIT hyperspectral data along with terrain variables derived from DEM for predicting soil concentrations of As, Cu, and Zn in a mountainous area of Southwest China. The results indicate that spectral information from EMIT provided useful but limited insights for predicting soil heavy metals. This aligns with earlier reviews that noted that hyperspectral remote sensing can be beneficial for monitoring soil heavy metals, but the spectral response tends to be indirect and dependent on the environment [29–31]. Most heavy metals do not have strong and unique absorption features in the visible–near-infrared and shortwave-infrared regions [32]. Instead, their spectral relationships are often influenced by soil color, iron oxides, clay minerals, organic matter, moisture, vegetation cover, and bare soil exposure [33–36]. Therefore, any interpretation linking the better As prediction to iron oxides, clay-related features, weathering conditions, or hydrological transport should be regarded as a possible indirect explanation rather than direct evidence of a confirmed geochemical mechanism. In this study, soil Fe, clay content, organic matter, pH, and detailed geological variables were not available; therefore, the specific roles of iron oxides, mineral weathering, soil chemical conditions, and hydrological transport could not be directly tested. The retained EMIT spectral predictors and DEM

elevation may reflect broad surface and topographic proxies associated with soil-forming environments, but they should not be interpreted as direct indicators of As-bearing minerals or specific transport pathways.

Previous studies have used different spectral platforms to estimate soil heavy metal concentrations, including field spectroscopy, multispectral Sentinel-2 imagery, Hyperion hyperspectral imagery, airborne hyperspectral data, and simulated satellite hyperspectral data [8,31,37,38]. These studies generally suggest that spectral remote sensing can provide useful indirect information for soil heavy metal mapping, but prediction performance varies considerably depending on sensor characteristics, spatial resolution, spectral resolution, soil exposure, sample size, environmental covariates, and validation strategy. Therefore, the validation performance obtained in the present study should be interpreted within the specific context of spaceborne EMIT observations, complex mountainous terrain, incomplete valid EMIT coverage, and spatially structured validation.

Among the three metals, As demonstrated the best validation performance, while Cu and Zn showed a weaker ability to generalize. This suggests that the spectral and terrain proxies in this study were more effective in representing the variability of As than for Cu and Zn. One possible reason is that As distribution might be more closely connected to mineralogical and terrain-related processes, such as its association with iron oxides, weathering, and hydrological movement. On the other hand, Cu and Zn could be more influenced by local human activities, land use, agriculture, transportation, or industrial sources that were not included in the current predictors [39]. However, the validation R^2 of the best As model remained below 0.5, indicating moderate rather than strong predictive performance. Compared with studies based on controlled field spectra, airborne hyperspectral data, or more favorable bare-soil conditions [8,31,37,38], the present study provides a more conservative assessment because it used spaceborne EMIT data with approximately 60 m spatial resolution in a humid mountainous region. Vegetation cover, soil moisture, mixed pixels, cloud-related data gaps, incomplete spatial coverage, and rugged topography all increase the difficulty of soil heavy metal prediction in this setting.

Thus, the contribution of this study lies not in demonstrating high-accuracy geochemical inversion, but in providing an exploratory and conservative assessment of the regional-scale applicability and limitations of EMIT hyperspectral data for soil heavy metal prediction in complex mountainous environments.

The moderate performance of the As model and the weak performance of the Cu and Zn models support the key finding of this study: EMIT hyperspectral data can offer useful indirect information for regional soil heavy metal prediction, but spectral data alone cannot fully explain the complex soil heavy metal patterns in mountainous areas. Therefore, mapping soil heavy metals using EMIT should be viewed as a regional-scale prediction method rather than a direct method for geochemical analysis.

4.2. Contribution of DEM-Derived Terrain Variables in a Mountainous Landscape

Terrain variables derived from DEM helped represent the environmental factors that control soil heavy metal distribution. Terrain influences erosion, runoff, sediment movement, soil moisture redistribution, and the local accumulation or loss of soil materials [20,21,40,41]. In mountainous areas, these processes can significantly affect how soil elements are distributed [42]. In digital soil mapping, terrain variables are commonly used as environmental covariates because they provide information on landscape position, drainage conditions, and material redistribution processes [43,44].

However, the present results provide only limited and element-specific evidence for the effectiveness of DEM-derived terrain variables. DEM_elevation was retained only for As, whereas no DEM-derived terrain variable passed the screening criterion for Cu or Zn.

Therefore, terrain variables should not be interpreted as broadly effective predictors for all three metals under the current modeling framework. The better performance of the As model using FS3 suggests that combining terrain information with EMIT spectral predictors improved the representation of As spatial variability. Nevertheless, this improvement should be interpreted cautiously because the best validation R^2 was still moderate, and terrain information was not evaluated as an independently strong predictor of soil heavy metal concentrations.

In contrast, Cu was modeled using FS1, while Zn used FS3 but still showed weak validation performance because FS3 was essentially the same as FS1 for Zn after screening for terrain. These outcomes indicate that the contribution of terrain variables depended on the specific element. For Cu and Zn, the absence of retained terrain variables does not necessarily mean that topography has no relevance. Rather, it means that no terrain variable showed a sufficiently strong marginal association with the response variable under the current filter-based screening rule. Because RF and XGBoost can capture nonlinear relationships and interactions, some terrain variables with weak univariate correlations may still have predictive value when combined with spectral variables. Therefore, the current feature-screening strategy may have underestimated the potential role of terrain–spectral interactions.

Accordingly, DEM-derived terrain variables should be interpreted as potentially supportive auxiliary predictors rather than as conclusive evidence of terrain control. Future studies should test full terrain-covariate sets, nested feature selection, wrapper-based or embedded variable-selection strategies, and interaction-aware modeling approaches to better evaluate the potential contribution of terrain variables in EMIT-based heavy metal mapping.

4.3. Interpretation of the Spatial Prediction Patterns

The spatial prediction maps displayed distinct patterns for As, Cu, and Zn. The As prediction map showed relatively high local variability, with localized high-value patches. This matches the moderate validation performance of the As model and indicates that the combined EMIT–terrain predictors captured some regional variation in As concentrations. However, the spatial pattern should still be interpreted as a regional prediction surface rather than as a direct map of geochemical mechanisms or contamination boundaries.

The Cu and Zn prediction maps exhibited more abrupt and large-scale spatial changes. The Cu map highlighted a clear difference between higher predicted concentrations in the northern and eastern regions and lower values in the central-western and southwestern areas. The Zn map revealed a smoother yet more distinctly partitioned spatial pattern, showing higher predicted concentrations in the east and lower values in the west and southwest. These large spatial changes should be interpreted cautiously. One potential reason for the abrupt transitions in Cu and Zn is the behavior of tree-based machine learning models. Random forest and XGBoost can capture nonlinear relationships, but they may produce segmented predictions when key thresholds among predictors divide the landscape into distinct groups [45–47]. This effect may be more pronounced when the final model uses a limited set of predictors or when predictor values vary sharply across space. For Zn, the final model utilized XGBoost and only a small number of predictors, which could have led to the smooth but strongly partitioned prediction surface. For Cu, reliance on FS1 suggests that the prediction primarily reflected the chosen EMIT spectral proxies, which might indicate broad land surface variations rather than detailed soil geochemical differences. Another factor is that the distribution of valid EMIT observations and the sampling sites was uneven. Several regions lacked valid EMIT data, leaving the prediction domain somewhat incomplete. Additionally, field sampling faced challenges due to accessibility and terrain conditions. These elements may have limited the models'

ability to recognize gradual transitions in underrepresented areas. As such, the sharp contrasts in Cu and Zn should not be viewed as exact boundaries of soil contamination; instead, they should be seen as regional-scale prediction patterns derived from the available EMIT, DEM, and sample data.

The weaker Cu and Zn prediction patterns may also reflect the absence of source-related predictors. In Renhuai, agricultural activity, transportation, construction land, and brewing-related industrial activity may introduce localized or non-point-source metal inputs. These anthropogenic signals are not directly represented by EMIT reflectance or DEM-derived terrain variables and may partly obscure natural spectral and topographic gradients. This interpretation is consistent with multi-source soil heavy metal mapping studies showing that remote sensing, topography, soil background, land use, and human-activity covariates can jointly influence prediction performance [48].

4.4. Model Generalization and Residual Behavior

The training–validation diagnostics indicated that model generalization differed substantially among the three metals. As had the highest validation R^2, while Cu showed weak predictive ability and Zn did not outperform the mean-based validation benchmark. These results are consistent with the validation residual maps (Figure 3) and the residual summary in Table A3. The As residuals were generally smaller, whereas Cu and Zn showed larger residual magnitudes and more evident spatial clustering.

The negative mean residuals for all three metals indicate a general tendency toward underestimation. This underestimation was especially evident for Zn, which also had the largest residual variability. A likely explanation is that tree-based models tend to smooth extreme values when high-concentration samples are rare or spatially clustered. This is common in environmental prediction, especially when the response variable is highly heterogeneous and the sample size is limited. The use of ln-transformed responses may further reduce sensitivity to high-concentration extremes, even though predictions were back-transformed to the original concentration scale for model evaluation and mapping. Therefore, high-concentration hotspots should be interpreted cautiously under the current modeling framework.

The residual patterns also show that the models were more effective in capturing broad spatial gradients than local concentration extremes. This finding agrees with the use of spatially structured validation, which generally provides a more conservative assessment of model performance than random cross-validation when spatial autocorrelation is present [49–53]. Therefore, the moderate-to-weak validation results should not be viewed only as a modeling limitation; they also reflect the difficulty of predicting soil heavy metals under realistic spatial validation conditions, limited sample size, incomplete EMIT coverage, mixed pixels, and heterogeneous mountainous terrain. Therefore, the moderate-to-weak validation results should not be viewed only as a modeling limitation; they also indicate that the spatial prediction task is challenging under realistic validation conditions.

4.5. Implications for EMIT-Based Soil Heavy Metal Mapping

The results have several implications for the use of EMIT in soil heavy metal prediction. First, EMIT provides valuable hyperspectral information for characterizing surface conditions, but heavy metal prediction remains indirect. The EMIT L2A reflectance product provides dense visible-to-shortwave infrared spectral information, but its moderate spatial resolution and partial coverage can limit fine-scale soil prediction in complex terrain. Mixed pixels, vegetation cover, bare soil exposure, and land surface heterogeneity may weaken the relationship between point-level soil measurements and pixel-level reflectance.

Second, DEM-derived terrain variables may provide useful auxiliary information, but their contribution was limited and element-specific in this study. Terrain predictors help represent physical redistribution processes in mountainous areas, but they cannot fully describe geological background, land use, agricultural inputs, road effects, or industrial sources. This is particularly relevant for Cu and Zn, whose spatial variability may be more strongly linked to human activities or local source inputs. In Renhuai, agricultural activities and the brewing-related industrial economy are important local geographical contexts that may influence soil Cu and Zn distributions. Non-point source inputs from fertilizer and pesticide use may introduce additional Cu- and Zn-related variability into cultivated or peri-urban soils, while point-source or localized disturbances associated with brewing industries, transportation, construction land, and related supporting activities may further modify the soil metal patterns. These anthropogenic inputs are not directly represented by EMIT spectral variables or DEM-derived terrain predictors and may therefore partly mask the natural spectral and topographic signals. This may help explain why Cu and Zn showed weaker validation performance than As under the current EMIT–terrain predictor framework.

Third, RF and XGBoost should be used with careful validation and interpretation. RF was generally more stable in this study, whereas XGBoost produced weak generalization for Zn. This does not mean that XGBoost is unsuitable for soil heavy metal prediction, but rather that it may be more sensitive to limited sample size, feature selection, and spatially uneven training data. For regional soil prediction with limited samples, model simplicity, predictor relevance, and validation design are as important as algorithm complexity.

Overall, EMIT-based soil heavy metal mapping should be developed as a multi-source environmental prediction framework rather than as a spectral-only inversion method. Future work should integrate EMIT hyperspectral predictors with lithology, soil type, soil pH, organic matter, Fe and clay content, land use, agricultural intensity, road density, settlement distribution, industrial-source indicators, and more systematic field sampling. Such integration would better represent both natural soil-forming processes and anthropogenic source effects, and may improve the prediction of Cu and Zn in complex mountainous regions.

4.6. Limitations and Future Research

Several limitations should be emphasized. First, the EMIT composite was constructed from scenes acquired over a relatively long period because usable observations were limited in the cloudy and mountainous study area. Although quality masking and median compositing were applied, temporal mismatch between soil sampling and EMIT acquisition may still affect surface reflectance. Second, the sample size was limited, and the sampling design was constrained by road access, terrain accessibility, land-use permission, and field safety. Third, mixed pixels were not explicitly removed, which may have weakened the relationship between field-measured soil concentrations and EMIT spectral predictors.

Another important limitation is the filter-based candidate-variable screening strategy. Because screening relied primarily on univariate associations between predictors and the response variable, some terrain or spectral variables with weak individual correlations may have been excluded even though they could contribute through multivariate or nonlinear interactions. This limitation is particularly relevant to the interpretation of terrain variables. Therefore, the present study should be understood as an evaluation of screened EMIT–terrain predictor sets rather than a complete assessment of all possible terrain–spectral interactions. Future studies should use larger and more systematically distributed sample datasets, seasonally matched imagery, pixel-purity screening, soil physicochemical variables, geological data, anthropogenic source indicators, and fully nested feature-selection procedures to improve model robustness and interpretability.

5. Conclusions

This study evaluated the potential of EMIT hyperspectral data combined with DEM-derived terrain variables for predicting soil As, Cu, and Zn concentrations in a mountainous region of Southwest China. The results showed that EMIT hyperspectral predictors provide useful but indirect information for soil heavy metal prediction. The inclusion of DEM-derived terrain information improved the representation of topographic controls for As, but terrain variables did not provide sufficient additional explanatory power for Cu and Zn under the current screening rule.

Among the three target metals, As showed the strongest validation performance, with an R^2 of 0.460. Cu performed much more weakly, with an R^2 of only 0.188. For Zn, the model failed to outperform the mean-based validation benchmark, as indicated by a negative R^2 value of -0.028. These differences suggest that the current EMIT–terrain predictor set was more effective in capturing As-related spatial variability than that of Cu and Zn. The poor performance for Cu and Zn also indicates that their spatial distributions may be controlled by factors not adequately represented in the present predictor set. Additional information on lithology, land use, agricultural inputs, traffic-related emissions, and other anthropogenic sources may therefore be necessary.

Under the current sample size and feature-set conditions, RF generally produced more stable results than XGBoost. The spatial prediction maps were able to reflect broad regional patterns of soil heavy metal concentrations. However, local concentration extremes and abrupt spatial changes should be interpreted with caution, especially for Cu and Zn. Overall, this study demonstrates the potential of EMIT hyperspectral data for regional soil heavy metal prediction, but it also reveals clear limitations. Future work should place greater emphasis on integrating multi-source environmental covariates, mapping prediction uncertainty, and improving the representativeness of field sampling in complex mountainous landscapes.

Author Contributions: Conceptualization, G.Q., G.W. and M.W.; methodology, G.Q. and Z.P.; software, G.Q. and Z.S.; validation, G.W. and M.W.; formal analysis, G.Q. and W.X.; investigation, G.W. and G.Q.; resources, G.Q.; data curation, G.Q. and Z.S.; writing—original draft preparation, G.Q.; writing—review and editing, G.W., Z.L. and M.W.; visualization, G.Q. and G.W.; supervision, G.W.; project administration, G.W., G.Q., Z.S., W.X. and M.W.; funding acquisition, G.W., G.Q., Z.S., W.X. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by 2025 Humanities and Social Sciences Research Project of Colleges and Universities of the Guizhou Provincial Department of Education (Grant No. 2025RW071); the 2025 Key Project of Zunyi Philosophy and Social Science Planning Program (Grant No. 25ZYZD03); Guizhou Provincial Theoretical Innovation Project (Tendered Project) (Grant No. GZLCZB2026-08); the Guizhou Provincial Science and Technology Program (Grant No. QN (2025) 293); High-Level Talent Start-up Project of Moutai Institute (Grant No. mygccrc[2023]042); High-Level Talent Start-up Project of Moutai Institute (Grant No. mygccrc[2022]125); and High-Level Talent Start-up Project of Moutai Institute (Grant No. mygccrc[2022]138).

Data Availability Statement: EMIT L2A surface reflectance data are publicly available through NASA Earthdata and the Google Earth Engine data catalogue. DEM-derived products, processed predictor layers, and model outputs can be made available from the corresponding author upon reasonable request, subject to data-use restrictions for field sampling locations.

Acknowledgments: The authors thank the field assistants for their support during soil sampling.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EMIT	Earth Surface Mineral Dust Source Investigation
DEM	Digital elevation model
VNIR	Visible-near-infrared
SWIR	Shortwave infrared
GEE	Google Earth Engine
RF	Random forest
XGBoost	Extreme gradient boosting
CV	Cross-validation
RMSE	Root mean square error
MAE	Mean absolute error
RPD	Ratio of performance to deviation
RPIQ	Ratio of performance to interquartile distance

Appendix A

The appendix provides additional information on EMIT scene selection, EMIT spectral-band metadata, residual summaries, DEM-derived terrain variables, and distributional diagnostics of soil heavy metal concentrations.

Table A1. Selected EMIT scenes used in this study.

No.	Acquisition Date (UTC)	Image ID	Orbit	Scene	Valid Pixel Ratio (%)	Valid ROI Coverage (%)	Cloud Flag Ratio (%)	Water Flag Ratio (%)
1	27 December 2023	EMIT_L2A_RFL_001_20231227T030234_2336102_002	2336102	002	49.91	1.28	0	0
2	27 December 2023	EMIT_L2A_RFL_001_20231227T030246_2336102_003	2336102	003	57.2	3.38	0	0
3	21 March 2024	EMIT_L2A_RFL_001_20240321T083524_2408106_005	2408106	005	99.84	10.09	0	0.172
4	22 August 2024	EMIT_L2A_RFL_001_20240822T042727_2423503_017	2423503	017	90.53	4.71	0.012	0
5	26 August 2024	EMIT_L2A_RFL_001_20240826T025021_2423902_013	2423902	013	85.9	46.34	0.078	0.264
6	26 August 2024	EMIT_L2A_RFL_001_20240826T025033_2423902_014	2423902	014	51.87	6.98	0.505	0.072
7	23 October 2024	EMIT_L2A_RFL_001_20241023T035511_2429703_004	2429703	004	72.98	8.04	0.071	0.075
8	21 February 2026	EMIT_L2A_RFL_001_20260221T032936_2605203_002	2605203	002	54.23	53.7	3.54	0.076

Table A2. Summary of EMIT spectral bands used in this study.

Spectral Range	Number of Bands	Band Index Range	Wavelength Range (nm)	Mean Wavelength (nm)	FWHM Range (nm)	Mean FWHM (nm)
Visible/VNIR lower range (<700 nm)	43	0–42	381.01–693.47	537.06	8.41–8.47	8.44
Near infrared (700–1000 nm)	41	43–83	700.93–999.37	850.14	8.47–8.53	8.50
SWIR (1000–2500 nm)	201	84–284	1006.83–2492.92	1751.18	8.53–8.81	8.67
Total EMIT bands used	285	0–284	381.01–2492.92	1438.37	8.41–8.81	8.61

Table A3. Residual distribution of the selected final models on the independent validation set.

Metal	Feature Set	Algorithm	*n*	Mean	SD	Median	Q1	Q3	Min	Max	Mean Abs.	Median Abs.
As	FS3	RF	139	−1.172	5.937	0.189	−3.520	2.681	−18.636	8.636	4.178	3.045
Cu	FS1	RF	136	−1.806	18.452	−0.010	−10.399	10.557	−90.575	32.549	9.000	10.498
Zn	FS3	XGBoost	130	−4.518	27.844	0.229	−18.950	12.195	−140.331	51.430	14.766	14.571

Note: Residuals were calculated as predicted minus observed values. Negative mean residuals indicate systematic underestimation.

Table A4. Candidate spectral indices/features identified from the available modeling table before feature selection.

Feature Group	Index/Feature	Formula or Spectral Definition	Intended Interpretation
Vegetation	NDVI	(NIR − Red)/(NIR + Red)	Vegetation greenness
Vegetation	EVI	2.5(NIR − Red)/(NIR + 6Red − 7.5Blue + 1)	Enhanced greenness
Vegetation	GNDVI	(NIR − Green)/(NIR + Green)	Chlorophyll-sensitive greenness
Vegetation/soil	SAVI	(1 + L)(NIR − Red)/(NIR + Red + L), L = 0.5	Soil-adjusted vegetation signal
Moisture	NDMI	(NIR − SWIR1)/(NIR + SWIR1)	Moisture-related signal
Moisture	LSWI	(NIR − SWIR1)/(NIR + SWIR1)	Land-surface water signal
Water/background	MNDWI	(Green − SWIR1)/(Green + SWIR1)	Water/background separation
Bare soil	BSI	[(SWIR1 + Red) − (NIR + Blue)]/[(SWIR1 + Red) + (NIR + Blue)]	Bare-soil exposure
Built-up/soil	NDBI	(SWIR1 − NIR)/(SWIR1 + NIR)	Built-up/dry-soil response
Soil/tillage	NDTI	(SWIR1 − SWIR2)/(SWIR1 + SWIR2)	SWIR contrast for soil/residue
Soil/background	SI	Workflow-defined soil index based on visible/NIR/SWIR reflectance	Soil brightness/background

Note: Blue, Green, Red, NIR, SWIR1, and SWIR2 denote surface reflectance in the corresponding bands. The SI definition should follow the original preprocessing workflow if the original code is retrieved later.

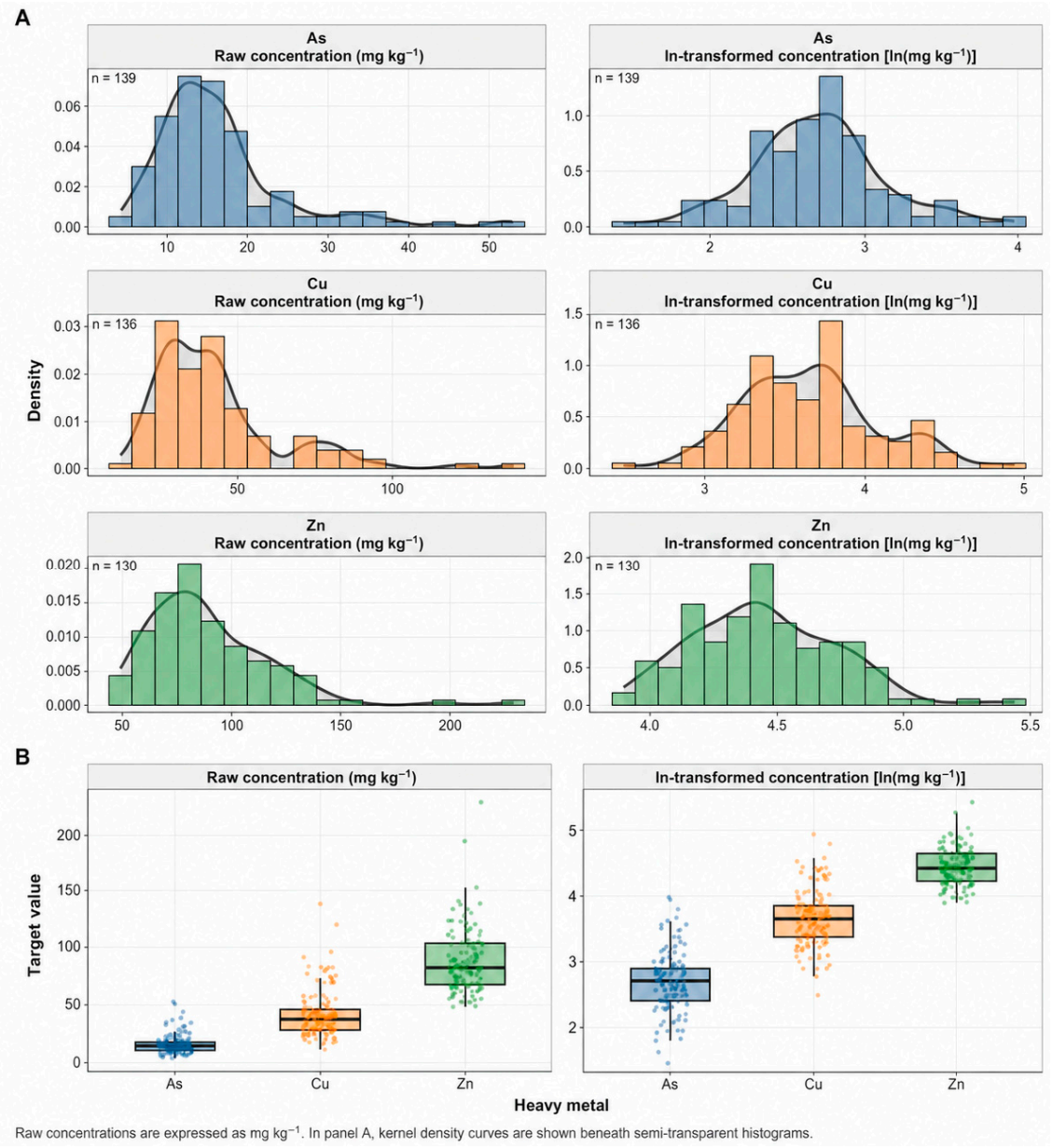


Figure A1. Distribution of heavy metal concentrations within the valid EMIT coverage area. (**A**) Histograms with overlaid kernel density curves showing the raw and natural-log-transformed concentrations of As, Cu, and Zn; *n* indicates the valid sample size for each metal. (**B**) Boxplots with jittered observations comparing the raw and natural-log-transformed target values of As, Cu, and Zn. Raw concentrations are expressed as mg kg^{-1}.

Table A5. Summary of final model-ready feature sets after candidate-variable screening.

Metal	Feature Set	Retained EMIT Predictors	Retained DEM-Derived Predictors	Total Predictors	Model Status
As	FS1	62	0	62	Used
As	FS2	0	1; DEM_elevation	1	Used
As	FS3	62	1; DEM_elevation	63	Used
Cu	FS1	13	0	13	Used
Cu	FS2	0	0	0	Excluded
Cu	FS3	13	0	13	Used; equivalent to FS1
Zn	FS1	7	0	7	Used
Zn	FS2	0	0	0	Excluded
Zn	FS3	7	0	7	Used; equivalent to FS1

Note: FS1 = selected EMIT spectral predictors; FS2 = selected DEM-derived terrain predictors; FS3 = selected EMIT spectral predictors plus selected DEM-derived terrain predictors. For Cu and Zn, FS3 was equivalent to FS1 because no DEM-derived terrain variable passed the screening criterion. This table summarizes the final model-ready feature-set composition reported in the manuscript; it does not list each individual EMIT band name.

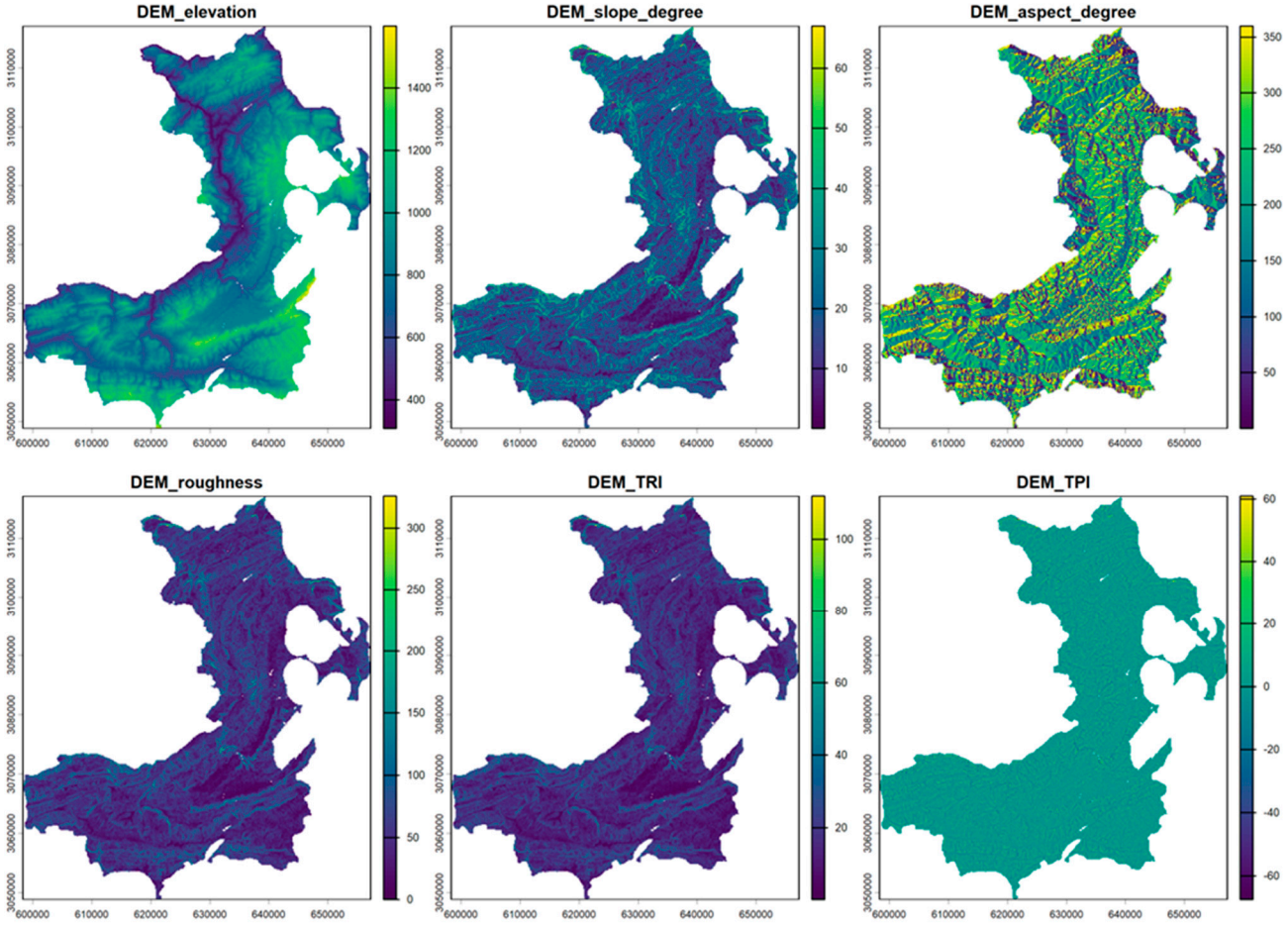


Figure A2. Representative DEM-derived terrain variables used to characterize topographic controls in the study area. The figure shows selected representative terrain variables derived from the DEM, including elevation, slope, aspect, roughness, terrain ruggedness index (TRI), and topographic position index (TPI). These variables were used to describe the complex mountainous terrain of the study area and to represent potential topographic controls on soil heavy metal redistribution, including erosion, runoff, sediment transport, local accumulation, and slope-position effects. Only part of the DEM-derived predictor set is displayed here for visual interpretation, while the full terrain-variable set was used according to the modeling framework described in the Methods section.

References

1. Alengebawy, A.; Abdelkhalek, S.T.; Qureshi, S.R.; Wang, M.-Q. Heavy Metals and Pesticides Toxicity in Agricultural Soil and Plants: Ecological Risks and Human Health Implications. *Toxics* **2021**, *9*, 42. [CrossRef] [PubMed]
2. Mishra, S.; Singh, G.; Gupta, A.; Tiwari, R.K. Heavy Metal/Metalloid Contamination: Their Sources in Environment and Accumulation in Food Chain. In *Heavy Metal Toxicity: Environmental Concerns, Remediation and Opportunities*; Singh, R.P., Singh, P., Srivastava, A., Eds.; Springer Nature: Singapore, 2023; pp. 19–47, ISBN 978-981-99-0397-9.
3. Szynkowska, M.I.; Pawlaczyk, A.; Maćkiewicz, E. Bioaccumulation and Biomagnification of Trace Elements in the Environment. In *Recent Advances in Trace Elements*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2018; pp. 251–276, ISBN 978-1-119-13378-0.
4. Sharma, V.; Dhaliwal, S.S.; Verma, V.; Kaur, M. Role of Micronutrients in Alleviating Abiotic Stresses to Improve the Productivity and Quality of Sugarcane. In *Climate-Smart Sugarcane Cultivation*; Apple Academic Press: Palm Bay, FL, USA, 2025.
5. Ersoy, A. Critical review of the environmental investigation on soil heavy metal contamination. *Appl. Ecol. Env. Res.* **2021**, *19*, 3853–3878. [CrossRef]
6. Chen, T.; Chang, Q.; Liu, J.; Clevers, J.G.P.W.; Kooistra, L. Identification of Soil Heavy Metal Sources and Improvement in Spatial Mapping Based on Soil Spectral Information: A Case Study in Northwest China. *Sci. Total Environ.* **2016**, *565*, 155–164. [CrossRef] [PubMed]
7. Xie, Y.; Chen, T.; Lei, M.; Yang, J.; Guo, Q.; Song, B.; Zhou, X. Spatial Distribution of Soil Heavy Metal Pollution Estimated by Different Interpolation Methods: Accuracy and Uncertainty Analysis. *Chemosphere* **2011**, *82*, 468–476. [CrossRef] [PubMed]
8. Wang, Y.; Zou, B.; Chai, L.; Lin, Z.; Feng, H.; Tang, Y.; Tian, R.; Tu, Y.; Zhang, B.; Zou, H. Monitoring of Soil Heavy Metals Based on Hyperspectral Remote Sensing: A Review. *Earth-Sci. Rev.* **2024**, *254*, 104814. [CrossRef]
9. Lovynska, V.; Bayat, B.; Bol, R.; Moradi, S.; Rahmati, M.; Raj, R.; Sytnyk, S.; Wiche, O.; Wu, B.; Montzka, C. Monitoring Heavy Metals and Metalloids in Soils and Vegetation by Remote Sensing: A Review. *Remote Sens.* **2024**, *16*, 3221. [CrossRef]
10. Sharma, S.; Tiwari, S.; Hasan, A.; Saxena, V.; Pandey, L.M. Recent Advances in Conventional and Contemporary Methods for Remediation of Heavy Metal-Contaminated Soils. *3 Biotech.* **2018**, *8*, 216. [CrossRef] [PubMed]
11. Wu, Z.; Chen, Y.; Han, Y.; Ke, T.; Liu, Y. Identifying the Influencing Factors Controlling the Spatial Variation of Heavy Metals in Suburban Soil Using Spatial Regression Models. *Sci. Total Environ.* **2020**, *717*, 137212. [CrossRef] [PubMed]
12. Bayer, A.D.; Bachmann, M.; Rogge, D.; Müller, A.; Kaufmann, H. Combining Field and Imaging Spectroscopy to Map Soil Organic Carbon in a Semiarid Environment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 3997–4010. [CrossRef]
13. Yu, H.; Kong, B.; Wang, G.; Du, R.; Qie, G. Prediction of Soil Properties Using a Hyperspectral Remote Sensing Method. *Arch. Agron. Soil Sci.* **2018**, *64*, 546–559. [CrossRef]
14. Wang, F.; Gao, J.; Zha, Y. Hyperspectral Sensing of Heavy Metals in Soil and Vegetation: Feasibility and Challenges. *ISPRS J. Photogramm. Remote Sens.* **2018**, *136*, 73–84. [CrossRef]
15. Qiu, W.; Tang, T.; He, S.; Zheng, Z.; Lv, J.; Guo, J.; Zeng, Y.; Lao, Y.; Wu, W. Inversion Studies on the Heavy Metal Content of Farmland Soils Based on Spectroscopic Techniques: A Review. *Agronomy* **2025**, *15*, 1678. [CrossRef]
16. Green, R. *EMIT L2A Estimated Surface Reflectance and Uncertainty and Masks 60 m V001 [Data Set]*; NASA EOSDIS Land Processes Distributed Active Archive Center (DAAC): Sioux Falls, South Dakota, 2022; EMITL2ARFL.001. [CrossRef]
17. Coleman, R.W.; Thompson, D.R.; Brodrick, P.G.; Dor, E.B.; Cox, E.; Pérez García-Pando, C.; Hoefen, T.; Kokaly, R.F.; Meyer, J.M.; Ochoa, F.; et al. An Accuracy Assessment of the Surface Reflectance Product from the EMIT Imaging Spectrometer. *Remote Sens. Environ.* **2024**, *315*, 114450. [CrossRef]
18. Hu, B.; Geng, Y.; Zou, Q.; Zhu, Q.; He, X.; Yang, J.; Wu, C.; Zhang, P.; Xia, F.; Shi, Z. Mapping Soil Heavy Metal Using an Interpretable Framework With Multi-Sources Data. *Land Degrad. Dev.* **2026**, *early view*. [CrossRef]
19. Gessler, P.E.; Moore, I.D.; McKenzie, N.J.; Ryan, P.J. Soil-Landscape Modelling and Spatial Prediction of Soil Attributes. *Int. J. Geogr. Inf. Syst.* **1995**, *9*, 421–432. [CrossRef]
20. Ågren, A.M.; Lidberg, W.; Strömgren, M.; Ogilvie, J.; Arp, P.A. Evaluating Digital Terrain Indices for Soil Wetness Mapping—A Swedish Case Study. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 3623–3634. [CrossRef]
21. Li, X.; McCarty, G.W.; Karlen, D.L.; Cambardella, C.A. Topographic Metric Predictions of Soil Redistribution and Organic Carbon in Iowa Cropland Fields. *CATENA* **2018**, *160*, 222–232. [CrossRef]
22. Qie, G.; Ye, J.; Wang, G.; Wang, M. Enhancing Urban Above-Ground Vegetation Carbon Density Mapping: An Integrated Approach Incorporating De-Shadowing, Spectral Unmixing, and Machine Learning. *Forests* **2024**, *15*, 480. [CrossRef]
23. Qie, G.; Wang, M.; Wang, G. Estimation of Urban Above-Ground Vegetation Carbon Density and Analysis of Topography-Modulated Spectral Responses in Shenzhen, China. *Remote Sens.* **2026**, *18*, 807. [CrossRef]
24. Li, J.; Li, X.; Wu, J.; Wang, C.; Liu, J.-Z.; Zang, F. Precision Mapping and Driving Factors of Heavy Metal(Loid)s in Agricultural Soils of the Yellow River: An Integrated Machine Learning and Geodetector Approach. *Environ. Res.* **2026**, *292*, 123681. [CrossRef] [PubMed]

25. Kafy, A.A.; Nowrin, M.T.; Altuwaijri, H.A.; Mukarram, M.M.T. Advanced Hybrid Ensemble Learning Models for Atmospheric Humidity Prediction: An Environmental Monitoring Solution in Arid Regions. *Int. J. Environ. Sci. Technol.* **2026**, *23*, 219. [CrossRef]
26. Liu, Y.; Lai, F.; Liu, N.; Sun, J.; Wang, R. Research on Mineral Calculation Based on Element Yield and XGBoost Optimization Algorithm: A Case Study of Shale Oil Reservoir in Weixinan Depression. *Geol. J.* **2026**, *61*, 1310–1324. [CrossRef]
27. Pohjankukka, J.; Pahikkala, T.; Nevalainen, P.; Heikkonen, J. Estimating the Prediction Performance of Spatial Models via Spatial K-Fold Cross Validation. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2001–2019. [CrossRef]
28. Wang, X.; Tan, K.; Du, P.; Pan, C.; Ding, J. A Unified Multiscale Learning Framework for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4508319. [CrossRef]
29. Shi, T.; Guo, L.; Chen, Y.; Wang, W.; Shi, Z.; Li, Q.; Wu, G. Proximal and Remote Sensing Techniques for Mapping of Soil Contamination with Heavy Metals. *Appl. Spectrosc. Rev.* **2018**, *53*, 783–805. [CrossRef]
30. Dai, X.; Wang, Z.; Liu, S.; Yao, Y.; Zhao, R.; Xiang, T.; Fu, T.; Feng, H.; Xiao, L.; Yang, X.; et al. Hyperspectral Imagery Reveals Large Spatial Variations of Heavy Metal Content in Agricultural Soil—A Case Study of Remote-Sensing Inversion Based on Orbita Hyperspectral Satellites (OHS) Imagery. *J. Clean. Prod.* **2022**, *380*, 134878. [CrossRef]
31. Wang, Y.; Zhang, X.; Sun, W.; Wang, J.; Ding, S.; Liu, S. Effects of Hyperspectral Data with Different Spectral Resolutions on the Estimation of Soil Heavy Metal Content: From Ground-Based and Airborne Data to Satellite-Simulated Data. *Sci. Total Environ.* **2022**, *838*, 156129. [CrossRef] [PubMed]
32. Shen, Q.; Zhang, S.; Xia, K. Spectral Heterogeneity Analysis and Soil Organic Matter Inversion across Differences in Soil Types and Organic Matter Content in Dryland Farmland in China. *Sustainability* **2023**, *15*, 16310. [CrossRef]
33. Schwertmann, U. Relations Between Iron Oxides, Soil Color, and Soil Formation. In *Soil Color*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1993; pp. 51–69, ISBN 978-0-89118-926-8.
34. Ibáñez-Asensio, S.; Marqués-Mateu, A.; Moreno-Ramón, H.; Balasch, S. Statistical Relationships between Soil Colour and Soil Attributes in Semiarid Areas. *Biosyst. Eng.* **2013**, *116*, 120–129. [CrossRef]
35. Poppiel, R.R.; Lacerda, M.P.C.; Rizzo, R.; Safanelli, J.L.; Bonfatti, B.R.; Silvero, N.E.Q.; Demattê, J.A.M. Soil Color and Mineralogy Mapping Using Proximal and Remote Sensing in Midwest Brazil. *Remote Sens.* **2020**, *12*, 1197. [CrossRef]
36. Escadafal, R. Soil Spectral Properties and Their Relationships with Environmental Parameters—Examples from Arid Regions. In *Imaging Spectrometry—A Tool for Environmental Observations*; Hill, J., Mégier, J., Eds.; Springer Netherlands: Dordrecht, The Netherlands, 1994; pp. 71–87, ISBN 978-0-585-33173-7.
37. Felegari, S.; Sharifi, A.; Khosravi, M.; Sabanov, S.; Tariq, A.; Karuppannan, S. Using Sentinel-2 Data to Estimate the Concentration of Heavy Metals Caused by Industrial Activities in Ust-Kamenogorsk, Northeastern Kazakhstan. *Heliyon* **2023**, *9*, e21908. [CrossRef] [PubMed]
38. Yang, L.Y.; Gao, X.H.; Zhang, W.; Shi, F.F.; He, L.H.; Jia, W. Estimating heavy metal concentrations in topsoil from vegetation reflectance spectra of Hyperion images: A case study of Yushu County, Qinghai, China. *Ying Yong Sheng Tai Xue Bao* **2016**, *27*, 1775–1784. [CrossRef] [PubMed]
39. *Encyclopedia of Global Environmental Change*; Munn, T., Ed.; Wiley: New York, NY, USA, 2002; ISBN 978-0-471-97796-4.
40. Pennock, D.J. Terrain Attributes, Landform Segmentation, and Soil Redistribution. *Soil Tillage Res.* **2003**, *69*, 15–26. [CrossRef]
41. Sørensen, R.; Zinko, U.; Seibert, J. On the Calculation of the Topographic Wetness Index: Evaluation of Different Methods Based on Field Observations. *Hydrol. Earth Syst. Sci.* **2006**, *10*, 101–112. [CrossRef]
42. Moore, I.D.; Grayson, R.B.; Ladson, A.R. Digital Terrain Modelling: A Review of Hydrological, Geomorphological, and Biological Applications. *Hydrol. Process.* **1991**, *5*, 3–30. [CrossRef]
43. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
44. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.M.; Gonzalez, M.R.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef] [PubMed]
45. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
46. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
47. Milà, C.; Ludwig, M.; Pebesma, E.; Tonne, C.; Meyer, H. Random Forests with Spatial Proxies for Environmental Modelling: Opportunities and Pitfalls. *Geosci. Model Dev.* **2024**, *17*, 6007–6033. [CrossRef]
48. Liu, Q.; Du, B.; He, L.; Zeng, Y.; Tian, Y.; Zhang, Z.; Wang, R.; Shi, T. Digital Soil Mapping of Heavy Metals Using Multiple Geospatial Data: Feature Identification and Deep Neural Network. *Ecol. Indic.* **2023**, *154*, 110863. [CrossRef]
49. Meyer, H.; Reudenbach, C.; Wöllauer, S.; Nauss, T. Importance of Spatial Predictor Variable Selection in Machine Learning Applications—Moving from Data Reproduction to Spatial Prediction. *Ecol. Model.* **2019**, *411*, 108815. [CrossRef]

50. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nat. Commun.* **2020**, *11*, 4540. [CrossRef] [PubMed]
51. Meyer, H.; Pebesma, E. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [CrossRef]
52. Wang, Y.; Khodadadzadeh, M.; Zurita-Milla, R. Spatial+: A New Cross-Validation Method to Evaluate Geospatial Machine Learning Models. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *121*, 103364. [CrossRef]
53. Stock, A. Choosing Blocks for Spatial Cross-Validation: Lessons from a Marine Remote Sensing Case Study. *Front. Remote Sens.* **2025**, *6*, 1531097. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.